Poznámky z přednášek
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Strojové učení

Peter Černo, 2010
petercerno@gmail.com

Garant: Mgr. Marta Vomlelová, Ph.D.
E-mail: Marta.Vomlelova@mff.cuni.cz
Domácí stránka: `http://ktiml.ms.mff.cuni.cz/~marta/`

18.01.2010 Mo
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

LEARNING PROBLEMS
- SUPERVISED - THE GOAL IS TO PREDICT THE VALUE OF AN OUTCOME MEASURE BASED ON A NUMBER OF INPUT MEASURES ...
- UNSUPERVISED - THERE IS NO OUTCOME MEASURE, AND THE GOAL IS TO DESCRIBE THE ASSOCIATIONS AND PATTERNS AMONG A SET OF INPUT MEASURES

OVERVIEW OF SUPERVISED ~~LE~~LEARNING

INPUTS ↦ OUTPUTS, GOAL: USE INPUTS TO PREDICT OUTPUTS

OUTPUTS ↳ DEPENDENT VARIABLES, RESPONSES

INPUTS ↑ PREDICTORS, INDEPENDENT VARIABLES, FEATURES

VARIABLE TYPES: QUANTITATIVE, ORDERED CATEGORICAL, QUALITATIVE (CATEGORICAL, DISCRETE, FACTORS) → CLASSES

- REGRESSION: WE PREDICT QUANTITATIVE OUTPUTS
- CLASSIFICATION: WE PREDICT QUALITATIVE OUTPUTS

TARGETS: NUMERIC CODES FOR ~~QUAL~~QUALITATIVE VARIABLES WITH TWO CLASSES / CATEGORIES. ("FAILURE"... 0 / "SUCCESS"... 1)

MORE THAN TWO CATEGORIES ⇒ DUMMY VARIABLES
K-LEVEL QUALITATIVE VARIABLE IS REPRESENTED BY A VECTOR OF K BINARY VARIABLES (BITS), ONLY ONE OF WHICH IS "ON" AT A TIME.

INPUTS ... DENOTED BY X (CAN BE A VECTOR)
QUANTITATIVE OUTPUTS ... Y
QUALITATIVE OUTPUTS ... G (FOR GROUP)

OBSERVED VALUES ARE WRITTEN IN LOWERCASE
iTH OBSERVED VALUE OF X ... x_i

MATRICES ... $\mathbb{X}$

SET OF N INPUT p-VECTORS WOULD BE REPRESENTED BY THE N×P MATRIX $\mathbb{X}$.

VECTORS WILL NOT BE BOLD, EXCEPT WHEN THEY HAVE N COMPONENTS

x_i ... P-VECTOR OF INPUTS FOR THE iTH OBSERVATION

$\mathbf{x}_j$ ... N-VECTOR ... ALL OBSERVATIONS ON VARIABLE X_j

ALL VECTORS ARE ASSUMED TO BE COLUMN VECTORS

⇒ THE iTH ROW OF $\mathbb{X}$ IS x_i^T.

<u>LEARNING TASK</u>: GIVEN THE VALUE OF AN INPUT VECTOR X, MAKE A GOOD PREDICTION OF THE OUTPUT Y, DENOTED BY $\hat{Y}$ ("Y-HAT")

FOR CATEGORICAL OUTPUTS, $\hat{G}$ SHOULD TAKE VALUES IN THE SAME SET $\mathcal{G}$ ASSOCIATED WITH G.

<u>TRAINING DATA</u>: A SET OF MEASUREMENTS (x_i, y_i) OR (x_i, g_i), $i = 1, \dots, N$

TWO SIMPLE APPROACHES: THE LINEAR MODEL AND k-NEAREST NEIGHBORS

LINEAR MODEL	k-NEAREST NEIGHBORS
HUGE ASSUMPTIONS ABOUT STRUCTURE	VERY MILD STRUCTURAL ASSUMPTIONS
YIELDS STABLE, BUT POSSIBLY INACCURATE PREDICTIONS	PREDICTIONS ARE OFTEN ACCURATE BUT CAN BE UNSTABLE

18.01.2010 r.
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

LINEAR MODELS

INPUTS: $X^T = (X_1, \dots, X_p)$

WE PREDICT THE OUTPUT Y VIA THE MODEL:

$$\boxed{\hat{Y} = \underset{\uparrow \text{ INTERCEPT (BIAS)}}{\hat{\beta}_0} + \sum_{j=1}^{p} X_j \hat{\beta}_j}$$

OFTEN IT IS CONVENIENT TO INCLUDE THE CONSTANT VARIABLE 1 IN X

VECTOR OF COEFFICIENTS: $\hat{\beta}^T = (\beta_0, \dots, \beta_p)$

LINEAR MODEL IN VECTOR FORM: $\hat{Y} = X^T \hat{\beta}$

IN THE $(p+1)$-DIMENSIONAL INPUT-OUTPUT SPACE, $(X, \hat{Y})$ REPRESENTS A HYPERPLANE

IF THE CONSTANT IS INCLUDED IN X, THEN THE HYPERPLANE INCLUDES THE ORIGIN AND IS A SUBSPACE

IF NOT, IT IS AN AFFINE SET CUTTING THE Y-AXIS AT THE POINT $(0, \hat{\beta}_0)$.

WE ASSUME THAT THE INTERCEPT IS INCLUDED IN $\hat{\beta}$. !

METHOD OF LEAST SQUARES

$$\min_{\beta} \quad RSS(\beta) = \sum_{i=1}^{N} (Y_i - x_i^T \beta)^2 =$$

$$= (\mathbb{Y} - \mathbb{X}\beta)^T (\mathbb{Y} - \mathbb{X}\beta)$$

DIFFERENCIATING W.R.T. β → NORMAL EQUATIONS:

$$\mathbb{X}^T (\mathbb{Y} - \mathbb{X}\beta) = 0$$

IF $\mathbb{X}^T \mathbb{X}$ IS NONSINGULAR $\Rightarrow$ $\underline{\underline{\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}}}$

LINEAR MODEL IN A CLASSIFICATION CONTEXT

TRAINING DATA $((X_1, X_2)^T, G)$ $G \in \{BLUE, ORANGE\}$

RESPONSE Y CODED AS 0 FOR BLUE (1 FOR ORANGE)

$$\hat{G} = \begin{cases} ORANGE & IF\ \hat{Y} > 0.5 \\ BLUE & IF\ \hat{Y} \leq 0.5 \end{cases}$$

DECISION BOUNDARY $\{x \mid x^T \hat{\beta} = 0.5\}$

SCENARIO 1 : TRAINING DATA IN EACH CLASS WERE GENERATED FROM BIVARIATE GAUSSIAN DISTRIBUTIONS WITH UNCORRELATED COMPONENTS AND DIFFERENT MEANS

SCENARIO 2 : TRAINING DATA IN EACH CLASS CAME FROM A MIXTURE OF LOW-VARIANCE GAUSSIAN DISTR.

IN SCENARIO 1 - LINEAR DECISION BOUNDARY IS THE BEST ONE CAN DO, THE REGION OF OVERLAP IS INEVITABLE

IN SCENARIO 2 - THE OPTIMAL DECISION BOUNDARY IS NONLINEAR AND DISJOINT

NEAREST - NEIGHBORS METHODS

USE THOSE OBSERVATIONS IN THE TRAINING SET τ CLOSEST IN INPUT SPACE TO x TO FORM $\hat{Y}$.

THE k- NEAREST NEIGHBOR FIT FOR $\hat{Y}$ IS DEFINED AS:

$$\boxed{\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i}$$

$N_k(x)$ - THE NEIGHBORHOOD OF x DEFINED BY THE k CLOSEST POINTS x_i IN THE TRAINING SAMPLE

METRIC - WE ASSUME EUCLIDEAN DISTANCE

1- NEAREST NEIGHBOR CLASS. -> VORONOI TESSELLATION

EXAM: NAIL029 STROJOVÉ UČENÍ

THE EFFECTIVE NUMBER OF PARAMETERS OF k-NEAREST NEIGHBORS IS N/k.

WE CANNOT USE SUM-OF-SQUARED ERRORS ON THE TRAINING SET AS A CRITERION FOR PICKING k, SINCE WE WOULD ALWAYS PICK $k=1$!

DECISION BOUNDARY OF

LINEAR METHODS	NEAREST NEIGHBORS
LOW VARIANCE AND POTENTIALLY HIGH BIAS	HIGH VARIANCE LOW BIAS

MORE COMPLEX METHODS:

- KERNEL METHODS – USE WEIGHTS THAT DECREASE SMOOTHLY TO ZERO WITH DISTANCE FROM THE TARGET P.
- IN HIGH-DIMENSIONAL SPACES THE DISTANCE KERNELS ARE MODIFIED TO EMPHASIZE SOME VARIABLE MORE THAN OTHERS
- LOCAL REGRESSION – FITS LINEAR MODELS BY LOCALLY WEIGHTED LEAST SQUARES
- LINEAR MODELS FIT TO A BASIS EXPANSION OF THE ORIGINAL INPUTS
- PROJECTION PURSUIT AND NEURAL NETWORK MODELS CONSIST OF SUMS OF NON-LINEARLY TRANSFORMED LINEAR MODELS

STATISTICAL DECISION THEORY

$X \in \mathbb{R}^p$ ... RANDOM INPUT VECTOR

$Y \in \mathbb{R}$ ... RANDOM OUTPUT VARIABLE

$Pr(X,Y)$ ... JOINT DISTRIBUTION

WE SEEK : FUNCTION $f(X)$ FOR PREDICTING Y

LOSS FUNCTION $L(Y, f(X))$ FOR PENALIZING ERRORS IN PR..

SQUARED ERROR LOSS: $L_2(Y,f(X)) = (Y-f(X))^2$

EXPECTED (SQUARED) PREDICTION ERROR :

$$EPE(f) = E(Y-f(X))^2 = \int [y-f(x)]^2 \, Pr(dx, dy)$$

BY CONDITIONING ON X :

$$EPE(f) = E_X \, E_{Y|X} \left([Y-f(X)]^2 \mid X\right)$$

$\Rightarrow$ IT SUFFICES TO MINIMIZE EPE POINTWISE :

$$f(x) = \operatorname{argmin}_c \, E_{Y|X}\left([Y-c]^2 \mid X=x\right)$$

THE SOLUTION IS $\underline{f(x) = E}(Y \mid X=x)$

$\equiv$ THE CONDITIONAL EXPECTATION, THE REGRESSION FUNC.

NEAREST-NEIGHBOR METHODS ATTEMPT TO DIRECTLY IMPLEMENT THIS RECIPE USING THE TRAINING DATA

UNDER MILD REGULARITY CONDITIONS ON THE JOINT PROBABILITY DISTRIBUTION $Pr(X,Y)$, ONE CAN SHOW THAT AS $N, k \to \infty$, $k/N \to 0$, $\hat{f}(x) \to E(Y \mid X=x)$.

LINEAR REGRESSION - A MODEL BASED APPROACH

WE ASSUME $f(x) \approx x^T \beta$, FROM EPE BY DIFFERENT..

$$\beta = [E(XX^T)]^{-1} E(XY)$$

18.01.2010 Mo
PETER CERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

LINEAR MODEL ASSUMES: $f(x)$ IS WELL APPROXIMATED BY A GLOBALLY LINEAR FUNCTION

k-NEAREST NEIGHBORS ASSUMES: $f(x)$ IS WELL APPROXIMATED BY A LOCALLY CONSTANT FUNCTION

IF WE REPLACE THE L_2 LOSS FUNCTION WITH THE L_1: $E|Y-f(X)|$, WE GET:

$$\underline{\underline{\hat{f}(x)}} = \text{median}(Y \mid X=x)$$

MORE ROBUST, BUT L_1 CRITERIA HAVE DISCONTINUITIES IN THEIR DERIVATES.

LOSS FUNCTION FOR CATEGORICAL VARIABLE G:
AN ESTIMATE $\hat{G}$ WILL ASSUME VALUES IN $\mathcal{G}$
LOSS FNC. REPRESENTED BY A $K \times K$ MATRIX $\mathbb{L}$, WHERE $K = \text{card}(\mathcal{G})$, $\text{diag}(\mathbb{L}) = 0$, $\mathbb{L}$ IS NONNEGATIVE
$\mathbb{L}(k,l)$ IS THE PRICE PAID FOR CLASSIFYING AN OBSERVATION BELONGING TO CLASS $\mathcal{G}_k$ AS $\mathcal{G}_l$.

$$EPE = E[L(G, \hat{G}(X)] =$$
$$= E_X \sum_{k=1}^{K} L(\mathcal{G}_k, \hat{G}(X)) \cdot Pr(\mathcal{G}_k | X)$$

$\Rightarrow$ AGAIN IT SUFFICES TO MINIMIZE EPE POINTWISE:

$$\hat{G}(x) = \text{argmin}_{g \in \mathcal{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, g) \cdot Pr(\mathcal{G}_k | X=x)$$

WITH 0-1 LOSS FUNCTION THIS SIMPLIFIES TO

$$\hat{G}(x) = \text{argmin}_{g \in \mathcal{G}} [1 - Pr(g | X=x)], \text{ I.E.}$$
$$\hat{G}(x) = \mathcal{G}_k \text{ IF } Pr(\mathcal{G}_k | X=x) = \max_{g \in \mathcal{G}} Pr(g | X=x)$$

BAYES CLASSIFIER - WE CLASSIFY TO THE MOST PROBABLE CLASS, USING THE CONDITIONAL (DISCRETE) DISTRIBUTION $Pr(G|X)$.

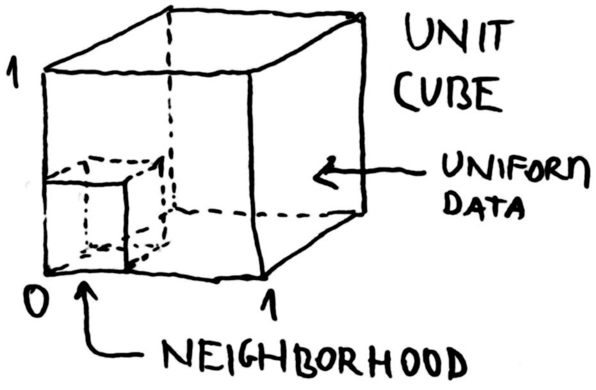
BAYES RATE - ERROR RATE OF THE BAYES CLASSIFIER

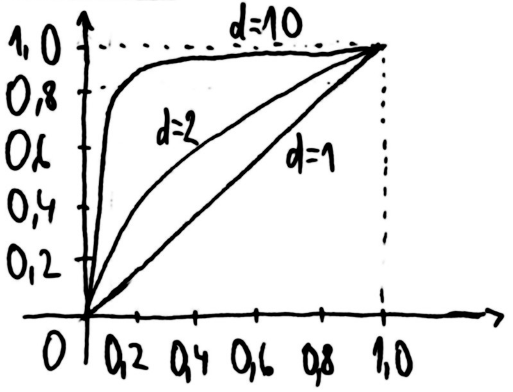
SUPPOSE FOR A TWO-CLASS PROBLEM WE HAD TAKEN THE DUMMY-VARIABLE APPROACH AND CODED G VIA A BINARY Y.

$$\hat{f}(X) = E(Y|X) = Pr(G = \mathcal{G}_1 | X)$$

... ANOTHER WAY OF REPRESENTING THE BAYES CLASSIFIER.

CURSE OF DIMENSIONALITY





IN TEN DIMENSIONS WE NEED TO COVER 80% OF THE RANGE OF EACH COORDINATE TO CAPTURE 10% OF THE DATA.

CONSIDER N DATA POINTS UNIFORMLY DISTRIBUTED IN A p- DIMENSIONAL UNIT BALL CENTERED AT ORIGIN

THE MEDIAN DISTANCE FROM THE ORIGIN TO THE CLOSEST DATA POINT IS (IN p- DIMENSIONS)

$$d(p, N) = \left(1 - \left(\frac{1}{2}\right)^{1/N}\right)^{1/p}$$

FOR $N = 500$, $p = 10$, $d(p, N) \approx 0.52$.

EXAM: NAIL029 STROJOVÉ UČENÍ

HENCE MOST DATA POINTS ARE CLOSER TO THE BOUNDARY OF THE SAMPLE SPACE THAN TO ANY OTHER DATA POINT.

SUPPOSE WE HAVE 1000 TRAINING SAMPLES x_i GENERATED UNIFORMLY ON $[-1, 1]^p$. ASSUME

$$Y = f(X) = e^{-8\|X\|^2}$$

WE USE 1-NEAREST-NEIGHBOR RULE TO PREDICT Y_0 AT $x_0 = 0$.

DENOTE THE TRAINING SET BY τ.

$$MSE(x_0) = E_\tau \left(f(x_0) - \hat{y}_0\right)^2 =$$

$$= E_\tau\left(\hat{y}_0 - E_\tau(\hat{y}_0)\right)^2 + \left(E_\tau(\hat{y}_0) - f(x_0)\right)^2 =$$

$$= \underline{\underline{Var_\tau(\hat{y}_0) + Bias^2(\hat{y}_0)}}$$

≡ BIAS-VARIANCE DECOMPOSITION

ON LARGE DIMENSIONS p THE MSE REACHES LEVELS NEAR 1.0, WHERE BIAS SIGNIFICANTLY DOMINATES

SUPPOSE THAT $Y = X^T\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

AND WE FIT THE MODEL BY LEAST SQUARES.

FOR AN ARBITRARY TEST POINT WE HAVE

$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + \sum_{i=1}^N l_i(x_0)\,\varepsilon_i,$$

WHERE $l_i(x_0)$ IS THE iTH ELEMENT OF $\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}x_0$,

⇒ LEAST SQUARE ESTIMATES ARE UNBIASED

$$EPE(x_0) = E_{y_0|x_0} E_\tau (y_0 - \hat{y}_0)^2 =$$

$$E_{y_0|x_0} E_\tau \left[x_0^T\beta + (y_0 - x_0^T\beta) - E_\tau\hat{y}_0 + E_\tau\hat{y}_0 - \hat{y}_0\right]^2 =$$

$$E_{y_0|x_0}\Big[\cancel{E_\tau}(x_0^T\beta - E_\tau\hat{y}_0)^2 + 2\cancel{E_\tau(x_0^T\beta - E_\tau\hat{y}_0)(E_\tau\hat{y}_0 - \hat{y}_0)} +$$

$$+ E_\tau(E_\tau\hat{y}_0 - \hat{y}_0)^2 + \cancel{E_\tau}(y_0 - x_0^T\beta)^2\Big] =$$

$$= (x_0^T\beta - E_\tau\hat{y}_0)^2 + E_\tau(E_\tau\hat{y}_0 - \hat{y}_0) + E_{y_0|x_0}(y_0 - x_0^T\beta)^2 =$$

$$= \mathrm{Bias}^2(\hat{y}_0) + \mathrm{Var}_\tau(\hat{y}_0) + \mathrm{Var}(y_0|x_0) =$$

$$= 0^2 + E_\tau x_0^T(\mathbb{X}^T\mathbb{X})^{-1}x_0\sigma^2 + \sigma^2$$

THERE IS NO BIAS, VARIANCE DEPENDS ON x_0

IF N IS LARGE, τ WERE SELECTED AT RANDOM, $E(X)=0$

$\Rightarrow E_{x_0}EPE(x_0) = \sigma^2(p/N) + \sigma^2$

STATISTICAL MODELS, SUPERVISED LEARNING AND FUNCTION APPROXIMATION

Θ ... SET OF PARAMETERS

LEAST SQUARES ... MINIMIZING THE RESIDUAL SUM OF SQ.

$RSS(\theta) = \sum_{i=1}^N (y_i - f_\theta(x_i))^2$

MAXIMUM LIKELIHOOD ESTIMATION

SUPPOSE WE HAVE RANDOM SAMPLE y_i, $i=1,\dots,N$ FROM A DENSITY $Pr_\theta(y)$

$L(\theta) = \sum_{i=1}^N \log Pr_\theta(y_i)$ – THE LOG-PROB. OF THE OBSERVED SAMPLE

THE MOST REASONABLE VALUES FOR Θ ARE THOSE FOR WHICH THE PROBABILITY OF THE OBSERVED SAMPLE IS LARGEST

EXAM: NAIL029 STROJOVÉ UČENÍ

LEAST SQUARES FOR THE ADDITIVE ERROR MODEL $Y = f_\theta(X) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ IS EQUIVALENT TO MAXIMUM LIKELIHOOD USING

$$Pr(Y \mid X, \theta) \sim N\left(f_\theta(X), \sigma^2\right)$$

THE LOG-LIKELIHOOD IS:

$$L(\theta) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{N}\left(y_i - f_\theta(x_i)\right)^2}_{= RSS(\theta)}$$

MULTINOMIAL LIKELIHOOD FOR THE REGRESSION FUNCTION $Pr(G \mid X)$ FOR A QUALITATIVE OUTPUT G

$$Pr(G = g_k \mid X = x) = p_{k,\theta}(x), \quad k = 1, \ldots, K.$$

LOG-LIKELIHOOD ($\equiv$ CROSS-ENTROPY):

$$L(\theta) = \sum_{i=1}^{N} \log p_{g_i,\theta}(x_i)$$

STRUCTURED REGRESSION MODELS

CONSIDER THE RSS CRITERION:

$$RSS(f) = \sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2$$

... MINIMIZING LEADS TO INFINITELY MANY SOLUTIONS. IN ORDER TO OBTAIN USEFUL RESULTS FOR FINITE N, WE MUST RESTRICT THE ELIGIBLE SOLUTIONS TO A SMALLER SET OF FUNCTIONS.

CLASSES OF RESTRICTED ESTIMATORS

① ROUGHNESS PENALTY & BAYESIAN METHODS

$$PRSS(f;\lambda) = RSS(f) + \lambda J(f)$$

↑ PENALIZED LEAST-SQUARES CRITERION

THE USER-SELECTED FUNCTIONAL $J(f)$ WILL BE LARGE FOR FUNCTIONS f THAT VARY TOO RAPIDLY OVER SMALL REGIONS OF INPUT SPACE.

CUBIC SMOOTHING SPLINE FOR 1-DIM. INPUTS:

$$PRSS(f;\lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$$

② KERNEL METHODS & LOCAL REGRESSION

KERNEL FUNCTION $K_\lambda(x_0, x)$ ASSIGNS WEIGHTS TO POINTS x IN A REGION AROUND x_0

GAUSSIAN KERNEL $K_\lambda(x_0,x) = \frac{1}{\lambda} \exp\left[- \frac{\|x-x_0\|^2}{2\lambda}\right]$

NADARAYA-WATSON WEIGHTED AVERAGE (KERNEL ESTIMATE):

$$\hat{f}(x_0) = \left(\sum_{i=1}^{N} K_\lambda(x_0, x_i)\, y_i\right) / \left(\sum_{i=1}^{N} K_\lambda(x_0, x_i)\right)$$

LOCAL REGRESSION ESTIMATE OF $f(x_0)$ AS $f_{\hat{\theta}}(x_0)$ WHERE $\hat{\theta}$ MINIMIZES

$$RSS(f_\theta, x_0) = \sum_{i=1}^{N} K_\lambda(x_0, x_i)\,(y_i - f_\theta(x_i))^2$$

$f_\theta(x) = \theta_0$ ... RESULTS IN THE NADARAYA-WATSON ESTIMAT

$f_\theta(x) = \theta_0 + \theta_1 x$ ... POPULAR LOCAL LINEAR REGRESSION MOD.

THE METRIC FOR k-NEAREST-NEIGHBORS

$$K_k(x, x_0) = I(\|x - x_0\| \leq \|x_{(k)} - x_0\|)$$

$x_{(k)}$ ... TRAINING OBSERVATION RANKED kTH IN DIST. FROM x_0

③ BASIS FUNCTIONS & DICTIONARY METHODS

THE MODEL FOR f IS A LINEAR EXPANSION OF BASIS F.

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x)$$

LINEAR SPLINES (FNC.): $b_1(x) = 1,\ b_2(x) = x,\ \dots,\ b_{m+2}(x) = (x - \overset{\overset{\text{KNOT}}{\downarrow}}{t_m})_+$

EXAM: NAIL029 STROJOVÉ UČENÍ

RADIAL BASIS FUNCTIONS ARE SYMMETRIC P-DIMENSIONAL KERNELS LOCATED AT PARTICULAR CENTROIDS.

$$f_\theta(x) = \sum_{m=1}^{M} K_{\lambda_m}(\mu_m, x)\, \theta_m$$

GAUSSIAN KERNEL $K_\lambda(\mu, x) = \frac{1}{\lambda} e^{-\|x-\mu\|^2/2\lambda}$

RADIAL BASIS FUNCTIONS HAVE CENTROIDS μ_m, SCALES λ_m
SPLINE BASIS FUNCTIONS HAVE KNOTS
... INCLUDING THESE AS PARAMETERS →
COMBINATORIALLY HARD NONLINEAR PROBLEM

MODEL SELECTION AND THE BIAS-VARIANCE TRADEOFF

SMOOTHING (COMPLEXITY) PARAMETER:

- THE MULTIPLIER OF THE PENALTY TERM
- THE WIDTH OF THE KERNEL
- THE NUMBER OF BASIS FUNCTIONS

CONSIDER k-NEAREST NEIGHBOR REGRESSION FIT $\hat{f}_k(x_0)$
AND MODEL $Y = f(X) + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$
FIX VALUES x_i IN THE SAMPLE

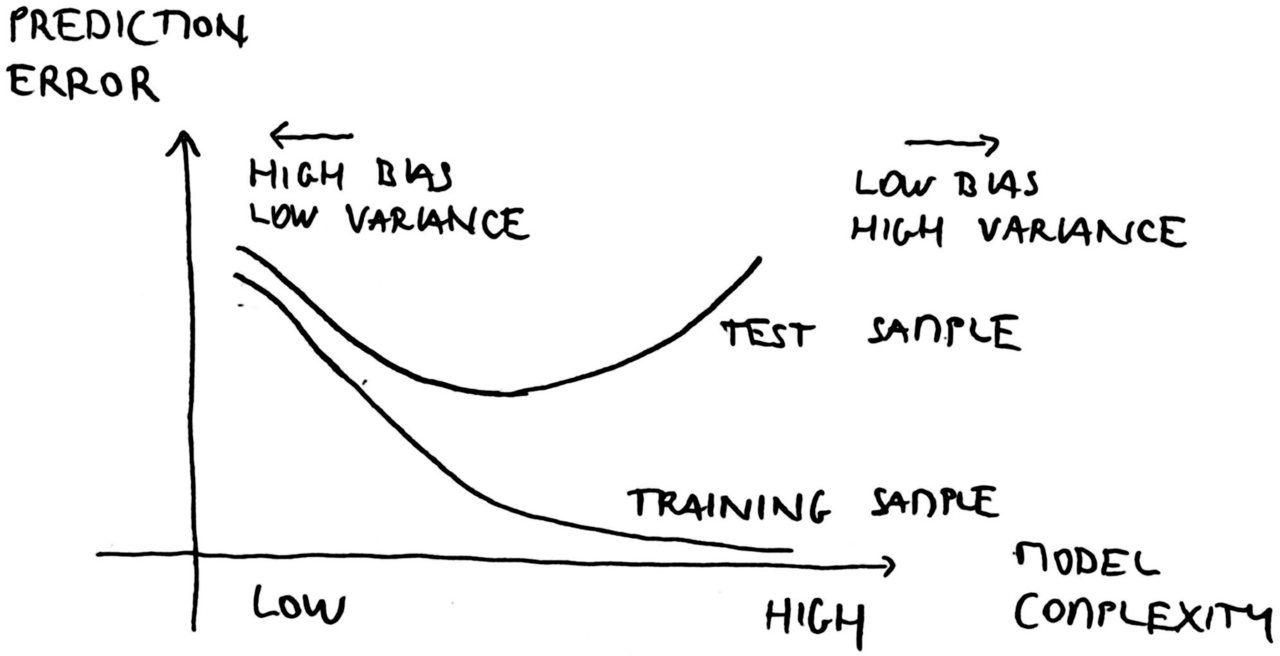
$$EPE_k(x_0) = E\left[(Y - \hat{f}_k(x_0))^2 \mid X = x_0\right] =$$

→ MSE OF $\hat{f}(x_0)$ IN ESTIMATING $f(x_0)$

$$= \sigma^2 + \boxed{Bias^2\left(\hat{f}_k(x_0)\right) + Var_\tau\left(\hat{f}_k(x_0)\right)} =$$

$$= \sigma^2 + \Big[f(x_0) - \underbrace{\frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})}_{E_\tau\left(\hat{f}_k(x_0)\right)}\Big]^2 + \frac{\sigma^2}{k}$$

AS k VARIES, THERE IS A BIAS-VARIANCE TRADEOFF


PREDICTION ERROR
HIGH BIAS
LOW VARIANCE
LOW BIAS
HIGH VARIANCE
TEST SAMPLE
TRAINING SAMPLE
LOW
HIGH
MODEL COMPLEXITY

19.01.2010 Tu
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

LINEAR METHODS OF REGRESSION

INPUT VECTOR $X^T = (X_1, \dots, X_p)$

WE ASSUME MODEL $f(X) = \beta_0 + \sum_{j=1}^{n} X_j \beta_j$

THE VARIABLES X_j CAN COME FROM DIFFERENT SOURCES
- QUANTITATIVE INPUTS
- TRANSFORMATIONS OF QUANTITATIVE INPUTS
- BASIS EXPANSIONS
- NUMERIC (DUMMY) CODING OF THE LEVELS OF QUAL. INP.
- INTERACTIONS BETWEEN VARIABLES

TRAINING DATA $(x_1, y_1), \dots, (x_N, y_N)$

$x_i = (x_{i1}, \dots, x_{ip})^T$ ... VECTOR OF FEATURE MEASUREMETS

ESTIMATION METHOD: LEAST SQUARES

$\mathbf{X}$ ... $N \times (p+1)$ MATRIX ... iTH ROW $= (1, x_i)$

$\mathbf{y}$ ... N-VECTOR OF OUTPUTS IN THE TRAINING SET

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \qquad \frac{\partial^2 RSS}{\partial \beta \, \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}$$

ASSUME ... $\mathbf{X}$ HAS FULL COLUMN RANK $\Rightarrow \mathbf{X}^T\mathbf{X} > 0$

$$\Rightarrow \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \Rightarrow \underline{\underline{\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}}} \mathbf{X}^T \mathbf{y}$$

PREDICTED VALUES AT INPUT VECTOR x_0:

$$\hat{f}(x_0) = (1, x_0)^T \hat{\beta}$$

FITTED VALUES AT THE TRAINING INPUTS:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \qquad \hat{y}_i = \hat{f}(x_i)$$

$\mathbb{H} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ ... HAT MATRIX

COLUMN VECTORS OF $\mathbb{X}$ SPAN A SUBSPACE OF $\mathbb{R}^N$
WE MINIMIZE $RSS(\beta) = \|y - \mathbb{X}\beta\|^2$ BY CHOOSING $\hat{\beta}$
SO THAT $y - \hat{y}$ IS ORTHOGONAL TO THIS SUBSPACE

ASSUME THAT THE OBSERVATIONS y_i ARE UNCORRELATED AND HAVE CONSTANT VARIANCE σ^2
FIX x_i.

$$\Rightarrow \quad Var(\hat{\beta}) = (\mathbb{X}^T\mathbb{X})^{-1}\sigma^2$$

σ^2 CAN BE ESTIMATED AS : $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^N (y_i - \hat{y}_i)^2$
$E(\hat{\sigma}^2) = \sigma^2$.

SUPPOSE $Y = \beta_0 + \sum_{j=1}^p X_j\beta_j + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

$$\Rightarrow \quad \hat{\beta} \sim N(\beta, (\mathbb{X}^T\mathbb{X})^{-1}\sigma^2)$$

LET US DENOTE
$\mathbb{V} = (\mathbb{X}^T\mathbb{X})^{-1}$

$$(N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$$

$\hat{\beta}, \hat{\sigma}^2$ ARE INDEPENDENT

$$z_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 v_{jj}}} \sim t_{N-p-1}$$

USED TO TEST HYPOTHES..
$H_0 : \beta_j = 0$

F STATISTICS : $F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1 / (N - p_1 - 1)} \sim F_{p_1 - p_0, N - p_1 - 1}$

RSS_1 ... RESIDUAL SUM OF SQUARE FOR THE LEAST SQUARES FIT OF THE BIGGER MODEL WITH p_1+1 PARAMETERS

RSS_0 ... —"— SMALLER MODEL, p_0+1 PARAMETERS

FOR LARGE N : $t_{N-p-1} \to N(0,1)$, $F_{p_1-p_0, N-p_1-1} \to \chi^2_{p_1-p_0}$

EXAM: NAIL029 STROJOVÉ UČENÍ

$1-2\alpha$ CONFIDENCE INTERVAL FOR β_j:

$$\left(\hat{\beta}_j - z^{(1-\alpha)} \sqrt{\hat{\sigma}^2 v_{jj}} \; , \; \hat{\beta}_j + z^{(1-\alpha)} \sqrt{\hat{\sigma}^2 v_{jj}} \right)$$

$z^{(1-\alpha)}$ ... $1-\alpha$ PERCENTIL OF t_{N-p-1} (OR $N(0,1)$)

$$z^{(1-0.025)} \approx 1.96$$
$$z^{(1-0.05)} \approx 1.645 \; \dots$$

APPROXIMATE CONFIDENCE SET FOR β:

$$C_\beta = \left\{ \beta \mid (\hat{\beta}-\beta)^T \mathbb{X}^T \mathbb{X} (\hat{\beta}-\beta) \le \hat{\sigma}^2 \chi^{2\,(1-\alpha)}_{p+1} \right\}$$

$\Rightarrow$ CONFIDENCE SET FOR THE TRUE FUNCTION: $x^T . C_\beta$

GAUSS - MARKOV THEOREM

THE LEAST SQUARE ESTIMATE OF $\Theta = a^T \beta$ IS

$\hat{\Theta} = a^T \hat{\beta} = a^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{y}$ (SUPPOSE $\mathbb{X}$ IS FIXED)

$a^T \hat{\beta}$ IS UNBIASED, SINCE

$$E(a^T \hat{\beta}) = a^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{X} \beta = a^T \beta$$

THE GAUSS-MARKOV THEOREM STATES THAT IF WE HAVE ANY OTHER LINEAR UNBIASED ESTIMATOR $\tilde{\Theta} = c^T \mathbb{y}$ FOR $a^T \beta$, THEN $Var(a^T \beta) \le Var(c^T \mathbb{y})$.

$$MSE(\tilde{\Theta}) = Var(\tilde{\Theta}) + [E(\tilde{\Theta}) - \Theta]^2$$

HOWEVER, THERE MAY WELL EXIST A BIASED ESTIMATOR WITH SMALLER MSE.

MULTIPLE REGRESSION FROM UNIVARIATE REGRESSION

UNIVARIATE $(p=1)$ LINEAR MODEL: $Y = X\beta + \varepsilon$

$\hat{\beta} = \frac{\langle \mathbb{x}, \mathbb{y} \rangle}{\langle \mathbb{x}, \mathbb{x} \rangle}$, $\mathbb{r} = \mathbb{y} - \mathbb{x}\hat{\beta}$

SUPPOSE THAT COLUMNS $\mathbb{x}_1, \dots, \mathbb{x}_p$ OF $\mathbb{X}$ ARE ORTHOGONAL

$\Rightarrow \hat{\beta}_j = \frac{\langle \mathbb{x}_j, \mathbb{y} \rangle}{\langle \mathbb{x}_j, \mathbb{x}_j \rangle}$

"REGRESS $\mathbb{b}$ ON $\mathbb{a}$" $\equiv$ COMPUTE $\hat{\gamma} = \frac{\langle \mathbb{a}, \mathbb{b} \rangle}{\langle \mathbb{a}, \mathbb{a} \rangle}$ AND RESIDUAL VEC.: $\mathbb{b} - \hat{\gamma}\mathbb{a}$

GRAM-SCHMIDT PROCEDURE:

```
1. $\mathbb{z}_0 = \mathbb{x}_0 = \mathbb{1}$
2. for $j = 1$ to $p$ do
   a) REGRESS $\mathbb{x}_j$ ON $\mathbb{z}_0, \dots, \mathbb{z}_{j-1}$ $\rightarrow \hat{\gamma}_{lj} = \frac{\langle \mathbb{z}_l, \mathbb{x}_j \rangle}{\langle \mathbb{z}_l, \mathbb{z}_l \rangle}$
   b) $\mathbb{z}_j \leftarrow \mathbb{x}_j - \sum_{l=0}^{j-1} \hat{\gamma}_{lj} \mathbb{z}_l$
```

RESULT: $\hat{\beta}_p = \frac{\langle \mathbb{z}_p, \mathbb{y} \rangle}{\langle \mathbb{z}_p, \mathbb{z}_p \rangle}$, $\mathrm{Var}(\hat{\beta}_p) = \frac{\sigma^2}{\langle \mathbb{z}_p, \mathbb{z}_p \rangle}$

$\mathbb{X} = \mathbb{Z}\Gamma$, $\mathbb{Z} = (\mathbb{z}_0, \dots, \mathbb{z}_p)$, $\Gamma = (\hat{\gamma}_{lj})$ UPPER TRIANG. MATRIX $\hat{\gamma}_{jj} = 1 \ \forall j$

SUPPOSE $\mathbb{D}$ DIAGONAL MATRIX $D_{jj} = \|\mathbb{z}_j\|$

$\mathbb{X} = (\mathbb{Z}\mathbb{D}^{-1})(\mathbb{D}\Gamma) = \mathbb{Q}\mathbb{R}$ ... QR-DECOMPOSITION

$\Rightarrow \hat{\beta} = \mathbb{R}^{-1}\mathbb{Q}^T\mathbb{y}$, $\hat{\mathbb{y}} = \mathbb{Q}\mathbb{Q}^T\mathbb{y}$

MULTIPLE OUTPUTS

$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{E}$, $\mathbb{Y}$ ... $N \times K$ RESPONSE MATRIX,
$\mathbb{B}$ ... $(p+1) \times K$ MATRIX OF PARAMETERS

$\mathrm{RSS}(\mathbb{B}) = \mathrm{tr}[(\mathbb{Y} - \mathbb{X}\mathbb{B})^T(\mathbb{Y} - \mathbb{X}\mathbb{B})$

LEAST SQUARE ESTIMATES: $\underline{\underline{\hat{\mathbb{B}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}}}$

19.01.2010 TU
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

SUBSET SELECTION

REASONS:

- PREDICTION ACCURACY: THE LEAST SQUARES OFTEN HAVE LOW BIAS BUT LARGE VARIANCE
- INTERPRETATION

1. BEST-SUBSET SELECTION

FINDS FOR EACH $k \in \{0, \dots, p\}$ THE SUBSET OF SIZE k THAT GIVES SMALLEST RESIDUAL SUM OF SQUARES.

ALGORITHM: LEAPS AND BOUNDS (FEASIBLE FOR $p \leq 40$)

THE QUESTION OF HOW TO CHOOSE k INVOLVES THE TRADEOFF BETWEEN BIAS AND VARIANCE.

THE AIC CRITERION IS A POPULAR ALTERNATIVE.

2. FORWARD- AND BACKWARD-STEPWISE SELECTION

FORWARD-STEPWISE SELECTION - STARTS WITH THE INTERCEPT, AND THEN SEQUENTIALLY ADDS INTO THE MODEL THE PREDICTOR THAT MOST IMPROVES THE FIT.

BACKWARD-STEPWISE SELECTION - STARTS WITH THE FULL MODEL, AND SEQUENTIALLY DELETES THE PREDICTOR THAT HAS THE LEAST IMPACT ON THE FIT. (I.E. THE VARIABLE WITH THE SMALLEST Z-SCORE)

SHRINKAGE METHODS

1. RIDGE REGRESSION - SHRINKS THE REGRESSION COEFFICIENTS BY IMPOSING A PENALTY ON THEIR SIZE

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

↑ NO INTERCEPT!

$\lambda \geq 0$ ... COMPLEXITY PARAMETER

INPUTS ARE NORMALLY STANDARDIZED

WE ASSUME CENTERING, I.E. $\mathbb{X}$ HAS p COLUMNS

$$RSS(\lambda) = (\mathbf{y} - \mathbb{X}\beta)^T (\mathbf{y} - \mathbb{X}\beta) + \lambda \beta^T \beta$$

$$\Rightarrow \hat{\beta}^{ridge} = (\mathbb{X}^T \mathbb{X} + \lambda \mathbf{I})^{-1} \mathbb{X}^T \mathbf{y}$$

2. THE LASSO (≡ BASIS PURSUIT)

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right]$$

SUBJECT TO $\sum_{j=1}^{p} |\beta_j| \leq t$

L_2 RIDGE PENALTY IS REPLACED WITH L_1 LASSO PENALTY

$\Rightarrow$ QUADRATIC PROGRAMMING PROBLEM

SHRINKAGE FACTOR $s = t / \sum_{j=1}^{p} |\hat{\beta}_j|$ ← LEAST SQUARES ESTIMATES (FULL MODEL)

4/7

19.01.2010 TU
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

LINEAR METHODS FOR CLASSIFICATION

DECISION BOUNDARIES ARE LINEAR

SUPPOSE THERE ARE K CLASSES (1, ..., K), AND THE FITTED LINEAR MODEL FOR THE kTH INDICATOR RESPONSE VARIABLE IS $\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$.

THE DECISION BOUNDARY BETWEEN CLASS k AND ℓ IS $\{x \mid \hat{f}_k(x) = \hat{f}_\ell(x)\} = \{x \mid (\hat{\beta}_{k0} - \hat{\beta}_{\ell 0}) + (\hat{\beta}_k - \hat{\beta}_\ell)x = 0\}$, I.E. AN AFFINE SET OR HYPERPLANE.

(a) DISCRIMINANT FUNCTIONS $\delta_k(x)$ FOR EACH CLASS CLASSIFY X TO THE CLASS WITH THE LARGEST VALUE FOR ITS DISCRIMINANT FUNCTION

(b) POSTERIOR PROBABILITIES $Pr(G=k \mid X=x)$

WE REQUIRE THAT SOME MONOTONE TRANSFORMATION OF δ_k OR $Pr(G=k \mid X=x)$ BE LINEAR.

FOR INSTANCE, FOR TWO CLASSES:

$$Pr(G=1 \mid X=x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}$$

$$Pr(G=2 \mid X=x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}$$

MONOTONE TRANSFORMATION: LOGIT TR. $\log[p/(1-p)]$

$$\log\left[\frac{Pr(G=1 \mid X=x)}{Pr(G=2 \mid X=x)}\right] = \beta_0 + \beta^T x$$

$\Rightarrow$ DECISION BOUNDARY $\{x \mid \beta_0 + \beta^T x = 0\}$.

LINEAR LOGITS METHODS:
- LINEAR DISCRIMINANT ANALYSIS
- LINEAR LOGISTIC REGRESSION

EXPLICIT MODELLING OF BOUNDARIES:
- PERCEPTRON (ROSENBLATT 1985, SEPARABLE TR. SET)
- OPTIMALLY SEPARATING HYPERPLANE (VAPNIK, 1996)

LINEAR REGRESSION OF AN INDICATOR MATRIX

EACH OF THE RESPONSE CATEGORIES ARE CODED VIA AN INDICATOR VARIABLE

$\mathcal{G}$ HAS K CLASSES, THERE ARE K INDICATORS Y_k

WITH $Y_{ik} = \begin{cases} 1 & G=k \\ 0 & G\neq k \end{cases}$

INDICATOR RESPONSE MATRIX $\mathbb{Y}$, $N\times K$ (EACH ROW HAS SINGLE 1)

WE FIT A LINEAR REGRESSION MODEL:

$$\hat{\mathbb{Y}} = \mathbb{X}\underbrace{(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}}_{\mathbb{B}\ \dots\ (p+1)\times K\ \text{COEFFICIENT MATRIX}}, \quad \mathbb{X}\ \dots\ N\times(p+1)\ \text{MODEL MATRIX}$$

CLASSIFICATION OF x:

1) COMPUTE THE FITTED OUTPUT $\hat{f}(x) = (1, x^T)\hat{\mathbb{B}}$
2) $\hat{G}(x) = \operatorname{argmax}_{k\in\mathcal{G}} \hat{f}_k(x)$

(ASSUME $t_k = e_k$)

A MORE SIMPLISTIC APPROACH ... CONSTRUCT TARGETS t_k

THE RESPONSE VECTOR y_i (ITH ROW OF $\mathbb{Y}$) FOR OBSERVATION i HAS THE VALUE $y_i = t_k$ IF $g_i = k$.

WE FIT THE LINEAR MODEL:

$$\min_{\mathbb{B}} \sum_{i=1}^{N} \| y_i - [(1, x_i^T)\mathbb{B}]^T \|^2$$

CLASSIFICATION: $\hat{G}(x) = \operatorname{argmin}_k \|\hat{f}(x) - t_k\|^2$.

PROBLEM: $K \geq 3 \Rightarrow$ CLASSES CAN BE MASKED BY OTHERS

EXAM : NAIL029 STROJOVÉ UČENÍ

LINEAR DISCRIMINANT ANALYSIS

SUPPOSE $f_k(x)$ IS THE CLASS-CONDITIONAL DENSITY OF X IN CLASS $G=k$

π_k ... PRIOR PROBABILITY OF CLASS k, $\sum_{k=1}^{K} \pi_k = 1$

BY THE BAYES THEOREM :

$$\Pr(G=k \mid X=x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

SUPPOSE THAT WE MODEL EACH CLASS DENSITY AS MULTIVARIATE GAUSSIAN :

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

LINEAR DISCRIMINANT ANALYSIS ARISES IN THE SPECIAL CASE WHEN $\Sigma_k = \Sigma \ \forall k$

$$\log \frac{\Pr(G=k \mid X=x)}{\Pr(G=l \mid X=x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} =$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l)$$

AN EQUATION LINEAR IN x.

LINEAR DISCRIMINANT FUNCTIONS :

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

ARE AN EQUIVALENT DESCRIPTION OF THE DECISION RULE, WITH $G(x) = \mathrm{argmax}_k \, \delta_k(x)$

IN PRACTICE WE DO NOT KNOW THE PARAMETERS OF THE GAUSSIAN DISTRIBUTIONS, AND WE WILL NEED TO ESTIMATE THEM USING OUR TRAINING DATA:

$\hat{\pi}_k = N_k / N$, N_k ... NUMBER OF CLASS-k OBSERVATIONS

$\hat{\mu}_k = \sum_{g_i = k} x_i / N_k$

$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$

LOGISTIC REGRESSION

... TO MODEL POSTERIOR PROBABILITIES OF THE K CLASSES VIA LINEAR FUNCTIONS IN x

$$\log \frac{Pr(G=1 \mid X=x)}{Pr(G=K \mid X=x)} = \beta_{10} + \beta_1^T x$$

...

$$\log \frac{Pr(G=K-1 \mid X=x)}{Pr(G=K \mid X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

A SIMPLE CALCULATION SHOWS THAT:

$$Pr(G=k \mid X=x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)} \qquad k=1,\dots,K-1$$

$$Pr(G=K \mid X=x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}$$

PARAMETER SET $\Theta = \{\beta_{10}, \beta_1^T, \dots, \beta_{(K-1)0}, \beta_{K-1}^T\}$

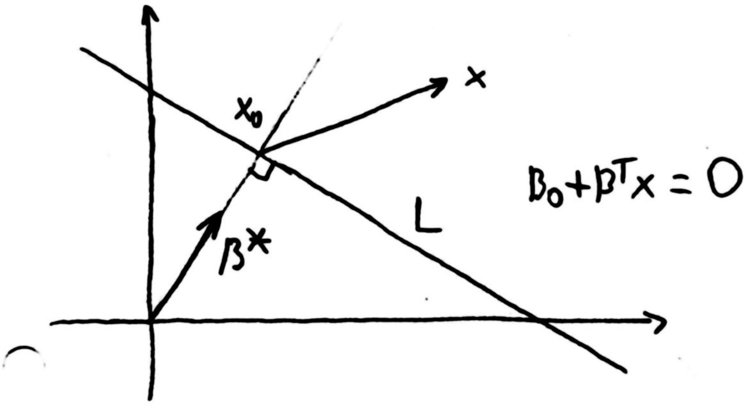
WE DENOTE $Pr(G=K \mid X=x) = p_k(x; \theta)$.

THE LOG-LIKELIHOOD FOR N OBSERVATIONS IS:

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)$$

EXAM: NAIL029 STROJOVÉ UCENÍ

SEPARATING HYPERPLANES



AFFINE SET L
DEFINED BY
$$f(x) = \beta_0 + \beta^T x = 0$$

1. FOR ANY TWO POINTS x_1 AND x_2 LYING IN L,
   $\beta^T(x_1 - x_2) = 0, \quad \Rightarrow \quad \beta^* = \beta / \|\beta\| \perp L$
2. FOR ANY POINT x_0 IN L : $\beta^T x_0 = -\beta_0$
3. THE SIGNED DISTANCE OF ANY POINT x TO L:
   $$\beta^{*T}(x - x_0) = \frac{1}{\|\beta\|} \cdot (\beta^T x + \beta_0) = \frac{1}{\|f'(x)\|} \cdot f(x)$$

$\Rightarrow$ $f(x)$ IS PROPORTIONAL TO THE SIGNED DISTANCE FROM x

ROSENBLATT'S PERCEPTRON LEARNING ALGORITHM

TRIES TO FIND A SEPARATING HYPERPLANE BY MINIMIZING THE DISTANCE OF MISCLASSIFIED POINTS TO THE DECISION BOUNDARY.

A RESPONSE $y_i = 1$ $(y_i = -1)$ IS MISCLASSIFIED IFF $x_i^T \beta + \beta_0 < 0$ (>0).

THE GOAL IS TO MINIMIZE
$$D(\beta, \beta_0) = -\sum_{i \in \mathcal{M}} y_i (x_i^T \beta + \beta_0)$$

$\mathcal{M}$ INDEXES THE SET OF MISCLASSIFIED POINTS.

THE GRADIENT (ASSUMING $\mathcal{M}$ IS FIXED) IS:

$$\frac{\partial D(\beta, \beta_0)}{\partial \beta} = -\sum_{i \in \mathcal{M}} y_i x_i \qquad \frac{\partial D(\beta, \beta_0)}{\partial \beta_0} = -\sum_{i \in \mathcal{M}} y_i$$

THE ALGORITHM USES STOCHASTIC GRADIENT DESCENT = RATHER THAN COMPUTING THE SUM OF THE GRADIENT CONTRIBUTIONS OF EACH OBSERVATION FOLLOWED BY A STEP IN THE NEGATIVE DIRECTION, A STEP IS TAKEN AFTER EACH OBSERVATION IS VISITED. MISCLASSIFIED OBSERVATIONS ARE VISITED IN SOME SEQUENCE, AND THE PARAMETERS β ARE UPDATED VIA

$$\begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} \leftarrow \begin{pmatrix} \beta \\ \beta_0 \end{pmatrix} + \rho \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix}, \quad \rho \text{ IS THE LEARNING RATE}$$

IF THE CLASSES ARE LINEARY SEPARABLE, IT CAN BE SHOWN THAT THE ALGORITHM CONVERGES TO A SEPARATING HYPERPLANE IN A FINITE NUMBER OF STEPS.

PROBLEMS:

- WHEN THE DATA ARE SEPARABLE, THERE ARE MANY SOLUTIONS
- THE "FINITE" NUMBER OF STEPS CAN BE VERY LARGE
- WHEN THE DATA ARE NOT SEPARABLE, THE ALGORITHM WILL NOT CONVERGE, AND CYCLES DEVELOP.

7/7

19.01.2010 Tu
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

OPTIMAL SEPARATING HYPERPLANES

SEPARATES THE TWO CLASSES AND MAXIMIZES THE DISTANCE TO THE CLOSEST POINT FROM EITHER CLASS (VAPNIK, 1996).

OPTIMIZATION PROBLEM:

$$\max_{\beta, \beta_0, \|\beta\|=1} M \qquad (*)$$

SUBJECT TO $y_i (x_i^T \beta + \beta_0) \geq M , \quad i=1,\dots,N$

THE SET OF CONDITIONS ENSURE THAT ALL THE POINTS ARE AT LEAST A SIGNED DISTANCE M FROM THE DECISION BOUNDARY DEFINED BY β AND β_0. WE CAN GET RID OF THE $\|\beta\|=1$ CONSTRAINT BY REPLACING THE CONDITIONS WITH:

$$\frac{1}{\|\beta\|} y_i \left(x_i^T \beta + \beta_0 \right) \geq M$$

(WHICH REDEFINES β_0), OR EQUIVALENTLY:

$$y_i \left(x_i^T \beta + \beta_0 \right) \geq M . \|\beta\|$$

SINCE FOR ANY β AND β_0 SATISFYING THESE INEQUALITIES, ANY POSITIVELY SCALED MULTIPLE SATISFIES THEM TOO, WE CAN SET $\|\beta\| = 1/M$. THUS $(*)$ IS EQUIVALENT TO:

$$\boxed{\begin{array}{l} \min_{\beta, \beta_0} \frac{1}{2} \|\beta\|^2 . \qquad (**) \\ \text{SUBJECT TO } y_i (x_i^T \beta + \beta_0) \geq 1 , \quad i=1,\dots,N \end{array}}$$

THE CONSTRAINTS (**) DEFINE AN EMPTY SLAB OR MARGIN AROUND THE LINEAR DECISION BOUNDARY OF THICKNESS $1/\|\beta\|$.

$\equiv$ CONVEX OPTIMIZATION PROBLEM

THE LAGRANGE FUNCTION TO BE MINIMIZED W.R.T β, β_0:

$$L_P = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i \left(x_i^T \beta + \beta_0 \right) - 1 \right]$$

SETTING THE DERIVATES TO ZERO, WE OBTAIN:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (1)$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i \qquad (2)$$

SUBSTITUING THESE INTO L_P WE OBTAIN SO-CALLED WOLFE DUAL:

$$\boxed{L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k} \, , \qquad \overset{(3):}{\alpha_i \geq 0}$$

THE SOLUTION IS OBTAINED BY MAXIMIZING L_D IN THE POSITIVE ORTHANT, A SIMPLER CONVEX OPTIMIZATION PROBLEM. IN ADDITION THE SOLUTION MUST SATISFY THE KARUSH-KUHN-TUCKER CONDITIONS:

(1), (2), (3), AND $\alpha_i \left[y_i \left(x_i^T \beta + \beta_0 \right) - 1 \right] = 0 \quad \forall i$.

FROM THESE WE CAN SEE THAT:

IF $\alpha_i > 0$ THEN $y_i (x_i^T \beta + \beta_0) = 1 \Rightarrow$
x_i IS ON THE BOUNDARY OF THE SLAB $\rightarrow$ SUPPORT POINT x_i

IF $y_i (x_i^T \beta + \beta_0) > 1$, x_i IS NOT ON THE BOUNDARY OF THE SLAB AND $\alpha_i = 0$

THE OPTIMAL SEPARATING HYPERPLANE PRODUCES A FUNCTION $\hat{f}(x) = x^T \hat{\beta} + \hat{\beta}_0$ FOR CLASSIFYING NEW OBSERVATIONS: $\hat{G}(x) = \text{sign}\, \hat{f}(x)$.

20.01.2010 We
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

BASIS EXPANSIONS AND REGULARIZATION

DENOTE $h_m(X) : \mathbb{R}^p \mapsto \mathbb{R}$ THE MTH TRANSFORMATION ON X, $m = 1, \dots, M$. THEN WE MODEL:

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

A LINEAR BASIS EXPANSION IN X.

WIDELY USED EXAMPLES OF THE h_m:

a) $h_m(X) = X_m$, $m = 1, \dots, p$ RECOVERS THE ORIGINAL MODEL.

b) $h_m(X) = X_j^2$, OR $X_j X_k$ ALLOWS US TO AUGUMENT THE INPUTS WITH POLYNOMIAL TERMS TO ACHIEVE HIGHER-ORDER TAYLOR EXPANSIONS.

c) $h_m(X) = \log(X_j), \sqrt{X_j}, \|X\|, \dots$ PERMITS OTHER NONLINEAR TRANSFORMATIONS

d) $h_m(X) = I(L_m \leq X_k \leq U_m)$ AN INDICATOR FOR A REGION OF X_k.

USEFUL FAMILIES: PIECEWISE-POLYNOMIALS, SPLINES, WAVELET BASES, ...

DICTIONARY $\mathcal{D}$ ... CONSISTING OF TYPICALLY A VERY LARGE NUMBER $|\mathcal{D}|$ OF BASIS FUNCTIONS

METHOD FOR CONTROLLING THE COMPLEXITY OF OUR MODEL:

(1) RESTRICTION METHODS - WE DECIDE BEFORE-HAND TO LIMIT THE CLASS OF FUNCTIONS

(2) SELECTION METHODS - WE ADAPTIVELY SCAN THE DICTIONARY AND INCLUDE ONLY SIGNIFICANT h_m

(3) REGULARIZATION METHODS - WE USE THE ENTIRE DICTIONARY, BUT RESTRICT THE COEFFICIENTS

PIECEWISE POLYNOMIALS

WE ASSUME THAT
X IS ONE-DIMENSIONAL

WE DIVIDE THE DOMAIN OF X INTO CONTINUOUS INTERVALS, AND REPRESENT f BY A SEPARATE POLYNOMIAL IN EACH INTERVAL.

a) PIECEWISE CONSTANT : ($\equiv$ ORDER-1 SPLINE)

$$h_1(X) = I(X < \xi_1),\ h_2(X) = I(\xi_1 \leq X < \xi_2), \dots,$$
$$h_{m-1}(X) = I(\xi_{m-2} \leq X < \xi_{m-1}),\ h_m(X) = I(\xi_{m-1} \leq X)$$

b) PIECEWISE LINEAR :

WE NEED TO ADD $h_{m+i}(X) = I(\dots)X$

c) PIECEWISE LINEAR, CONTINUOUS : ($\equiv$ ORDER-2 SPLINE)

$$h_1(X) = 1,\quad h_2(X) = X,\quad h_3(X) = (X - \xi_1)_+,\ h_4(X) = (X - \xi_2)_+, \dots$$

$t_+ \overset{def}{:=} \max(0, t)$, ξ_i ... KNOTS

d) CUBIC SPLINE - PIECEWISE-CUBIC POLYNOMIALS, CONTINUOUS, WITH CONTINUOUS FIRST AND SECOND DERIVATES AT THE KNOTS : ($\equiv$ ORDER-4 SPLINE)

$$h_1(X) = 1,\quad h_2(X) = X,\quad h_3(X) = X^2,\quad h_4(X) = X^3,$$
$$h_5(X) = (X - \xi_1)^3,\quad h_6(X) = (X - \xi_2)^3, \dots$$

e) ORDER M- SPLINE WITH KNOTS ξ_j, $j = 1, \dots, K$ - PIECEWISE-POLYNOMIAL OF ORDER M, WITH CONTINUOUS DERIVATES UP TO ORDER $M-2$:

$$h_j(X) = X^{j-1},\quad j = 1, \dots, M$$
$$h_{M+\ell}(X) = (X - \xi_\ell)_+^{M-1}\qquad \ell = 1, \dots, K$$

EXAM: NAIL029 STROJOVÉ UČENÍ

5) NATURAL CUBIC SPLINE - ADDS ADDITIONAL CONSTRAINTS - FUNCTION IS LINEAR BEYOND THE BOUNDARY KNOTS.

A NATURAL CUBIC SPLINE WITH K KNOTS IS REPRESENTED BY K BASIS FUNCTIONS.

$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{K-1}(X)$$

$$d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}$$

MULTIDIMENSIONAL SPLINES

SUPPOSE $X \in \mathbb{R}^2$ AND WE HAVE A BASIS OF FUNCTIONS $h_{1k}(X_1)$, $k = 1, \dots, M_1$, $h_{2k}(X_2)$, $k = 1, \dots, M_2$.

THEN THE $M_1 \times M_2$ DIMENSIONAL TENSOR PRODUCT BASIS DEFINED BY: $g_{jk}(X) = h_{1j}(X_1) \cdot h_{2k}(X_2)$

CAN BE USED FOR REPRESENTING A 2-DIM. FNC.

$$g(X) = \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \Theta_{jk} \, g_{jk}(X)$$

KERNEL SMOOTHING METHODS

A CLASS OF REGRESSION TECHNIQUES THAT ACHIEVE FLEXIBILITY IN ESTIMATING THE REGRESSION FUNCTION $f(X)$ OVER THE DOMAIN $\mathbb{R}^p$ BY FITTING A DIFFERENT BUT SIMPLE MODEL SEPARATELY AT EACH QUERY POINT x_0. THIS IS DONE BY USING ONLY THOSE OBSERVATIONS CLOSE TO THE TARGET POINT x_0 TO FIT THE SIMPLE MODEL.

KERNEL $K_\lambda(x_0, x_i)$ ... ASSIGNS A WEIGHT TO x_i BASED ON A DISTANCE FROM x_0

MEMORY-BASED METHODS
REQUIRE IN PRINCIPLE LITTLE OR NO TRAINING (ONLY λ)
THE MODEL $\equiv$ THE ENTIRE TRAINING DATA SET

ONE-DIMENSIONAL KERNEL SMOOTHERS

k-NEAREST-NEIGHBOR AVERAGE :

$$\hat{f}(x) = \text{Ave}\left(y_i \mid x_i \in N_k(x)\right)$$

THE AVERAGE CHANGES IN A DISCRETE WAY, LEADING TO A DISCONTINUOUS $\hat{f}(x)$

NADARAYA-WATSON KERNEL-WEIGHTED AVERAGE :

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i)\, y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

WITH EPANECHNIKOV QUADRATIC KERNEL :

$$K_\lambda(x_0, x) = D\left(\frac{|x - x_0|}{\lambda}\right), \quad D(t) = \begin{cases} \frac{3}{4}(1 - t^2) & |t| \leq 1 \\ 0 & \text{OTHERWISE} \end{cases}$$

EXAM: NAIL029 STROJOVÉ UČENÍ

ADAPTIVE NEIGHBORHOODS - FOR k-NEAREST NEIGHBORS WE HAVE $h_k(x_0) = |x_0 - x_{[k]}|$, $x_{[k]}$ ... kTH CLOSEST x_i TO x_0

ISSUES:

- THE SMOOTHING PARAMETER λ HAS TO BE DETERMINED
- METRIC WINDOW WIDTHS $h_\lambda(x)$
- OBSERVATION WEIGHTS
- BOUNDARY ISSUES - THE METRIC NEIGHBORHOODS TEND TO CONTAIN LESS POINTS ON THE BOUNDARIES, WHILE THE NEAREST-NEIGHBORHOODS GET WIDER

COMPACT SUPPORT FOR KERNELS:

EPANECHNIKOV KERNEL ... $D(t) = \frac{3}{4}(1-t^2) \quad |t| \leq 1$

TRI-CUBE KERNEL ... $D(t) = (1-|t|^3)^3 \quad |t| \leq 1$

NONCOMPACT KERNEL:

GAUSSIAN KERNEL ... $D(t) = \phi(t) = \frac{1}{2\pi} e^{-\frac{t^2}{2}}$

LOCAL LINEAR REGRESSION

LOCALLY WEIGHTED AVERAGES CAN BE BADLY BIASED ON THE BOUNDARIES OF THE DOMAIN, BECAUSE OF THE ASSYMMETRY OF THE KERNEL IN THAT REGION. BY FITTING STRAIGHT LINES RATHER THAN CONSTANTS LOCALLY, WE CAN REMOVE THIS BIAS EXACTLY TO FIRST ORDER.

LOCALLY WEIGHTED REGRESSION SOLVES A SEPARATED WEIGHTED LEAST SQUARES PROBLEM AT EACH TARGET POINT x_0:

$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \, [y_i - \alpha(x_0) - \beta(x_0) x_i]^2 .$$

THE ESTIMATE IS THEN: $\hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0) x_0$.

DEFINE $b(x)^T = (1, x)$, $\mathbb{B}$ BE THE REGRESSION MATRIX WITH ITH ROW $b(x_i)^T$, AND $\mathbb{W}(x_0)$ THE N×N DIAGONAL MATRIX, $W_{ii}(x_0) = K_\lambda(x_0, x_i)$.

$$\Rightarrow \hat{f}(x_0) = b(x_0)^T \left(\mathbb{B}^T \mathbb{W}(x_0) \mathbb{B}\right)^{-1} \mathbb{B}^T \mathbb{W}(x_0)\, y =$$

$$= \sum_{i=1}^{N} l_i(x_0)\, y_i$$

$l_i(x_0)$ ... SO-CALLED EQUIVALENT KERNEL

LOCAL POLYNOMIAL REGRESSION

$$\min_{\substack{\alpha(x_0),\, \beta_j(x_0)\\ j=1,\dots,d}} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \left[y_i - \alpha(x_0) - \sum_{j=1}^{d} \beta_j(x_0)\, x_i^j \right]$$

WITH ESTIMATE $\hat{f}(x_0) = \hat{\alpha}(x_0) + \sum_{j=1}^{d} \hat{\beta}_j(x_0)\, x_0^j$.

LOCAL LINEAR FITS TEND TO BE BIASED IN REGIONS OF CURVATURE OF THE TRUE FUNCTION, A PHENOMENON REFFERED TO AS TRIMMING THE HILLS AND FILLING THE VALLEYS.

LOCAL QUADRATIC REGRESSION IS GENERALLY ABLE TO CORRECT THIS BIAS

- LOCAL LINEAR FITS CAN HELP BIAS DRAMATICALLY AT THE BOUNDARIES AT A MODEST COST IN VARIANCE
- LOCAL QUADRATIC FITS TEND TO BE MOST HELPFUL IN REDUCING BIAS DUE TO CURVATURE IN THE INTERIOR OF THE DOMAIN
- ASYMPTOTIC ANALYSIS SUGEST THAT LOCAL POLYNOMIALS OF ODD DEGREE DOMINATE THOSE OF EVEN DEGREE

4/4

20.01.2010 We
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

SELECTING THE WIDTH OF THE KERNEL

- FOR EPANECHNIKOV OR TRI-CUBE KERNEL, λ IS THE RADIUS OF THE SUPPORT REGION
- FOR GAUSSIAN KERNEL, λ IS THE STANDARD DEVIATION
- λ IS THE NUMBER k OF NEAREST NEIGHBORS IN k-NN

BIAS-VARIANCE TRADEOFF

- IF THE WINDOW IS NARROW, $\hat{f}(x_0)$ IS AN AVERAGE OF A SMALL NUMBER OF y_i CLOSE TO x_0, AND ITS VARIANCE WILL BE RELATIVELY LARGE. THE BIAS WILL TEND TO BE SMALL.
- IF THE WINDOW IS WIDE, THE VARIANCE OF $\hat{f}(x_0)$ WILL BE SMALL, BECAUSE OF THE EFFECT OF AVERAGING. THE BIAS WILL BE LARGER.

LOCAL REGRESSION IN $\mathbb{R}^p$

LET $b(X)$ BE A VECTOR OF POLYNOMIAL TERMS IN X OF MAXIMUM DEGREE d. AT EACH $x_0 \in \mathbb{R}^p$ SOLVE

$$\min_{\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i)\left(y_i - b(x_i)^T \beta(x_0)\right)^2$$

TO PRODUCE THE FIT: $\hat{f}(x_0) = b(x_0)^T \hat{\beta}(x_0)$

$$K_\lambda(x_0, x) = D\left(\frac{\|x - x_0\|}{\lambda}\right)$$

FIRST WE STANDARDIZE EACH PREDICTOR.

LOCAL REGRESSION BECOMES LESS USEFUL IN DIMENSIONS MUCH HIGHER THAN TWO / THREE. IT IS IMPOSSIBLE TO SIMULTANEOUSLY MAINTAIN LOCALNESS ($\Rightarrow$ LOW BIAS) AND SIZEABLE SAMPLE IN THE NEIGHB. ($\Rightarrow$ LOW VAR.) AS THE DIM. p INCREASES, WITHOUT N INCREASING $\sim \exp(p)$.

STRUCTURED KERNELS

... WE USE A POSITIVE SEMIDEFINITE MATRIX A TO WEIGH DIFFERENT COORDINATES :

$$K_{\lambda,A}(x_0,x) = D\left(\frac{(x-x_0)^T A (x-x_0)}{\lambda}\right)$$

STRUCTURED REGRESSION FUNCTIONS

VARYING COEFFICIENT MODELS - WE DIVIDE THE p PREDICTORS IN X INTO A SET $(X_1, \dots, X_q)$, $q<p$, AND THE REMAINDER OF THE VARIABLES ... Z

WE THEN ASSUME THE CONDITIONALLY LINEAR MODEL :

$$f(X) = \alpha(Z) + \beta_1(Z) X_1 + \dots + \beta_q(Z) X_q$$

FOR GIVEN Z THIS IS A LINEAR MODEL.

21.01.2010 Th
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

MODEL ASSESSMENT AND SELECTION

THE GENERALIZATION PERFORMANCE OF A LEARNING METHOD RELATES TO ITS PREDICTION CAPABILITY ON INDEPENDENT TEST DATA.

BIAS, VARIANCE AND MODEL COMPLEXITY

WE HAVE A TARGET VARIABLE Y, A VECTOR OF INPUTS X, AND A PREDICTION MODEL $\hat{f}(X)$ THAT HAS BEEN ESTIMATED FROM A TRAINING SET $\mathcal{T}$. THE LOSS FUNCTION FOR MEASURING ERRORS BETWEEN Y AND $\hat{f}(X)$ ... $L(Y, \hat{f}(X))$

TYPICAL CHOICES: SQUARED ERROR, ABSOLUTE ERROR

TEST ERROR (GENERALIZATION ERROR) - IS THE PREDICTION ERROR OVER AN INDEPENDENT TEST SAMPLE $Err_{\mathcal{T}} = E[L(Y, \hat{f}(X)) \mid \mathcal{T}]$

WHERE BOTH X AND Y ARE DRAWN RANDOMLY FROM THEIR JOINT DISTRIBUTION (POPULATION), $\mathcal{T}$ IS FIXED.

EXPECTED PREDICTION ERROR (EXPECTED TEST ERROR):

$$Err = E[L(Y, \hat{f}(X))] = E[Err_{\mathcal{T}}]$$

TRAINING ERROR IS THE AVERAGE LOSS OVER THE TRAINING SAMPLE:

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

TRAINING ERROR IS NOT A GOOD ESTIMATE OF THE TEST ERROR. TRAINING ERROR CONSISTENTLY DECREASES WITH MODEL COMPLEXITY.

SUPPOSE A QUALITATIVE OR CATEGORICAL RESPONSE G TAKING ONE OF K VALUES IN A SET $\mathcal{G}$, LABELED $1, \dots, K$. TYPICALLY WE MODEL THE PROBABILITIES $p_k(X) = Pr(G = k \mid X)$ (OR SOME MONOTONE TRANSFORMATIONS $f_k(X)$), AND THEN:

$\hat{G}(X) = \text{argmax}_k \; \hat{p}_k(X)$.

TYPICAL LOSS FUNCTIONS:

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X)) \quad (\text{0-1 LOSS})$$

$$L(G, \hat{G}(X)) = -2 \sum_{k=1}^{K} I(G=k) \log \hat{p}_k(X) =$$

$$= -2 \log \hat{p}_G(X) \quad (-2\text{x LOG-LIKELIHOOD})$$

TEST ERROR $Err_\tau = E(L(G, \hat{G}(X)) \mid \tau)$

Err ... THE EXPECTED MISCLASSIFICATION ERROR

TRAINING ERROR, FOR EXAMPLE: $\overline{err} = -\frac{2}{N} \sum_{i=1}^{N} \log \hat{p}_{g_i}(x_i)$

MODEL SELECTION: ESTIMATING THE PERFORMANCE OF DIFFERENT MODELS IN ORDER TO CHOOSE THE BEST ONE

MODEL ASSESSMENT: HAVING CHOOSEN A FINAL MODEL, ESTIMATING ITS PREDICTION ERROR ON NEW DATA

2/6

21.01.2010 Th
PETER ČERNO

EXAM: NAIL 029 STROJOVÉ UČENÍ

DATA-RICH SITUATION → RANDOMLY DIVIDE THE DATASET INTO THREE PARTS : A TRAINING SET, A VALIDATION SET, AND A TEST SET

TRAINING SET - USED TO FIT THE MODELS

VALIDATION SET - USED TO ESTIMATE PREDICTION ERROR FOR MODEL SELECTION

TEST SET - USED FOR ASSESSMENT OF THE GENERALIZATION ERROR OF THE FINAL CHOSEN MODEL

WE CAN APPROXIMATE THE VALIDATION STEP :
- ANALYTICALLY (AIC, BIC, MDL, SRM)
- BY EFFICIENT SAMPLE RE-USE (CROSS-VALIDATION, BOOTSTRAP)

THE BIAS-VARIANCE DECOMPOSITION

WE ASSUME $Y = f(X) + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma_\varepsilon^2$

EXPRESSION FOR THE EXPECTED PREDICTION ERROR OF A REGRESSION FIT $\hat{f}(X)$ AT AN INPUT POINT x_0, USING SQUARED-ERROR LOSS :

$$Err(x_0) = E\left[(Y - \hat{f}(x_0))^2 \mid X = x_0\right] =$$
$$= \sigma_\varepsilon^2 + \left[E\hat{f}(x_0) - f(x_0)\right]^2 + E\left[\hat{f}(x_0) - E\hat{f}(x_0)\right]^2 =$$
$$= \sigma_\varepsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0)) =$$
$$= \text{IRREDUCIBLE ERROR} + \text{BIAS}^2 + \text{VARIANCE}$$

ASSUME THAT TRAINING INPUTS x_i ARE FIXED, AND THE RANDOMNESS ARISES FROM THE y_i

FOR THE k-NEAREST-NEIGHBOR REGRESSION FIT:

$$\mathrm{Err}(x_0) = E\left[(Y - \hat{f}_k(x_0))^2 \mid X = x_0\right] =$$

$$= \sigma_\varepsilon^2 + \left[f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right] + \frac{\sigma_\varepsilon^2}{k}$$

FOR A LINEAR MODEL FIT $\hat{f}_p(x) = x^T\hat{\beta}$:

$$\mathrm{Err}(x_0) = E\left[(Y - \hat{f}_p(x_0))^2 \mid X = x_0\right] =$$

$$= \sigma_\varepsilon^2 + \left[f(x_0) - E\hat{f}_p(x_0)\right]^2 + \|h(x_0)\|^2\sigma_\varepsilon^2$$

$$\hat{f}_p(x_0) = x_0^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y \Rightarrow h(x_0) = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}x_0$$

AND $\mathrm{Var}[\hat{f}_p(x_0)] = \|h(x_0)\|^2\sigma_\varepsilon^2$

$$\Rightarrow \frac{1}{N}\sum_{l=1}^{N}\mathrm{Err}(x_i) = \sigma_\varepsilon^2 + \frac{1}{N}\sum_{l=1}^{N}\left[f(x_i) - E\hat{f}(x_i)\right]^2 + \frac{p}{N}\sigma_\varepsilon^2$$

NOTE: BIAS-VARIANCE TRADEOFF BEHAVES DIFFERENTLY FOR 0-1 LOSS THAN IT DOES FOR SQUARED ERROR LOSS.

OPTIMISM OF THE TRAINING ERROR RATE

GIVEN A TRAINING SET $\mathcal{T} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ THE GENERALIZATION ERROR OF A MODEL $\hat{f}$ IS

$$\mathrm{Err}_{\mathcal{T}} = E_{X^0, Y^0}\left[L(Y^0, \hat{f}(X^0)) \mid \mathcal{T}\right]$$

$\mathcal{T}$ IS FIXED, THE POINT (X^0, Y^0) IS A NEW TEST DATA POINT, DRAWN FROM F, THE JOINT DISTRIBUTION OF THE DATA.

3/6

21.01.2010 Th
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

AVERAGING OVER TRAINING SETS YIELDS THE EXPECTED ERROR

$$Err = E_{\tau} E_{X^0, Y^0} [L(Y^0, \hat{f}(X^0)) | \tau]$$

MOST METHODS EFFECTIVELY ESTIMATE THE EXPECTED ERROR RATHER THAN $E\tau$.

TYPICALLY, THE TRAINING ERROR

$$\overline{err} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

WILL BE LESS THAN THE TRUE ERROR Err_{τ}.

IN-SAMPLE ERROR $Err_{in} = \frac{1}{N} \sum_{i=1}^{N} E_{Y^0} [L(Y_i^0, \hat{f}(x_i)) | \tau]$

OPTIMISM $op = Err_{in} - \overline{err}$.

AVERAGE OPTIMISM $\omega = E_{y}(op)$

THE PREDICTORS IN THE TRAINING SET ARE FIXED, AND THE EXPECTATION IS OVER TR. SET OUTCOME VALUES.

AN OBVIOUS WAY TO ESTIMATE PREDICTION ERROR IS TO ESTIMATE THE OPTIMISM AND THEN ADD IT TO THE TRAINING ERROR $\overline{err}$. (C_p, AIC, BIC, ..)

CROSS-VALIDATION, BOOTSTRAP METHODS ARE DIRECT ESTIMATES OF THE EXTRA SAMPLE ERROR Err.

IN-SAMPLE ERROR IS NOT USUALLY OF DIRECT INTEREST SINCE FUTURE VALUES ARE NOT LIKELY TO COINCIDE WITH THEIR TRAINING SET VALUES.
BUT IN-SAMPLE ERROR IS CONVENIENT FOR COMPARISON BETWEEN MODELS AND OFTEN LEADS TO EFFECTIVE MODEL SELECTION.

ESTIMATES OF IN-SAMPLE PREDICTION ERROR

AKAIKE INFORMATION CRITERION (AIC)

GIVEN A SET OF MODELS $f_\alpha(x)$ INDEXED BY A TUNING PARAMETER α, DENOTE BY $\overline{err}(\alpha)$, $d(\alpha)$ THE TRAINING ERROR, NUMBER OF PARAMETERS FOR EACH MODEL.

$$\boxed{AIC(\alpha) = \overline{err}(\alpha) + 2 \cdot \frac{d(\alpha)}{N} \cdot \hat{\sigma}_\varepsilon^2}$$

THE FUNCTION $AIC(\alpha)$ PROVIDES AN ESTIMATE OF THE TEST ERROR CURVE, AND WE FIND THE TUNNING PARAMETER $\hat{\alpha}$ THAT MINIMIZES IT.

THE BAYESIAN APPROACH AND BIC

THE BAYESIAN INFORMATION CRITERION, LIKE AIC, IS APPLICABLE IN SETTINGS WHERE THE FITTING IS CARRIED OUT BY MAXIMIZATION OF A LOG-LIKELIHOOD.

$$\boxed{BIC(\alpha) = \frac{N}{\hat{\sigma}_\varepsilon^2}\left[\overline{err}(\alpha) + (\log N) \cdot \frac{d(\alpha)}{N}\hat{\sigma}_\varepsilon^2\right]}$$

FOR MODEL SELECTION PURPOSES, THERE IS NO CLEAR CHOICE BETWEEN AIC AND BIC.

BIC IS ASSYMPTOTICALLY CONSISTENT AS A SELECTION CRITERION - GIVEN A FAMILY OF MODELS, INCLUDING THE TRUE MODEL, THE PROBABILITY THAT BIC WILL SELECT THE CORRECT MODEL APPROACHES ONE AS THE SAMPLE SIZE $N \to \infty$. THIS IS NOT THE CASE FOR AIC, WHICH TENDS TO CHOOSE MODELS WHICH ARE TOO COMPLEX AS $N \to \infty$.

4/6

21.01.2010Th
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

CROSS-VALIDATION

PROBABLY THE SIMPLEST AND MOST WIDELY USED METHOD FOR ESTIMATING PREDICTION ERROR. THIS METHOD DIRECTLY ESTIMATES THE EXPECTED EXTRA-SAMPLE ERROR $Err = E[L(Y, \hat{f}(X))]$.

K-FOLD CROSS-VALIDATION

WE SPLIT THE DATA INTO K ROUGHLY EQUAL-SIZED PARTS. FOR THE kTH PART, WE FIT THE MODEL TO THE OTHER K-1 PARTS OF THE DATA, AND CALCULATE THE PREDICTION ERROR OF THE FITTED MODEL WHEN PREDICTING THE kTH PART. WE DO THIS FOR $k = 1, \dots, K$ AND COMBINE THE K ESTIMATES OF PREDICTION ERROR.

LET $\kappa: \{1, \dots, N\} \to \{1, \dots, K\}$ BE AN INDEXING FUNCTION, $\kappa(i)$ = PARTITION OF iTH OBSERVATION

$\hat{f}^{-k}(x)$ ... FITTED FUNCTION WITHOUT kTH PARTITION

CROSS-VALIDATION ESTIMATE OF PREDICTION ERROR:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right)$$

THE CASE $K = N$ IS KNOWN AS LEAVE-ONE-OUT CROSS-VALIDATION, $\kappa(i) = i$.

GIVEN A SET OF MODELS f_α, α - TUNING PARAMETER, WE CHOOSE $\hat{\alpha} = \arg\min_{\alpha} CV(f_\alpha)$.

BOOTSTRAP METHODS

THE BOOTSTRAP IS A GENERAL TOOL FOR ASSESSING STATISTICAL ACCURACY.

LET US DENOTE THE TRAINING SET BY $\mathbb{Z} = (z_1, \dots, z_N)$ WHERE $z_i = (x_i, y_i)$. THE BASIC IDEA IS TO RANDOMLY DRAW DATASETS WITH REPLACEMENT FROM THE TRAINING DATA, EACH SAMPLE (=DATASET) THE SAME SIZE AS THE ORIGINAL TRAINING SET.

THIS IS DONE B TIMES → B BOOTSTRAP DATASETS.

THEN WE FIT THE BODEL TO EACH OF THE BOOTSTRAP DATASETS.

LET $S(\mathbb{Z})$ BE ANY QUANTITY COMPUTED FROM THE DATA $\mathbb{Z}$ (FOR EXAMPLE, THE PREDICTION AT SOME INPUT POINT). FROM THE BOOTSTRAP SAMPLING WE CAN ESTIMATE ANY ASPECT OF THE DISTRIBUTION OF $S(\mathbb{Z})$. FOR INSTANCE,

$$\widehat{\mathrm{Var}}(S(\mathbb{Z})) = \frac{1}{B-1} \sum_{b=1}^{B} \left(S(\mathbb{Z}^{*b}) - \bar{S}^{*} \right)^2,$$

WHERE $\bar{S}^{*} = \sum_{b=1}^{B} S(\mathbb{Z}^{*b}) / B$.

HOW TO ESTIMATE PREDICTION ERROR?

$$\widehat{\mathrm{Err}}_{boot} = \xcancel{\frac{1}{B} \frac{1}{N} \sum_{b=1}^{B} \sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))}$$

$\widehat{\mathrm{Err}}_{boot}$ DOES NOT PROVIDE A GOOD ESTIMATE: BOOTSTRAP DATASETS ARE ACTING AS THE TRAINING SAMPLES, WHILE THE ORIGINAL TRAINING SET IS ACTING AS THE TEST SAMPLE, AND THESE TWO SAMPLES HAVE OBSERVATIONS IN COMMON.

EXAM: NAIL 029 STROJOVÉ UČENÍ

Pr (OBSERVATION $i \in$ BOOTSTRAP SAMPLE b) =

$$= 1 - \left(1 - \frac{1}{N}\right)^N \approx 1 - e^{-1} = 0.632$$

BY MIMICKING CROSS-VALIDATION, A BETTER BOOTSTRAP ESTIMATE CAN BE OBTAINED. FOR EACH OBSERVATION, WE ONLY KEEP TRACK OF PREDICTIONS FROM BOOTSTRAP SAMPLES NOT CONTAINING THAT OBS. LET C^{-i} BE A SET OF INDICES OF BOOTSTRAP SAMPLES b THAT DO NOT CONTAIN OBSERVATION i. THE LEAVE-ONE-OUT BOOTSTRAP ESTIMATE OF PREDICTION ERROR IS:

$$\widehat{Err}^{(1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|C^{-1}|} \cdot \sum_{b \in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

WE LEAVE OUT THE TERMS CORRESPONDING TO $|C^{-1}| = 0$

THE .632 ESTIMATOR: $\widehat{Err}^{(.632)} = .368\,\overline{err} + .632\,\widehat{Err}^{(1)}$.

FOR MANY ADAPTIVE, NONLINEAR TECHNIQUES (LIKE TREES), ESTIMATION OF THE EFFECTIVE NUMBER OF PARAMETERS IS VERY DIFFICULT. THIS MAKES METHODS LIKE AIC IMPRACTICAL AND LEAVES US WITH CROSS-VALIDATION OR BOOTSTRAP AS THE METHODS OF CHOICE.

MODEL INFERENCE AND AVERAGING

THE BOOTSTRAP METHOD DESCRIBED ABOVE, IN WHICH WE SAMPLE WITH REPLACEMENT FROM THE TRAINING DATA, IS CALLED THE NONPARAMETRIC BOOTSTRAP.

THIS REALLY MEANS THAT THE METHOD IS "MODEL-FREE", SINCE IT USES THE RAW DATA.

CONSIDER A VARIATION OF THE BOOTSTRAP, CALLED THE PARAMETRIC BOOTSTRAP, IN WHICH WE SIMULATE NEW RESPONSES BY ADDING GAUSSIAN NOISE TO THE PREDICTED VALUES:

$$y_i^* = \hat{\mu}(x_i) + \varepsilon_i^*, \quad \varepsilon_i^* \sim N(0, \hat{\sigma}^2); \quad i=1,\dots,N$$

THIS PROCESS IS REPEATED B TIMES. THE RESULTING DATASETS HAVE THE FORM $(x_1, y_1^*), \dots, (x_N, y_N^*)$.

IN GENERAL, THE PARAMETRIC BOOTSTRAP AGREES WITH MAXIMUM LIKELIHOOD.

MAXIMUM LIKELIHOOD INFERENCE

FIRST WE SPECIFY A PROBABILITY DENSITY OR PROBABILITY MASS FUNCTION FOR OUR OBSERVATIONS

$$z_i \sim g_\theta(z)$$

θ ... ONE OR MORE UNKNOWN PARAMETERS

A SO-CALLED PARAMETRIC MODEL FOR Z

EXAMPLE: FOR $Z \sim N(\mu, \sigma^2)$ WE HAVE $\theta = (\mu, \sigma^2)$

AND $g_\theta(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(z-\mu)^2/\sigma^2}$

6/6

21.01.2010 Th
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

LIKELIHOOD FUNCTION : $L(\theta, \mathbb{Z}) = \prod_{i=1}^{N} g_\theta(z_i)$.

WE THINK OF $L(\theta, \mathbb{Z})$ AS A FUNCTION OF θ, WITH OUR DATA $\mathbb{Z}$ FIXED.

DENOTE THE LOGARITHM OF $L(\theta, \mathbb{Z})$ AS

$$\ell(\theta, \mathbb{Z}) = \sum_{i=1}^{N} \ell(\theta, z_i) = \sum_{i=1}^{N} \log g_\theta(z_i)$$

A SO-CALLED LOG-LIKELIHOOD.

THE METHOD OF MAXIMUM LIKELIHOOD CHOOSES THE VALUE $\hat{\theta}$ WHICH MAXIMIZES $\ell(\theta; \mathbb{Z})$

SCORE FUNCTION : $\dot{\ell}(\theta, \mathbb{Z}) = \sum_{i=1}^{N} \dot{\ell}(\theta, z_i)$

WHERE $\dot{\ell}(\theta, z_i) = \partial \ell(\theta, z_i) / \partial\theta$

WE ASSUME THAT THE LIKELIHOOD FUNCTION TAKES ITS MAXIMUM IN THE INTERIOR OF THE PARAMETER SPACE, I.E. $\dot{\ell}(\hat{\theta}; \mathbb{Z}) = 0$.

THE INFORMATION MATRIX :

$$\mathbf{I}(\theta) = - \sum_{i=1}^{N} \frac{\partial^2 \ell(\theta, z_i)}{\partial\theta \, \partial\theta^T}$$

FISHER INFORMATION (EXPECTED INFORMATION) :

$$i(\theta) = E_\theta[\mathbf{I}(\theta)]$$

LET θ_0 DENOTE THE TRUE VALUE OF θ.

THE SAMPLING DISTRIBUTION OF THE MAXIMUM LIKELIHOOD ESTIMATOR HAS A LIMITING NORMAL DIST.

$$\hat{\theta} \to N(\theta_0, i(\theta_0)^{-1}) \quad \text{AS} \quad N \to \infty$$

THE SAMPLING DIST OF $\hat{\theta}$ CAN BE APPROXIMATED BY

$$N(\hat{\theta}, i(\hat{\theta})^{-1}) \quad \text{OR} \quad N(\hat{\theta}, \mathbf{I}(\hat{\theta})^{-1}) .$$

BAYESIAN METHODS

IN THE BAYESIAN APPROACH TO INFERENCE, WE SPECIFY A SAMPLING MODEL $Pr(\mathbb{Z}|\theta)$ FOR OUR DATA GIVEN THE PARAMETERS AND A PRIOR DISTRIBUTION FOR THE PARAMETERS $Pr(\theta)$ REFLECTING OUR KNOWLEDGE ABOUT θ BEFORE WE SEE THE DATA. WE THEN COMPUTE THE POSTERIOR DISTRIBUTION:

$$Pr(\theta|\mathbb{Z}) = \frac{Pr(\mathbb{Z}|\theta) . Pr(\theta)}{\int Pr(\mathbb{Z}|\theta). Pr(\theta) d\theta}$$

WHICH REPRESENTS OUR UPDATED KNOWLEDGE ABOUT θ AFTER WE SEE THE DATA.

THE POSTERIOR DISTRIBUTION ALSO PROVIDES THE BASIS FOR PREDICTING THE VALUES OF FUTURE OBSERVATION z^{new}, VIA THE PREDICTIVE DISTRIBUTION:

$$Pr(z^{new}|\mathbb{Z}) = \int Pr(z^{new}|\theta) \cdot Pr(\theta|\mathbb{Z}) d\theta$$

IN CONTRAST, MAXIMUM LIKELIHOOD APPROACH WOULD USE $Pr(z^{new}|\hat{\theta})$, THE DATA DENSITY EVALUATED AT THE MAXIMUM LIKELIHOOD ESTIMATE, TO PREDICT FUTURE DATA.

IN GAUSSIAN MODELS, MAXIMUM LIKELIHOOD AND PARAMETRIC BOOTSTRAP ANALYSES TEND TO AGREE WITH BAYESIAN ANALYSES THAT USE A NONINFORMATIVE PRIOR FOR THE FREE PARAMETERS. THIS CORRESPONDENCE ALSO EXTENDS TO THE NONPARAMETRIC CASE, WHERE THE NONPARAMETRIC BOOTSTRAP APPROXIMATES A NONINFORMATIVE BAYES ANALYSIS.

22.01.2010 Fr
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

THE EM ALGORITHM

... IS A POPULAR TOOL FOR SIMPLIFYING DIFFICULT MAXIMUM LIKELIHOOD PROBLEMS.

SUPPOSE Y IS A MIXTURE OF TWO NORMAL DISTRIBUTIONS:

$$Y_1 \sim N(\mu_1, \sigma_1^2), \quad Y_2 \sim N(\mu_2, \sigma_2^2)$$

$$Y = (1-\Delta)Y_1 + \Delta Y_2, \quad \text{WHERE } \Delta \sim Alt(\pi)$$

LET $\phi_\theta(x)$ DENOTE THE NORMAL DENSITY WITH PARAMETERS $\theta = (\mu, \sigma^2)$. THEN THE DENSITY OF Y IS:

$$g_Y(x) = (1-\pi)\phi_{\theta_1}(y) + \pi \phi_{\theta_2}(y)$$

UNKNOWN PARAMETERS $\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

THE LOG-LIKELIHOOD BASED ON THE N TRAINING CASES:

$$\ell(\theta; \mathbb{Z}) = \sum_{i=1}^N \log\left[(1-\pi)\phi_{\theta_1}(y_i) + \pi \phi_{\theta_2}(y_i)\right]$$

DIRECT MAXIMIZATION OF $\ell(\theta; \mathbb{Z})$ IS QUITE DIFFICULT NUMERICALLY, BECAUSE OF THE SUM OF TERMS INSIDE THE LOGARITHM.

A SIMPLER APPROACH: CONSIDER UNOBSERVED LATENT VARIABLES Δ_i : IF $\Delta_i = 0$ THEN Y_i COMES FROM MODEL 1, OTHERWISE FROM MODEL 2.

THEN THE LOG-LIKELIHOOD WOULD BE:

$$\boxed{\begin{aligned} \ell_0(\theta, \mathbb{Z}, \boldsymbol{\Delta}) = {} & \sum_{i=1}^N \left[(1-\Delta_i)\log \phi_{\theta_1}(y_i) + \Delta_i \log_{\theta_2}(y_i)\right] + \\ & + \underline{\underline{\sum_{i=1}^N}} \left[(1-\Delta_i)\log(1-\pi) + \Delta_i \log \pi\right] \end{aligned}}$$

→ THE MAXIMUM LIKELIHOOD ESTIMATES OF μ_1, σ_1^2 (μ_2, σ_2^2, RESP.) WOULD BE THE SAMPLE MEAN AND VARIANCE FOR THOSE DATA WITH $\Delta_i = 0$ ($\Delta_i = 1$) THE ESTIMATE OF π WOULD BE THE PROPORTION OF $\Delta_i = 1$.

SINCE THE VALUES OF Δ_i ARE UNKNOWN, WE PROCEED IN AN ITERATIVE FASHION, SUBSTITUTING FOR EACH Δ_i ITS *EXPECTED VALUE*:

$$\gamma_i(\theta) = E(\Delta_i \mid \theta, \mathbb{Z}) = Pr(\Delta_i = 1 \mid \theta, \mathbb{Z}).$$

ALSO CALLED *RESPONSIBILITY* OF MODEL 2 FOR OBS. i.

EM ALGORITHM FOR TWO-COMPONENT GAUSSIAN MIX.

1. TAKE INITIAL GUESSES FOR THE PARAMETERS $\hat{\theta}$

$\hat{\mu}_1, \hat{\mu}_2$ ... CHOOSE TWO OF THE y_i AT RANDOM

$\hat{\sigma}_1^2 = \hat{\sigma}_2^2 =$ OVERALL SAMPLE VARIANCE $\sum_{i=1}^N (y_i - \bar{y})^2 / N$

THE MIXING PROPORTION CAN BE STARTED AT $\hat{\pi} = 0.5$.

2. *EXPECTATION STEP*: COMPUTE THE RESPONSIBILITIES:

$$\hat{\gamma}_i = \frac{\hat{\pi}\, \phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\, \phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\, \phi_{\hat{\theta}_2}(y_i)} \qquad i = 1, \dots, N$$

3. *MAXIMIZATION STEP*: COMPUTE WEIGTED MEANS & VARIANCES:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)\, y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)} \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i\, y_i}{\sum_{i=1}^N \hat{\gamma}_i} \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i}$$

MIXING PROBABILITY $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. ITERATE STEPS 2. AND 3. UNTIL CONVERGENCE.

2/7

22.01.2010 Fr
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

BAGGING

HOW TO USE BOOTSTRAP TO IMPROVE THE ESTIMATE OR PREDICTION ITSELF:

CONSIDER FIRST THE REGRESSION PROBLEM. SUPPOSE WE FIT A MODEL TO OUR TRAINING DATA $\mathbb{Z} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, OBTAINING THE PREDICTION $\hat{f}(x)$ AT INPUT x. BOOTSTRAP AGGREGATING OR BAGGING AVERAGES THIS PREDICTION OVER A COLLECTION OF BOOTSTRAP SAMPLES, THEREBY REDUCING ITS VARIANCE.

FOR EACH BOOTSTRAP SAMPLE $\mathbb{Z}^{*b}$, $b = 1, 2, \ldots, B$ WE FIT OUR MODEL, GIVING PREDICTION $\hat{f}^{*b}(x)$.

THE BAGGING ESTIMATE:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

IN FACT, THE TRUE BAGGING ESTIMATE IS DEFINED BY $E_{\hat{P}} \hat{f}^*(x)$, WHERE $\mathbb{Z}^* = (x_1^*, y_1^*), \ldots, (x_N^*, y_N^*)$ AND EACH $(x_i^*, y_i^*) \sim \hat{P}$.

BAGGING CAN DRAMATICALLY REDUCE THE VARIANCE OF UNSTABLE PROCEDURES LIKE TREES, LEADING TO IMPROVED PREDICTION. MORE STABLE PROCEDURES LIKE NEAREST NEIGHBORS ARE TYPICALLY NOT AFFECTED MUCH BY BAGGING.

MODEL AVERAGING AND STACKING

WE HAVE A SET OF CANDIDATE MODELS $\mathcal{M}_m$, $m=1,\dots,M$ FOR OUR TRAINING SET $\mathbb{Z}$. SUPPOSE ζ IS SOME QUANTITY OF INTEREST, FOR EXAMPLE, A PREDICTION $f(x)$ AT SOME FIXED FEATURE VALUE x.

THE POSTERIOR DISTRIBUTION OF ζ IS:

$$\Pr(\zeta \mid \mathbb{Z}) = \sum_{m=1}^{M} \Pr(\zeta \mid \mathcal{M}_m, \mathbb{Z}) \cdot \Pr(\mathcal{M}_m \mid \mathbb{Z})$$

WITH POSTERIOR MEAN

$$E(\zeta \mid \mathbb{Z}) = \sum_{m=1}^{M} E(\zeta \mid \mathcal{M}_m, \mathbb{Z}) \cdot \Pr(\mathcal{M}_m \mid \mathbb{Z})$$

THIS BAYESIAN PREDICTION IS A WEIGHTED AVERAGE OF THE INDIVIDUAL PREDICTIONS, WITH WEIGHTS PROPORTIONAL TO THE POSTERIOR PROB. OF EACH MODEL.

GIVEN PREDICTIONS $\hat{f}_1(x), \dots, \hat{f}_M(x)$, UNDER SQUARED-ERROR LOSS, WE CAN SEEK THE WEIGHTS $w=(w_1,\dots,w_M)$ SUCH THAT $\hat{w} = \underset{w}{\operatorname{argmin}}\ E_{\mathcal{P}} \left[Y - \sum_{m=1}^{M} w_m \hat{f}_m(x) \right]^2$.

HERE THE INPUT VALUE x IS FIXED AND THE N OBSERVATIONS IN THE DATASET $\mathbb{Z}$ (AND THE TARGET Y) ARE DISTRIBUTED ACCORDING TO $\mathcal{P}$.

THE SOLUTION IS POPULATION LINEAR REGRESSION OF Y ON $\hat{F}(x)^T = (\hat{f}_1(x), \dots, \hat{f}_M(x))$.

HOWEVER, THE POPULATION LINEAR REGRESSION IS NOT AVAILABLE.

EXAM: NAIL029 STROJOVÉ UCENÍ

LET $\hat{f}_m^{-i}(x)$ BE THE PREDICTION AT x, USING MODEL m, APPLIED TO THE DATASET WITH iTH TRAINING OBSERVATION REMOVED. THE STACKING ESTIMATE OF THE WEIGHTS:

$$\boxed{\hat{w}^{st} = \underset{w}{\arg\min} \sum_{i=1}^{N} \left[y_i - \sum_{m=1}^{M} w_m \hat{f}_m^{-i}(x_i) \right]^2}$$

THE FINAL PREDICTION IS $\sum_{m=1}^{M} \hat{w}_m^{st} \hat{f}_m(x)$. !

BY USING CROSS-VALIDET PREDICTIONS $\hat{f}_m^{-i}(x)$, STACKING AVOIDS GIVING UNFAIRLY HIGH WEIGHT TO MODELS WITH HIGHER COMPLEXITY.

BETTER RESULTS CAN BE OBTAINED BY RESTRICTING THE WEIGHTS TO BE NONNEGATIVE, AND TO SUM TO 1 ≡ TRACTABLE QUADRATIC PROGRAMMING PROBLEM.

STOCHASTIC SEARCH: BUMPING

BUMPING USES BOOTSTRAP SAMPLING TO MOVE RANDOMLY THROUGH MODEL SPACE. FOR PROBLEMS WHERE FITTING METHOD FINDS MANY LOCAL MINIMA, BUMPING CAN HELP THE METHOD TO AVOID GETTING STUCK IN POOR SOLUTIONS.

WE DRAW BOOTSTRAP SAMPLES $Z^{*1}, \dots, Z^{*B}$ AND FIT OUR MODEL TO EACH: $\hat{f}^{*b}(x)$, $b = 1, \dots, B$.

WE THEN CHOOSE THE MODEL THAT PRODUCES THE SMALLEST PREDICTION ERROR OVER ORIGINAL TR SET

$$\boxed{\hat{b} = \underset{b}{\arg\min} \sum_{i=1}^{N} \left[y_i - \hat{f}^{*b}(x_i) \right]^2}$$

ADDITIVE MODELS, TREES, ...

GENERALIZED ADDITIVE MODEL:

$$E(Y|X_1, \dots, X_p) = \alpha + f_1(X_1) + \dots + f_p(X_p)$$

WE FIT EACH FUNCTION USING A SCATTERPLOT SMOOTHER, AND PROVIDE AN ALGORITHM FOR SIMULTANEOUSLY ESTIMATING ALL p FUNCTIONS.

ADDITIVE LOGISTIC REGRESSION MODEL FOR TWO-CLASS CLASSIFICATION REPLACES EACH LINEAR TERM BY A MORE GENERAL FUNCTIONAL FORM

$$\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = \alpha + f_1(X_1) + \dots + f_p(X_p)$$

IN GENERAL, THE CONDITIONAL MEAN $\mu(X)$ OF A RESPONSE Y IS RELATED TO AN ADDITIVE FUNCTION OF THE PREDICTORS VIA A LINK FUNCTION g:

$$g[\mu(X)] = \alpha + f_1(X_1) + \dots + f_p(X_p)$$

$g(\mu) = \mu$ ... IDENTITY LINK, USED FOR LINEAR AND ADDITIVE MODELS FOR GAUSSIAN RESPONSE DATA

$g(\mu) = \text{logit}(\mu)$, OR $g(\mu) = \text{probit}(\mu)$ ... PROBIT LINK, USED FOR MODELLING BINOMIAL PROBABILITIES

$\text{probit}(\mu) \overset{\text{def}}{:=} \Phi^{-1}(\mu)$

$g(\mu) = \log(\mu)$ FOR LOG-LINEAR OR LOG-ADDITIVE MODELS FOR POISSON COUNT DATA

4/7

22. 01. 2010 Fr
PETER ČERNO

EXAM: NAIL 029 STROJOVÉ UCENÍ

A SIMPLE ITERATIVE PROCEDURE EXISTS FOR FINDING THE SOLUTION. WE SET $\hat{\alpha} = \mathrm{ave}(y_i)$, AND IT NEVER CHANGES. WE APPLY A CUBIC SMOOTHING SPLINE S_j TO THE TARGETS $\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^N$ AS A FUNCTION OF x_{ij}, TO OBTAIN A NEW ESTIM. $\hat{f}_j$. THIS IS DONE FOR EACH PREDICTOR IN TURN, UNTIL THE ESTIMATES $\hat{f}_j$ STABILIZE.

THE BACKFITTING ALGORITHM FOR ADDITIVE MODELS

1. INITIALIZE: $\hat{\alpha} \leftarrow \frac{1}{N} \sum_{i=1}^N y_i$, $\hat{f}_j(x_i) \equiv 0 \quad \forall i, j$
2. CYCLE: $j = 1, \dots, p, \dots, 1, \dots, p, \dots$

$$\hat{f}_j \leftarrow S_j \left[\{ y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik}) \}_1^N \right]$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^N \hat{f}_j(x_{ij})$$

TREE-BASED METHODS

TREE-BASED METHODS PARTITION THE FEATURE SPACE INTO A SET OF RECTANGLES, AND THEN FIT A SIMPLE MODEL (LIKE A CONSTANT) IN EACH ONE.

CART - POPULAR METHOD FOR TREE-BASED REGRESSION AND CLASSIFICATION

TO SIMPLIFY MATTERS, WE RESTRICT ATTENTION TO RECURSIVE BINARY PARTITIONS.

REGRESSION TREES

OUR DATA CONSISTS OF p INPUTS AND A RESPONSE, I.E. (x_i, y_i), $i=1,\dots,N$, WITH $x_i = (x_{i1}, \dots, x_{ip})$.

SUPPOSE THAT WE HAVE PARTITION INTO M REGIONS $R_1, \dots, R_M$, AND WE MODEL THE RESPONSE AS A CONSTANT c_m IN EACH REGION:

$$f(x) = \sum_{m=1}^{M} c_m \, I(x \in R_m).$$

IF WE ADOPT AS OUR CRITERION MINIMIZATION OF THE SUM OF SQUARES $\sum_{i=1}^{N} (y_i - f(x_i))^2$, IT IS EASY TO SEE THAT THE BEST $\hat{c}_m$ IS JUST THE AVERAGE IN REGION R_m:

$$\hat{c}_m = \text{ave}(y_i \mid x_i \in R_m)$$

FINDING THE BEST BINARY PARTITION IS GENERALLY COMPUTATIONALLY INFEASIBLE. HENCE WE PROCEED WITH A GREEDY ALGORITHM.

CONSIDER A SPLITTING VARIABLE j AND SPLIT POINT s.

$$R_1(j,s) = \{X \mid X_j \leq s\} \qquad R_2(j,s) = \{X \mid X_j > s\}$$

WE SEEK j AND s THAT SOLVE:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

... APPARENTLY $\hat{c}_1 = \text{ave}(y_i \mid x_i \in R_1(j,s))$, AND SIMILARLY $\hat{c}_2$

HAVING FOUND THE BEST SPLIT, WE PROCEED RECURSIVELY ON EACH REGION.

TREE SIZE IS A TUNNING PARAMETER.

22.01.2010 Fr
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

PREFFERED STRATEGY: GROW A LARGE TREE T_0, STOPPING THE SPLITTING PROCESS ONLY WHEN SOME MINIMUM NODE SIZE IS REACHED. THEN THIS TREE IS PRUNNED USING <u>COST-COMPLEXITY PRUNNING</u>:

$T \subseteq T_0$ ... ANY TREE THAT CAN BE OBTAINED BY PRUNNING T_0, THAT IS, COLLAPSING ANY NUMBER OF ITS INTERNAL NODES

WE INDEX TERMINAL NODES WITH m, REGIONS R_m

$|T|$ ... NUMBER OF TERMINAL NODES IN T

$$N_m := | \{ x_i \in R_m \} |$$

$$\hat{c}_m := \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$\underset{\sim}{Q}_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2 \quad \text{(NODE IMPURITY)}$$

COST COMPLEXITY CRITERION $C_\alpha(T) = \sum_{m=1}^{|T|} N_m \underset{\sim}{Q}_m(T) + \alpha |T|$

IDEA: FOR EACH α FIND $T_\alpha \subseteq T_0$ WHICH MINIMIZES $C_\alpha(T)$.

$\alpha \geq 0$ GOVERNS THE TRADEOFF BETWEEN TREE SIZE AND ITS GOODNESS OF FIT TO THE DATA

<u>WEAKEST LINK PRUNNING</u>: WE SUCCESSIVELY COLLAPSE THE INTERNAL NODE THAT PRODUCES THE SMALLEST PER-NODE INCREASE IN $\sum_m N_m Q_m(T)$, AND CONTINUE UNTIL WE PRODUCE THE SINGLE NODE-ROOT. THIS GIVES A FINITE SEQUENCE OF SUBTREES, AND ONE CAN SHOW THIS SEQUENCE MUST CONTAIN T_α.

<u>ESTIMATION OF α</u>: FIVE/TEN-FOLD CROSS-VALIDATION

CLASSIFICATION TREES

TARGET IS A CLASSIFICATION OUTCOME $\in \{1, \dots, K\}$

LET $\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$

$\equiv$ THE PROPORTION OF CLASS k OBSERVATIONS IN NODE m.
WE CLASSIFY THE OBSERVATIONS IN NODE m TO CLASS $k(m) = \operatorname{argmax}_k \hat{p}_{mk}$, THE MAJORITY CLASS IN NODE m.

$$Q_m(T) \left\{ \begin{array}{l} \text{MISCLASSIFICATION ERROR}: \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{m k(m)} \\ \text{GINI INDEX}: \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk}) \\ \text{CROSS-ENTROPY (DEVIANCE)}: -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \end{array} \right.$$

FOR TWO CLASSES, IF p IS THE PROPORTION OF THE SECOND CLASS, THESE MEASURES ARE:

$$1 - \max(p, 1-p) \quad ; \quad 2p(1-p) \quad ; \quad -p \log p - (1-p) \log(1-p)$$

CROSS-ENTROPY AND GINI INDEX ARE DIFFERENTIABLE

CONSIDER A TWO-CLASS PROBLEM WITH 400 OBSERVATIONS IN EACH CLASS (400, 400). SUPPOSE:

1. SPLIT : (300, 100) (100, 300)
2. SPLIT : (200, 400) (200, 0)

BOTH GINI INDEX AND CROSS-ENTROPY ARE LOWER FOR THE SECOND SPLIT $\Rightarrow$ THEY SHOULD BE USED WHEN GROWING THE TREE.

INTERPRETATION OF GINI INDEX: WE CLASSIFY THE OBSERVATION TO CLASS k WITH PROB. $\hat{p}_{mk}$
$\Rightarrow$ THE TRAINING ERROR RATE = GINI INDEX

22.01.2010 Fr
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

CATEGORICAL PREDICTORS

WHEN SPLITTING A PREDICTOR HAVING q POSSIBLE UNORDERED VALUES, THERE ARE $2^{q}-1$ POSSIBLE PARTITIONS. HOWEVER, WITH A 0-1 OUTCOME, IF WE ORDER THE PREDICTOR CLASSES ACCORDING TO THE PROPORTION FALLING IN OUTCOME CLASS 1, THEN WE CAN SPLIT THIS PREDICTOR AS IF IT WERE AN ORDERED PREDICTOR. THIS GIVES THE OPTIMAL SPLIT, IN TERMS OF CROSS-ENTROPY OR GINI INDEX. THIS RESULT HOLDS ALSO FOR A QUANTITATIVE OUTCOME AND SQUARE ERROR LOSS, IF CATEGORIES ARE ORDERED BY INCREASING MEAN OF THE OUTCOME.
FOR MULTICATEGORY OUTCOMES, NO SUCH SIMPL. ARE POSSIBLE.

THE LOSS MATRIX

IN CLASSIFICATION PROBLEMS, THE CONSEQUENCES OF MISCLASSIFYING OBSERVATIONS ARE MORE SERIOUS IN SOME CLASSES THAN OTHERS.

$K \times K$ LOSS MATRIX $\mathbb{L}$, $L_{kk'}$ ... LOSS INCURED FOR CLASSIFYING A CLASS k OBSERVATION AS CLASS k'.

TYPICALLY $L_{kk} = 0 \ \forall k$.

GINI INDEX: $\sum_{k \neq k'} L_{kk'} \hat{p}_{mk} p_{mk'}$

FOR TWO-CLASS A BETTER APPROACH IS TO WEIGH THE OBSERVATIONS IN CLASS k BY $L_{kk'}$

WE CLASSIFY TO CLASS $\boxed{k(m) = \arg\min_k \sum_\ell L_{\ell k} \hat{p}_{m\ell}}$

MISSING PREDICTOR VALUES

TWO APPROACHES:

(1) FOR CATEGORICAL PREDICTOR - SIMPLY MAKE A NEW CATEGORY FOR "MISSING"

(2) MORE GENERAL APPROACH - THE CONSTRUCTION OF SURROGATE VARIABLES

OTHER TREE-BUILDING PROCEDURES

THE DISCUSSION ABOVE FOCUSES ON CART (CLASSIFICATION AND REGRESSION TREE). THE OTHER POPULAR METHODOLOGY IS ID3 AND ITS LATER VERSIONS, C4.5 AND C5.0.

LINEAR COMBINATION SPLITS

RATHER THAN RESTRICTING SPLITS TO $X_j \leq s$, ONE CAN ALLOW $\sum a_j X_j \leq s$. WHILE THIS CAN IMPROVE THE PREDICTIVE POWER OF THE TREE, IT CAN HURT INTERPRETABILITY.

INSTABILITY OF TREES

ONE MAJOR PROBLEM WITH TREES IS THEIR HIGH VARIANCE. BAGGING AVERAGES MANY TREES TO REDUCE THIS VARIANCE.

LACK OF SMOOTHNESS

THE MARS PROCEDURE CAN BE VIEWED AS MODIFICATION OF CART DESIGNED TO ALLEVIATE THE LACK OF SMOOTHNESS.

DIFFICULTY IN CAPTURING ADDITIVE STRUCTURE

AGAIN THE MARS METHOD ...

7/7

22.01.2010 Fr
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

IN MEDICAL CLASSIFICATION PROBLEMS:
SENSITIVITY: PROBABILITY OF PREDICTING DISEASE GIVEN TRUE STATE IS DISEASE
SPECIFICITY: PROBABILITY OF PREDICTING NON-DISEASE GIVEN TRUE STATE IS NON-DISEASE

ROC ... RECEIVER OPERATING CHARACTERISTIC CURVE IS A COMMONLY USED SUMMARY FOR ASSESSING THE TRADEOFF BETWEEN SENSITIVITY AND SPECIFICITY IT IS A PLOT OF THE SENSITIVITY VERSUS SPECIFICITY AS WE VARY THE PARAMETERS OF A CLASSIFICATION RULE.

PRIM: BUMP HUNTING

TREE-BASED METHODS (FOR REGRESSION) PARTITION THE FEATURE SPACE INTO BOX-SHAPED REGIONS, TO TRY TO MAKE THE RESPONSE AVERAGES IN EACH BOX AS DIFFERENT AS POSSIBLE.
THE PATIENT RULE INDUCTION METHOD (PRIM) SEEKS BOXES IN WHICH THE RESPONSE AVERAGE IS HIGH.

1. START WITH ALL OF THE TRAINING DATA, AND A MAXIMAL BOX CONTAINING ALL OF THE DATA.
2. CONSIDER SHRINKING THE BOX BY COMPRESSING ONE FACE, SO AS TO PEEL OFF THE PROPORTION α OF OBSERVATIONS, CHOOSE THE PEELING THAT PRODUCES THE HIGHEST RESPONSE MEAN IN THE REMAINING BOX. (TYPICALLY $\alpha=0.05$, OR $\alpha=0.10$)

3. REPEAT STEP 2 UNTIL SOME MINIMAL NUMBER OF OBSERVATIONS (SAY 10) REMAIN IN THE BOX.

4. EXPAND THE BOX ALONG ANY FACE, AS LONG AS THE RESULTING BOX MEAN INCREASES.

5. STEPS 1-4. GIVE A SEQUENCE OF BOXES - USE CROSS-VALIDATION TO CHOOSE A MEMBER OF THE SEQUENCE → B_1

6. REMOVE THE DATA IN THE BOX B_1 FROM THE DATASET AND REPEAT 2.-5. TO OBTAIN A SECOND BOX, ETC.

PRIM CAN HANDLE CATEGORICAL PREDICTOR BY CONSIDERING ALL PARTITIONS OF THE PREDICTOR, AS IN CART. PRIM IS DESIGNED FOR REGRESSION. A TWO-CLASS OUTCOME CAN BE HANDLED SIMPLY BY CODING IT AS 0 AND 1. THERE IS NO SIMPLE WAY TO DEAL WITH $K > 2$ CLASSES SIMULTANEOUSLY.

1/8

23.01.2010 Sa
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

MARS: MULTIVARIATE ADAPTIVE REGRESSION SPLINES

MARS IS AN ADAPTIVE PROCEDURE FOR REGRESSION, AND IS WELL SUITED FOR HIGH-DIMENSIONAL PROBLEMS.

REFLECTED PAIR: $(x-t)_+$, $(t-x)_+$, KNOT t

THE IDEA IS TO FORM REFLECTED PAIRS FOR EACH INPUT X_j WITH KNOTS AT EACH OBSERVED VALUE x_{ij}.

THE COLLECTION OF BASIS FUNCTIONS IS

$$\mathcal{C} = \{(X_j - t)_+, (t - X_j)_+\}_{\substack{t \in \{x_{1j}, \dots, x_{Nj}\} \\ j \in \{1, \dots, p\}}}$$

$\equiv 2Np$ BASIS FUNCTIONS.

THE MODEL HAS THE FORM: $f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$

h_m IS A FUNCTION IN $\mathcal{C}$, OR A PRODUCT OF TWO OR MORE SUCH FUNCTIONS.

GIVEN A CHOICE FOR THE h_m, COEFFICIENTS β_m ARE ESTIMATED BY MINIMIZING THE RSS.

THE REAL ART IS THE CONSTRUCTION OF THE FNCS. $h_m(x)$.

WE START WITH ONLY $h_0(X) = 1$.

AT EACH STAGE WE ADD TO THE MODEL $\mathcal{M}$ THE TERM

$$\hat{\beta}_{M+1} h_e(X)(X_j - t)_+ + \hat{\beta}_{M+2} h_e(X)(t - X_j)_+,$$

WHERE $h_e \in \mathcal{M}$ AND $(X_j - t)_+, (t - X_j)_+ \in \mathcal{C}$

THAT PRODUCES THE LARGEST DECREASE IN TRAINING ERR.

$\hat{\beta}_{M+1}, \hat{\beta}_{M+2}$ ARE LINEAR LEAST SQUARES ESTIMATES

AT THE END OF THIS PROCEDURE WE HAVE A LARGE MODEL, SO A BACKWARD DELETION PROCEDURE IS APPLIED. THE TERM WHOSE REMOVAL CAUSES THE SMALLEST INCREASE IN RESIDUAL SQUARED ERROR IS DELETED FROM THE MODEL AT EACH STAGE, PRODUCING AN ESTIMATED BEST MODEL $\hat{f}_\lambda$ OF EACH SIZE λ. ONE COULD USE CROSS-VALIDATION TO ESTIMATE λ, BUT FOR COMPUTATIONAL SAVINGS THE MARS PROCEDURE INSTEAD USES GENERALIZED CROSS-VALIDATION $GCV(\lambda)$.

RELATIONSHIP OF MARS TO CART

- REPLACE THE PIECEWISE LINEAR BASIS FUNCTIONS BY STEP FUNCTIONS $I(x-t > 0)$, $I(x-t \leq 0)$
- WHEN A MODEL TERM IS INVOLVED IN A MULTIPLICATION BY A CANDIDATE TERM, IT GETS REPLACED BY THE INTERACTION, AND HENCE IS NOT AVAILABLE FOR FURTHER INTERACTIONS

WITH THESE CHANGES, THE MARS FORWARD PROCEDURE IS THE SAME AS THE CART TREE-GROWING ALGORITHM.

GENERALIZED CROSS-VALIDATION:

$$\boxed{GCV(\lambda) = \frac{\sum_{i=1}^{N} (y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}}$$

$M(\lambda)$ ... THE EFFECTIVE NUMBER OF PARAMETERS:

$$\boxed{\begin{array}{l} M(\lambda) = r + c.K \\ c = 3 \end{array}}$$

, r ... # LINIARY INDEP. BASIS FNCS. IN $\hat{f}_\lambda$

K ... # OF KNOTS IN $\hat{f}_\lambda$

23.01.2010 Sa
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

BOOSTING AND ADDITIVE TREES

BOOSTING METHODS

BOOSTING IS ONE OF THE MOST POWERFUL LEARNING IDEAS INTRODUCED IN THE LAST TWENTY YEARS. THE MOTIVATION FOR BOOSTING WAS A PROCEDURE THAT COMBINES THE OUTPUTS OF MANY "WEAK" CLASSIFIERS TO PRODUCE A POWERFUL "COMMITTEE".

MOST POPULAR BOOSTING ALGORITHM:
FREUND, SCHAPIRE (1997) - ADABOOST. M1

CONSIDER A TWO-CLASS PROBLEM, WITH THE OUTPUT VARIABLE CODED AS $Y \in \{-1, +1\}$.

CLASSIFIER $G(X)$ PRODUCES A PREDICTION $\in \{-1, +1\}$

THE ERROR RATE ON THE TRAINING SAMPLE

$$\overline{err} = \frac{1}{N} \sum_{i=1}^{N} I(y_i \neq G(x_i))$$

EXPECTED ERROR RATE ... $E_{XY}\, I(Y \neq G(X))$

WEAK CLASSIFIER ... ERROR RATE IS SLIGHTLY BETTER THAN RANDOM GUESSING

THE PURPOSE OF BOOSTING IS TO SEQUENTIALLY APPLY THE WEAK CLASSIFICATION ALGORITHM TO REPEATEDLY MODIFIED VERSIONS OF THE DATA.

$\rightsquigarrow$ SEQUENCE OF WEEK CLASSIFIERS $G_m(X)$, $m = 1, \dots, M$

THE FINAL PREDICTION:

$$\boxed{G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)}$$

$\alpha_1, \dots, \alpha_M$ ARE COMPUTED BY THE BOOSTING ALGORITHM

THE DATA MODIFICATIONS AT EACH BOOSTING STEP CONSISTS OF APPLYING WEIGHTS $w_1, \dots, w_N$ TO THE TRAINING OBSERVATIONS (x_i, y_i). (INITIALLY $w_i = \frac{1}{N}$)

AT STEP m, THOSE OBSERVATIONS THAT WERE MISSCLASSIFIED BY $G_{m-1}(x)$ HAVE THEIR WEIGHTS INCREASED, WHEREAS THE WEIGHTS ARE DECREASED FOR THOSE THAT WERE CLASSIFIED CORRECTLY.

AdaBoost. M1

```
1. INITIALIZE THE OBSERVATION WEIGHTS $w_i = \frac{1}{N} \ \forall i$
2. FOR $m = 1$ to $M$:
    (a) FIT A CLASSIFIER $G_m(x)$ TO THE TRAINING DATA USING WEIGHTS $w_i$
    (b) COMPUTE $err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$
    (c) COMPUTE $\alpha_m = \log\left(\frac{1 - err_m}{err_m}\right)$
    (d) SET $w_i \leftarrow w_i \cdot \exp\left(\alpha_m \cdot I(y_i \neq G_m(x_i))\right)$
3. OUTPUT $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$
```

3/8

23.01.2010 Sa
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

BOOSTING IS A WAY OF FITTING AN ADDITIVE EXPANSION IN A SET OF ELEMENTARY BASIS FUNCTIONS. BASIS FUNCTION EXPANSION TAKE THE FORM $f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m)$.
TYPICALLY THESE MODELS ARE FIT BY MINIMIZING A LOSS FUNCTION AVERAGED OVER THE TRAINING DATA, SUCH AS THE SQUARED-ERROR OR A LIKELIHOOD-BASED LOSS FUNCTION,

$$\min_{\{\beta_m, \gamma_m\}_1^M} \sum_{i=1}^{N} L\left(y_i, \sum_{m=1}^{M} \beta_m b(x_i; \gamma_m)\right)$$

$\therefore$ OFTEN THIS REQUIRES COMPUTATIONALLY INTENSIVE NUMERICAL OPTIMIZATION TECHNIQUES.
SIMPLE ALTERNATIVE: FITTING A SINGLE BASIS FUNCTION:

$$\min_{\beta, \gamma} \sum_{i=1}^{N} L(y_i, \beta b(x_i; \gamma))$$

FORWARD STAGEWISE ADDITIVE MODELING

1. INITIALIZE $f_0(x) = 0$
2. FOR $m = 1$ to M:
   (a) COMPUTE

   $$(\beta_m, \gamma_m) = \arg\min_{(\beta, \gamma)} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta b(x_i, \gamma))$$

   (b) SET $f_m(x) := f_{m-1}(x) + \beta_m b(x, \gamma_m)$

I.E. PREVIOUSLY ADDED TERMS ARE NOT MODIFIED ...

EXPONENTIAL LOSS AND ADABOOST

WE SHOW THAT ADABOOST.M1 IS EQUIVALENT TO FORWARD STAGEWISE ADDITIVE MODELLING USING THE LOSS FUNCTION: $L(y, f(x)) = \exp(-y\, f(x))$.

FOR ADABOOST THE BASIS FUNCTIONS ARE THE INDIVIDUAL CLASSIFIERS $G_m(x) \in \{-1, 1\}$.

USING THE EXPONENTIAL LOSS FUNCTION, ONE MUST SOLVE:

$$(\beta_m, G_m) = \underset{\beta, G}{\arg\min} \sum_{i=1}^{N} \exp\left[-y_i\left(f_{m-1}(x_i) + \beta G(x_i)\right)\right]$$

$$(\beta_m, G_m) = \underset{\beta, G}{\arg\min} \sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i\, G(x_i)\right) \qquad (*)$$

WHERE $w_i^{(m)} = \exp\left(-y_i\, f_{m-1}(x_i)\right)$.

SINCE $w_i^{(m)}$ DEPENDS NEITHER ON β NOR $G(x)$, IT CAN BE REGARDED AS A WEIGHT THAT IS APPLIED TO EACH OBSERVATION. FOR ANY $\beta > 0$:

$$\sum_{i=1}^{N} w_i^{(m)} \exp\left(-\beta y_i\, G(x_i)\right) =$$

$$= e^{-\beta} \sum_{y_i = G(x_i)} w_i^{(m)} + e^{\beta} \sum_{y_i \neq G(x_i)} w_i^{(m)} =$$

$$= \left(e^{\beta} - e^{-\beta}\right) \sum_{i=1}^{N} w_i^{(m)} I\left(y_i \neq G(x_i)\right) + e^{-\beta} \sum_{i=1}^{N} w_i^{(m)} \qquad (**)$$

THUS $G_m = \underset{G}{\arg\min} \sum_{i=1}^{N} w_i^{(m)} I\left(y_i \neq G(x_i)\right)$

PLUGGING THIS G_m INTO (*) AND SOLVING FOR β:

$$\beta_m = \frac{1}{2} \log\left(\frac{1-err_m}{err_m}\right)$$

WHERE $err_m = \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G_m(x_i)) / \sum_{i=1}^{N} w_i^{(m)}$

THE APPROXIMATION IS THEN UPDATED:

$$f_m(x) = f_{m-1}(x) + \beta_m G_m(x)$$

WHICH CAUSES THE WEIGHTS FOR THE NEXT ITERATION TO BE:

$$w_i^{(m+1)} = \exp(-y_i f_m(x_i)) = \exp\left[-y_i \left(f_{m-1}(x) + \beta_m G_m(x)\right)\right] =$$

$$= w_i^{(m)} \cdot \exp(-\beta_m y_i G_m(x_i)).$$

USING THE FACT $-y_i G_m(x_i) = 2 \cdot I(y_i \neq G_m(x_i)) - 1$

$$w_i^{(m+1)} = w_i^{(m)} \cdot e^{\alpha_m I(y_i \neq G_m(x_i))} \cdot e^{-\beta_m}$$

WHERE $\alpha_m = 2\beta_m$ IS THE QUANTITY DEFINED AT 2c). THE FACTOR $e^{-\beta_m}$ MULTIPLIES ALL WEIGHTS, SO IT HAS NO EFFECT, AND IS EQUIVALENT TO 2d).

ONE CAN VIEW 2a) AS A METHOD FOR APPROXIMATELY SOLVING THE MINIMIZATION OF (**), AND HENCE (*).

HENCE WE CONCLUDE THAT ADABOOST.M1 MINIMIZES THE EXPONENTIAL LOSS CRITERION

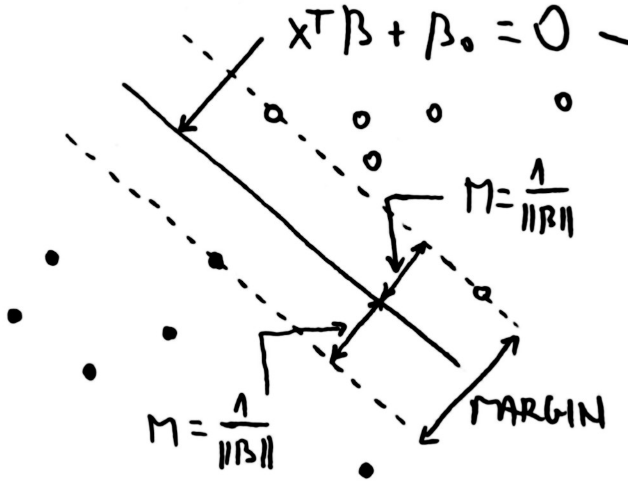
$$L(y, f(X)) = \exp(-y\, f(x))$$

VIA A FORWARD-STAGEWISE ADDITIVE MODELING APPROACH.

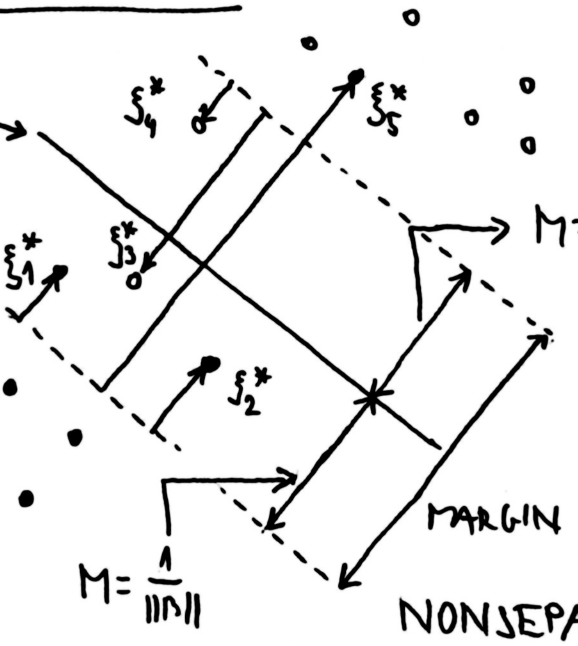
SUPPORT VECTOR MACHINES AND FLEXIBLE DISCRIMINANTS

GENERALIZATIONS OF LINEAR DECISION BOUNDARIES FOR CLASSIFICATION.

SUPPORT VECTOR CLASSIFIER



SEPARABLE CASE



NONSEPARABLE CASE

THE POINTS LABELED ξ_j^* ARE ON THE WRONG SIDE OF THEIR MARGIN BY AN AMOUNT $\xi_j^* = M\xi_j$, POINTS ON THE CORRECT SIDE HAVE $\xi_j^* = 0$.

THE MARGIN (2M) IS MAXIMIZED SUBJECT TO A TOTAL BUDGET $\sum \xi_j \leq$ CONSTANT.

OUR TRAINING DATA CONSISTS OF N PAIRS $(x_1, y_1), \dots (x_N, y_N)$, WITH $x_i \in \mathbb{R}^p$ AND $y_i \in \{-1, +1\}$.

DEFINE A HYPERPLANE BY $\{x \mid f(x) = x^T\beta + \beta_0 = 0\}$, WHERE $\|\beta\| = 1$. A CLASSIFICATION RULE INDUCED BY $f(x)$ IS $\boxed{G(x) = \text{sign}\left[x^T\beta + \beta_0\right]}$.

5/8

23. 01. 2010 Sa
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

THE SEPARABLE CASE CAPTURES THE FOLLOWING PROB.:

$$\max_{\beta, \beta_0, ||\beta||=1} M \quad \text{SUBJECT TO } y_i (x_i^T \beta + \beta_0) \geq M \quad \forall i$$

WHICH IS EQUIVALEN TO:

$$\min_{\beta, \beta_0} ||\beta|| \quad \text{SUBJECT TO } y_i (x_i^T \beta + \beta_0) \geq 1$$

VIA $M = 1/||\beta||$. (CONVEX OPTIMIZATION PROBLEM)

SUPPOSE THAT THE CLASSES OVERLAP IN FEATURE SPACE. ONE WAY TO DEAL WITH THE OVERLAP IS TO MAXIMIZE M, BUT ALLOW FOR SOME POINTS TO BE ON THE WRONG SIDE OF THE MARGIN.

DEFINE THE SLACK VARIABLES $\xi = (\xi_1, \dots, \xi_N)$.

TWO NATURAL WAYS:

a) $y_i (x_i^T \beta + \beta_0) \geq M - \xi_i$

b) $y_i (x_i^T \beta + \beta_0) \geq M(1 - \xi_i)$

→ LEAD TO DIFFERENT SOLUTIONS

$\forall i \ \xi_i \geq 0, \ \sum_{i=1}^N \xi_i \leq$ CONSTANT.

THE FIRST CHOICE a) RESULTS IN A NONCONVEX OPTIMIZATION PROBLEM, WHILE THE SECOND b) IS CONVEX.

THE VALUE ξ_i IN THE CONSTRAINT $y_i (x_i^T \beta + \beta_0) \geq$ $\geq M(1 - \xi_i)$ IS THE PROPORTIONAL AMOUNT BY WHICH PREDICTIONS FALL ON THE WRONG SIDE. MISCLASSIFICATION OCCURS WHEN $\xi_i > 1$.

EQUIVALENT FORM:

$$\min \|\beta\| \quad \text{SUBJECT TO} \quad \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0, \ \sum \xi_i \leq \text{CONSTANT} \end{cases}$$

COMPUTING THE SUPPORT VECTOR CLASSIFIER

COMPUTATIONALLY IT IS CONVENIENT TO RE-EXPRESS THE PROBLEM IN THE EQUIVALENT FORM:

$$(*) \quad \boxed{\min_{\beta, \beta_0} \frac{1}{2}\|\beta\|^2 + \underbrace{C}_{\text{TUNNING PARAMETER}}\sum_{i=1}^{N} \xi_i \quad \text{SUBJECT TO} \quad \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0 & \forall i \end{cases}}$$

THE SEPARABLE CASE CORRESPONDS TO $C = \infty$.

THE LAGRANGE (PRIMAL) FUNCTION IS:

$$\boxed{L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)\right] - \sum_{i=1}^{N}\mu_i\xi_i, \quad \text{WHICH WE MINIMIZE W.R.T. } \beta, \beta_0, \xi_i.}$$

SETTING THE RESPECTIVE DERIVATES TO ZERO:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i$$

$$\alpha_i = C - \mu_i, \quad \forall i$$

$$\alpha_i, \mu_i, \xi_i \geq 0 \quad \forall i$$

BY SUBSTITUING THESE INTO L_p WE OBTAIN THE LAGRANGIAN (WOLFE) DUAL OBJECTIVE FUNCTION

$$\boxed{L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'} x_i^T x_{i'}}$$

WHICH GIVES A LOWER BOUND ON THE OBJ.FNC. (*).

23.01.2010 Sa
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

WE MAXIMIZE L_D SUBJECT TO $0 \leq \alpha_i \leq C$, AND $\sum_{i=1}^{N} \alpha_i y_i = 0$. IN ADDITION, THE KARUSH-KUHN-TUCKER CONDITIONS INCLUDE:

$$\left.\begin{array}{ll} (1) & \alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)] = 0 \\ (2) & \mu_i \xi_i = 0 \\ (3) & y_i (x_i^T \beta + \beta_0) - (1 - \xi_i) \geq 0 \end{array}\right\} \forall i = 1, \ldots, N$$

TOGETHER THESE EQUATIONS UNIQUELY CHARACTERIZE THE SOLUTION TO THE PRIMAL AND DUAL PROBLEM. THE SOLUTION FOR β HAS THE FORM

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i$$

WITH NONZERO COEFFICIENTS $\hat{\alpha}_i$ ONLY FOR THOSE OBSERVATIONS i FOR WHICH THE CONSTRAINTS IN (3) ARE EXACTLY MET. (DUE TO (1)) THESE OBSERVATIONS ARE CALLED THE SUPPORT VECTORS, SINCE $\hat{\beta}$ IS REPRESENTED IN TERMS OF THEM ALONE. AMONG THESE SUPPORT VECTORS, SOME WILL LIE ON THE EDGE OF THE MARGIN ($\hat{\xi}_i = 0$), AND HENCE $0 < \hat{\alpha}_i < C$. THE REMAINDER ($\hat{\xi}_i > 0$) HAVE $\hat{\alpha}_i = C$. FROM (2) WE CAN SEE THAT ANY OF THESE MARGIN POINTS ($0 < \hat{\alpha}_i$, $\hat{\xi}_i = 0$) CAN BE USED TO SOLVE FOR β_0. DECISION FNC.: $\hat{G}(x) = \text{sign}[\hat{f}(x)]$.

SUPPORT VECTOR MACHINES AND KERNELS

WE CAN MAKE THE PROCEDURE MORE FLEXIBLE BY ENLARGING THE FEATURE SPACE USING BASIS EXPANSIONS SUCH AS POLYNOMIALS OR SPLINES: $h_m(x)$, $m = 1, \dots, M$.

WE CAN REPRESENT THE OPTIMIZATION PROBLEM AND ITS SOLUTION IN A SPECIAL WAY THAT ONLY INVOLVES THE INPUT FEATURES VIA INNER PRODUCTS.

THE LAGRANGE DUAL FUNCTION HAS THE FORM:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'}) \rangle$$

AND FROM $\beta = \sum_{i=1}^{N} \alpha_i y_i h(x_i)$ WE CAN SEE THAT

$$f(x) = h(x)^T \beta + \beta_0 = \sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0 .$$

$\Rightarrow$ WE DO NOT NEED TO SPECIFY THE TRANSFORMATION $h(x)$ AT ALL, BUT REQUIRE ONLY KNOWLEDGE OF THE KERNEL FUNCTION $K(x, x') = \langle h(x), h(x') \rangle$.

dTH-DEGREE POLYNOMIAL: $K(x, x') = (1 + \langle x, x' \rangle)^d$

RADIAL BASIS: $K(x, x') = \exp(-\gamma \|x - x'\|^2)$

NEURAL NETWORK: $K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$

7/8

23.01.2010 Sa
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

REDUCED-RANK LINEAR DISCRIMINANT AN.

SUPPOSE WE MODEL EACH CLASS DENSITY AS MULTIVARIATE GAUSSIAN:

$$\boxed{f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)}$$

LINEAR DISCRIMINANT ANALYSIS (LDA) ARISES IN THE SPECIAL CASE WHEN WE ASSUME THAT THE CLASSES HAVE A COMMON COVARIANCE MATRIX Σ.

SINCE $\boxed{Pr(G=k \mid X=x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}}$

WE SEE THAT: $\log \frac{Pr(G=k \mid X=x)}{Pr(G=l \mid X=x)} = \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} =$

$$= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T\Sigma^{-1}(\mu_k-\mu_\ell) + x^T\Sigma^{-1}(\mu_k-\mu_\ell)$$

CORRESPONDING LINEAR DISCRIMINANT FUNCTIONS:

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k$$

IF THE Σ_k ARE NOT ASSUMED TO BE EQUAL, WE GET QUADRATIC DISCRIMINANT FUNCTIONS (QDA)

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

THE DECISION BOUNDARY BETWEEN EACH PAIR OF CLASSES k AND ℓ: $\{x : \delta_k(x) = \delta_\ell(x)\}$

COMPUTATIONS FOR LDA

SUPPOSE WE COMPUTE THE EIGEN-DECOMPOSITION FOR EACH $\boxed{\hat{\Sigma}_k = \mathbb{U}_k \mathbb{D}_k \mathbb{U}_k^T}$ WHERE $\mathbb{U}_k$ IS $p \times p$ ORTHONORMAL, AND $\mathbb{D}_k$ A DIAGONAL MATRIX OF POSITIVE EIGENVALUES $d_{k\ell}$. THEN:

$$(x-\hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x-\hat{\mu}_k) = [\mathbb{U}_k^T (x-\hat{\mu}_k)]^T \mathbb{D}_k^{-1} [\mathbb{U}_k^T (x-\hat{\mu}_k)]$$

$$\log |\hat{\Sigma}_k| = \sum_\ell \log d_{k\ell} .$$

THE LDA CLASSIFIER CAN BE IMPLEMENTED BY THE FOLLOWING PAIR OF STEPS:

1. SPHERE THE DATA WITH RESPECT TO THE COMMON COVARIANCE ESTIMATE $\hat{\Sigma}$:

   $X^* \leftarrow \mathbb{D}^{-1/2} \mathbb{U}^T X$, WHERE $\hat{\Sigma} = \mathbb{U}\mathbb{D}\mathbb{U}^T$.

   $\Rightarrow$ THE COMMON COVARIANCE ESTIMATE OF X^* WILL NOW BE THE IDENTITY

2. CLASSIFY TO THE CLOSEST CLASS CENTROID IN THE TRANSFORMED SPACE, MODULO THE EFFECT OF THE CLASS PRIOR PROBABILITIES.

REDUCED-RANK LINEAR DISCRIMINANT ANALYSIS

THE K CENTROIDS IN p-DIMENSIONAL INPUT SPACE LIE IN AN AFFINE SUBSPACE OF DIMENSION $\leq K-1$. THUS IF p IS MUCH LARGER THAN K, WE MIGHT PROJECT THE X^* ONTO THIS CENTROID-SPANNING SUBSPACE H_{K-1}, AND MAKE DISTANCE COMPARISIONS THERE.

EXAM: NAIL029 STROJOVÉ UČENÍ

WE MIGHT ASK FOR $L < K-1$ DIMENSIONAL SUBSPACE $H_L \subseteq H_{K-1}$ OPTIMAL FOR LDA IN SOME SENSE. FISHER DEFINED OPTIMAL TO MEAN THAT THE PROJECTED CENTROIDS WERE SPREAD OUT AS MUCH AS POSSIBLE IN TERMS OF VARIANCE. THIS AMOUNTS TO FINDING PRINCIPAL COMPONENT SUBSPACES OF THE CENTROIDS THEMSELVES.

- COMPUTE THE $K \times p$ MATRIX OF CLASS CENTROIDS $\mathbb{M}$ AND THE COMMON COVARIANCE MATRIX $\mathbb{W}$ (FOR <u>W</u>ITHIN-CLASS COVARIANCE)
- COMPUTE $\mathbb{M}^* = \mathbb{M}\mathbb{W}^{-1/2}$ USING THE EIGEN-DECOMPOSITION OF $\mathbb{W} = \mathbb{U}\mathbb{D}\mathbb{U}^T$, I.E. $\mathbb{W}^{-1/2} = \mathbb{U}\mathbb{D}^{-1/2}$
- COMPUTE $\mathbb{B}^*$, THE COVARIANCE MATRIX OF $\mathbb{M}^*$ (FOR <u>B</u>ETWEEN-CLASS COVARIANCE), AND ITS EIGEN-DECOMPOSITION $\mathbb{B}^* = \mathbb{V}^* \mathbb{D}_B \mathbb{V}^{*T}$. THE COLUMNS v_ℓ^* OF $\mathbb{V}^*$ IN SEQUENCE FROM FIRST TO LAST DEFINE THE COORDINATES OF THE OPTIMAL SUBSPACES.

THE lTH DISCRIMINANT VARIABLE IS GIVEN BY

$Z_\ell = v_\ell^T X$ WITH $v_\ell = \mathbb{W}^{-1/2} v_\ell^*$, I.E.

$Z_\ell = v_\ell^{*T} \mathbb{D}^{-1/2} \mathbb{U}^T X = v_\ell^* X^*$

1/7

24.01.2010 Su
PETER CERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

UNSUPERVISED LEARNING

≡ LEARNING WITHOUT A TEACHER. WE HAVE SET OF N OBSERVATIONS $(x_1, \dots, x_N)$ OF A RANDOM p-VECTOR X HAVING JOINT DENSITY $Pr(X)$. THE GOAL IS TO DIRECTLY INFER THE PROPERTIES OF THIS PROBABILITY DENSITY WITHOUT THE HELP OF A SUPERVISOR OR TEACHER PROVIDING CORRECT ANSWERS OR DEGREE-OF-ERROR FOR EACH OBSERVATION.

IN LOW-DIMENSIONAL PROBLEMS, THERE ARE A VARIETY OF EFFECTIVE NONPARAMETRIC METHODS FOR DIRECTLY ESTIMATING THE DENSITY $Pr(X)$ ITSELF. OWING TO THE CURSE OF DIMENSIONALITY, THESE METHODS FAIL IN HIGH DIMENSIONS.

IN THE CONTEXT OF UNSUPERVISED LEARNING, THERE IS NO DIRECT MEASURE OF SUCCESS.

ASSOCIATION RULES

POPULAR TOOL FOR MINING COMMERCIAL DATA BASES. THE GOAL IS TO FIND JOINT VALUES OF THE VARIABLES $X = (X_1, X_2, \dots, X_p)$ THAT APPEAR MOST FREQUENTLY IN THE DATA BASE. IT IS MOST OFTEN APPLIED TO BINARY-VALUED DATA $X_j \in \{0, 1\}$, WHERE IT IS REFERRED TO AS "MARKET BASKET" ANALYSIS.

LET S_j REPRESENT THE SET OF ALL POSSIBLE VALUES OF THE JTH VARIABLE (ITS SUPPORT), AND LET $s_j \subseteq S_j$ BE A SUBSET OF THESE VALUES.

GOAL: FIND SUBSETS $s_1, \dots, s_p$, SUCH THAT PROBABILITY

$$Pr\left[\bigcap_{j=1}^{p} (X_j \in s_j)\right]$$

IS RELATIVELY LARGE.

$\bigcap_{j=1}^{p} (X_j \in s_j)$ IS CALLED A CONJUNCTIVE RULE.

IF THE SUBSET $s_j = S_j$, THE VARIABLE X_j IS SAID NOT TO APPEAR IN THE RULE.

MARKET BASKET ANALYSIS

WE SUPPOSE VERY LARGE DATA BASES $(p \approx 10^4, N \approx 10^8)$

SIMPLIFICATIONS:

1. $$s_j = \begin{cases} \{v_{0j}\} & \text{... SINGLE VALUE} \\ S_j & \text{... ENTIRE SET} \end{cases}$$

$\Rightarrow$ GOAL IS TO FIND $J \subseteq \{1, \dots, p\}$ AND $v_{0j}, j \in J$:

$$Pr\left[\bigcap_{j \in J} (X_j = v_{0j})\right] \text{ IS LARGE}$$

2. WE CAN APPLY THE TECHNIQUE OF DUMMY VARIABLES $\Rightarrow$ WE GET A PROBLEM INVOLVING ONLY BINARY-VALUED VARIABLES.

THE TOTAL NUMBER OF DUMMY VARIABLES $K = \sum_{j=1}^{p} |S_j|$

$\Rightarrow$ GOAL IS TO FIND $\mathcal{K} \subseteq \{1, \dots, K\} \equiv$ ITEM SET, S.T.:

$$Pr\left[\bigcap_{k \in \mathcal{K}} (Z_k = 1)\right] = Pr\left[\prod_{k \in \mathcal{K}} Z_k = 1\right] \text{ IS LARGE.}$$

EXAM: NAIL029 STROJOVÉ UČENÍ

THE ESTIMATED VALUE FOR THE ITEM SET $\mathcal{K}$:

$$\hat{Pr}\left[\prod_{k\in\mathcal{K}}(Z_k=1)\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\mathcal{K}} z_{ik}$$

THIS IS CALLED THE "SUPPORT" OR "PREVALENCE" $T(\mathcal{K})$ OF THE ITEM SET $\mathcal{K}$.

A LOWER SUPPORT BOUND t IS SPECIFIED, AND WE SEEK ALL ITEM SETS $\mathcal{K}_l$, S.T. $T(\mathcal{K}_l) > t$.

THE APRIORI ALGORITHM

- THE CARDINALITY $|\{\mathcal{K} \mid T(\mathcal{K}) > t\}|$ IS RELATIVELY SMALL
- $\mathcal{L} \subseteq \mathcal{K} \Rightarrow T(\mathcal{L}) \geq T(\mathcal{K})$.

THE FIRST PASS OVER THE DATA COMPUTES THE SUPPORT OF ALL SINGLE-ITEM SETS. THOSE WHOSE SUPPORT IS LESS THAN THE THRESHOLD ARE DISCARDED. THE SECOND PASS COMPUTES THE SUPPORT OF ALL ITEM SETS OF SIZE TWO THAT CAN BE FORMED FROM PAIRS OF THE SINGLE ITEMS SURVIVING THE FIRST PASS. EACH SUCCESSIVE PASS OVER THE DATA CONSIDERS ONLY THOSE ITEM SETS THAT CAN BE FORMED BY COMBINING THOSE THAT SURVIVED THE PREVIOUS PASS WITH THOSE RETAINED FROM THE FIRST PASS.

THE APRIORI ALGORITHM REQUIRES ONLY ONE PASS OVER THE DATA FOR EACH VALUE OF $|\mathcal{K}|$.

EACH HIGH SUPPORT ITEM SET $\mathcal{K}$ RETURNED BY THE APRIORI ALGORITHM IS CAST INTO A SET OF "ASSOCIATION RULES". THE ITEMS Z_k, $k \in \mathcal{K}$ ARE PARTITIONED: $A \dot{\cup} B = \mathcal{K}$, AND WRITE

$$\underset{\text{ANTECEDENT}}{A} \Rightarrow \underset{\text{CONSEQUENT}}{B}$$

THE SUPPORT OF $A \Rightarrow B$: $T(A \Rightarrow B) = T(\mathcal{K})$

THE CONFIDENCE (PREDICTABILITY) : $C(A \Rightarrow B) = \dfrac{T(A \Rightarrow B)}{T(A)}$

THE EXPECTED CONFIDENCE : $T(B)$

THE LIFT : $L(A \Rightarrow B) = \dfrac{C(A \Rightarrow B)}{T(B)}$

A CONFIDENCE THRESHOLD c IS SET, AND ALL RULES THAT CAN BE FORMED FROM COMPUTED $\mathcal{K}_\ell$, $T(\mathcal{K}_\ell) > t$ WITH CONFIDENCE GREATER THEN c :

$$\{ A \Rightarrow B \mid C(A \Rightarrow B) > c \}$$

ARE REPORTED.

THE OUTPUT OF THE ENTIRE ANALYSIS IS A COLLECTION OF ASSOCIATION RULES THAT SATISFY :

$$T(A \Rightarrow B) > t \quad \text{AND} \quad C(A \Rightarrow B) > c$$

ASSOCIATION RULES ARE AMONG DATA MINING'S BIGGEST SUCCESSES.

3/7

24.01.2010 Su
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

UNSUPERVISED AS SUPERVISED LEARNING

LET $g(x)$ BE THE UNKNOWN DATA PROBABILITY DENSITY TO BE ESTIMATED, AND $g_0(x)$ BE A SPECIFIED PROBABILITY DENSITY FUNCTION USED FOR REFERENCE. FOR EXAMPLE, $g_0(x)$ MIGHT BE THE UNIFORM DENSITY OVER THE RANGE OF VARIABLES.

SUPPOSE $x_1, \dots, x_N \overset{i.i.d.}{\sim} g(x)$

A SAMPLE OF SIZE N_0 CAN BE DRAWN FROM $g_0(x)$ USING MONTE CARLO METHODS.

POOLING THESE TWO DATA SETS, AND ASSIGNING MASS MASS $w = N_0/(N+N_0)$ TO THOSE DRAWN FROM $g(x)$, AND $w_0 = N/(N+N_0)$ –״– FROM $g_0(x)$

RESULTS IN A RANDOM SAMPLE DRAWN FROM THE MIXTURE DENSITY $(g(x) + g_0(x))/2$.

IF ONE ASSIGNS THE VALUE $Y=1$ TO EACH SAMPLE POINT DRAWN FROM $g(x)$ AND $Y=0$ TO THOSE DRAWN FROM $g_0(x)$, THEN

$$\mu(x) = E(Y|x) = \frac{g(x)}{g(x)+g_0(x)}$$

CAN BE ESTIMATED BY SUPERVISED LEARNING USING THE COMBINED SAMPLE $(y_1, x_1), \dots (y_{N+N_0}, x_{N+N_0})$ AS TRAINING DATA. $\Rightarrow$ $\hat{g}(x) = g_0(x) \cdot \frac{\hat{\mu}(x)}{1-\hat{\mu}(x)}$.

CLUSTER ANALYSIS

≡ DATA SEGMENTATION, GROUPING OR SEGMENTING A COLLECTION OF OBJECTS INTO SUBSETS OR "CLUSTERS", SUCH THAT THOSE WITHIN EACH CLUSTER ARE MORE CLOSELY RELATED TO ONE ANOTHER THAN OBJECTS ASSIGNED TO DIFFERENT CLUSTERS.

PROXIMITY MATRICES

$N \times N$ MATRIX $\mathbb{D}$, WHERE N IS THE NUMBER OF OBJECTS

$d_{ii'}$ ... PROXIMITY BETWEEN THE iTH AND i'TH OBJECTS.

IF THE ORIGINAL DATA WERE COLLECTED AS SIMILARITIES, A SUITABLE MONOTONE - DECREASING FUNCTION CAN BE USED TO CONVERT THEM TO DISSIMILARITIES.

IF THE ORIGINAL MATRIX $\mathbb{D}$ IS NOT SYMMETRIC, IT CAN BE REPLACED BY $(\mathbb{D} + \mathbb{D}^T)/2$.

IN GENERAL, TRIANGLE INEQUALITY $d_{ii'} \leq d_{ik} + d_{ki'}$ IS NOT GUARANTEED.

DISSIMILARITIES BASED ON ATTRIBUTES

FIRST WE DEFINE A DISSIMILARITY $d_j(x_{ij}, x_{i'j})$ BETWEEN VALUES OF jTH ATTRIBUTE, AND THEN DEFINE

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} d_j(x_{ij}, x_{ij'}) \quad \text{(CAN BE WEIGHTED)}$$

AS THE DISSIMILARITY BETWEEN OBJECTS i AND i'.

MOST COMMON CHOICE: SQUARED DISTANCE:

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2.$$

24.01.2010 Su
PETER CERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

CLUSTERING ALGORITHMS

1. COMBINATORIAL ALGORITHMS - WORK DIRECTLY ON THE OBSERVED DATA WITH NO DIRECT REFERENCE TO AN UNDERLYING PROBABILITY MODEL.
2. MIXTURE MODELING - SUPPOSES THAT THE DATA IS AN i.i.d SAMPLE FROM SOME POPULATION DESCRIBED BY A PARAMETERIZED MODEL TAKEN TO BE A MIXTURE OF COMPONENT DENSITY FUNCTIONS.
3. MODE SEEKERS ("BUMP HUNTERS") - TAKE A NONPARAMETRIC PERSPECTIVE, ATTEMPTING TO DIRECTLY ESTIMATE DISTINCT MODES OF THE PROBABILITY DENSITY FUNCTION.

COMBINATORIAL ALGORITHMS

EACH OBSERVATION $i \in \{1, \dots, N\}$ IS ASSIGNED TO THE CLUSTER $C(i) \in \{1, \dots, K\}$ (ENCODER C)
WE SEEK THE PARTICULAR ENCODER $C^*(i)$ WHICH MINIMIZES A "LOSS" FUNCTION THAT CHARACTERIZES THE DEGREE TO WHICH THE CLUSTERING GOAL IS NOT MET.

NATURAL LOSS ("ENERGY")

$$\boxed{W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})}$$

$\equiv$ "WITHIN CLUSTER" POINT SCATTER

"BETWEEN - CLUSTER" POINT SCATTER :

$$\boxed{B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d_{ii'}}$$

"TOTAL" POINT SCATTER:

$$\boxed{T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{ii'} = W(C) + B(C)}$$

PRACTICAL CLUSTERING ALGORITHMS ARE ABLE TO EXAMINE ONLY A VERY SMALL FRACTION OF ALL POSSIBLE ENCODERS $k = C(i)$.

K-MEANS

ONE OF THE MOST POPULAR ITERATIVE DESCENT CLUSTERING METHODS. IT IS INTENDED FOR SITUATIONS IN WHICH ALL VARIABLES ARE OF THE QUANTITATIVE TYPE, AND SQUARED EUCLIDEAN DIST.:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \| x_i - x_{i'} \|^2$$

WITHIN-POINT SCATTER :

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} \| x_i - x_{i'} \|^2 =$$

$$\boxed{= \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| x_i - \bar{x}_k \|^2 ,}$$

WHERE $\bar{x}_k = (\bar{x}_{1k}, \dots, \bar{x}_{pk})$ IS THE MEAN VECTOR ASSOCIATED WITH THE kTH CLUSTER,

AND $N_k = \sum_{i=1}^{N} I(C(i) = k)$.

WE SEEK :

$$C^* = \min_{C} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| x_i - \bar{x}_k \|^2 .$$

EXAM: NAIL 029 STROJOVÉ UČENÍ

NOTE THAT $\bar{x}_S = \arg\min_m \sum_{i \in S} \| x_i - m \|^2$.

HENCE WE OBTAIN C^* BY SOLVING THE ENLARGED OPTIMIZATION PROBLEM:

$$\boxed{\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} \| x_i - m_k \|^2} \qquad (*)$$

K-MEANS CLUSTERING

1. FOR A GIVEN CLUSTER ASSIGNMENT C, MINIMIZE THE TOTAL CLUSTER VARIANCE (*) WITH RESPECT TO $\{m_1, \dots, m_K\}$ $O(N)$
2. GIVEN A CURRENT SET OF MEANS $\{m_1, \dots, m_K\}$, (*) IS MINIMIZED BY ASSIGNING EACH OBSERVATION TO THE CLOSEST (CURRENT) CLUSTER MEAN, I.E.

$$C(i) \leftarrow \arg\min_{1 \le k \le K} \| x_i - m_k \|^2 \qquad O(KN)$$

3. ITERATE STEPS 1 AND 2 UNTIL THE ASSIGNMENTS DO NOT CHANGE.

VECTOR QUANTIZATION

THE K-MEANS CLUSTERING ALGORITHM REPRESENTS A KEY TOOL IN THE AREA OF IMAGE AND SIGNAL COMPRESSION, PARTICULARY IN VECTOR QUANTIZATION OR VQ.

FIRST BREAK THE IMAGE INTO SMALL BLOCKS, SAY 2x2 BLOCKS OF PIXELS, EACH REGARDED AS A VECTOR IN $\mathbb{R}^4$. A K-MEANS CLUSTERING ALGORITHM (LLOYD'S ALGORITHM IN THIS CONTEXT) IS RUN IN THIS SPACE. EACH OF THE BLOCK IS APPROXIMATED BY ITS CLOSEST CLUSTER CENTROID, KNOWN AS A CODEWORD.

THE COLLECTION OF CODEWORDS → CODEBOOK.

K-MEDOIDS

1. FOR A GIVEN CLUSTER ASSIGNMENT C FIND THE OBSERVATION IN THE CLUSTER MINIMIZING TOTAL DISTANCE TO OTHER POINTS IN THAT CLUSTER:

$$i_k^* = \underset{\{i : C(i)=k\}}{\arg\min} \sum_{C(i')=k} D(x_i, x_{i'}) \qquad O\left(\sum_{k=1}^{K} N_k^2\right)$$

THEN $m_k \leftarrow x_{i_k^*}$, $k = 1, \dots, K$.

2. GIVEN A CURRENT SET OF CLUSTER CENTERS $\{m_1, \dots, m_K\}$, MINIMIZE THE TOTAL ERROR BY ASSIGNING EACH OBSERVATION TO THE CLOSEST (CURRENT) CLUSTER CENTER:

$$C(i) \leftarrow \underset{1 \le k \le K}{\arg\min} \, D(x_i, m_k) \qquad O(KN)$$

3. ITERATE STEPS 1 AND 2 UNTIL THE ASSIGNMENTS DO NOT CHANGE.

THIS APPROACH CAN BE APPLIED TO DATA DESCRIBED ONLY BY PROXIMITY MATRICES. K-MEDOIDS IS FAR MORE COMPUT. INTENSIVE THAN K-MEANS.

6/7

24.01.2010 Su
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

A CHOICE FOR THE NUMBER OF CLUSTERS K DEPEND ON THE GOAL. DATA-BASED METHODS FOR ESTIMATING K^* TYPICALLY EXAMINE THE WITHIN-CLUSTER DISSIMILARITY W_K AS A FUNCTION OF THE NUMBER OF CLUSTERS K. THE CORRESPONDING VALUES $\{W_1, \dots, W_{K_{max}}\}$ GENERALLY DECREASE WITH INCREASING K (EVEN WHEN EVALUATED ON THE INDEPENDENT TEST SET). THE CROSS-VALIDATION TECHNIQUES CANNOT BE UTILIZED IN THIS CONTEXT.

GAB STATISTICS (TIBSHIRANI ET AL., 2001) - COMPARES THE CURVE $\log W_K$ TO THE CURVE OBTAINED FROM DATA UNIFORMLY DISTRIBUTED OVER A RECTANGLE CONTAINING THE DATA. IT ESTIMATES THE OPTIMAL NUMBER OF CLUSTERS TO BE THE PLACE WHERE THE GAP BETWEEN THE TWO CURVES IS LARGEST.

HIERARCHICAL CLUSTERING

FIRST WE SPECIFY THE MEASURE OF DISSIMILARITY BETWEEN (DISJOINT) GROUPS OF OBSERVATIONS. THESE METHODS PRODUCE HIERARCHICAL REPRESENTATIONS IN WHICH THE CLUSTERS AT EACH LEVEL OF THE HIERARCHY ARE CREATED BY MERGING CLUSTERS AT THE NEXT LOWER LEVEL. AT THE LOWEST LEVEL, EACH CLUSTER CONTAINS A SINGLE OBSERVATION. AT THE HIGHEST LEVEL THERE IS ONLY ONE CLUSTER CONTAINING ALL OF THE DATA.

STRATEGIES $\begin{cases} \nearrow \text{AGGLOMERATIVE (BOTTOM-UP)} \\ \searrow \text{DIVISIVE (TOP-DOWN)} \end{cases}$

AGGLOMERATIVE - THE PAIR CHOSEN FOR MERGING CONSISTS OF THE TWO GROUPS WITH THE SMALLEST INTERGROUP DISSIMILARITY.

DIVISIVE - A CLUSTER AND SPLIT ARE CHOSEN TO PRODUCE TWO NEW GROUPS WITH THE LARGEST BETWEEN-GROUP DISSIMILARITY

... THERE ARE N-1 LEVELS IN THE HIERARCHY

GAP STATISTIC CAN BE USED TO CHOOSE THE LEVEL.

RECURSIVE BINARY SPLITTING/ AGGLOMERATION CAN BE REPRESENTED BY A ROOTED BINARY TREE. ALL AGGLOMERATIVE AND SOME DIVISIVE METHODS POSSESS A MONOTONICITY PROPERTY - THE DISSIMILARITY BETWEEN MERGED CLUSTERS IS MONOTONE INCREASING WITH THE LEVEL. ⇒ THE BINARY TREE CAN BE PLOTTED SO THAT THE HEIGHT OF EACH NODE IS PROPORTIONAL TO THE VALUE OF THE INTERGROUP DISSIMILARITY BETWEEN ITS TWO SONS. - A SO-CALLED DENDROGRAM.

AGGLOMERATIVE CLUSTERING

BEGINS WITH EVERY OBSERVATION REPRESENTING A SINGLE CLUSTER. AT EACH OF THE N-1 STEPS THE CLOSEST TWO (LEAST DISSIMILAR) CLUSTERS ARE MERGED INTO A SINGLE CLUSTER:

A MEASURE OF DISSIMILARITY BETWEEN TWO CLUSTERS MUST BE DEFINED.

7/7

24.01.2010 Su
PETER CERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

a) SINGLE LINKAGE (SL):

$$d_{SL}(G,H) = \min_{i \in G \; i' \in H} d_{ii'}$$

NEAREST-NEIGHBOR TECHNIQUE

b) COMPLETE LINKAGE (CL):

$$d_{CL}(G,H) = \max_{i \in G \; i' \in H} d_{ii'}$$

FURTHEST-NEIGHBOR TECHNIQUE

c) GROUP AVERAGE (GA):

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

SINGLE LINKAGE - A PHENOMENON, REFFERED TO AS <u>CHAINING</u>, IS OFTEN CONSIDERED A DEFECT OF THIS METHOD → CAN VIOLATE "COMPACTNESS" PROPERTY (≡ OBS. WITHIN EACH CLUSTER TEND TO BE SIMILAR)

COMPLETE LINKAGE - IT CAN PRODUCE CLUSTERS THAT VIOLATE THE "CLOSENESS" PROPERTY, I.E. OBSERVATIONS ASSIGNED TO A CLUSTER CAN BE MUCH CLOSER TO MEMBERS OF OTHER CLUSTERS.

GROUP AVERAGE - IS NOT INVARIANT TO MONOTONE STRICTLY INCREASING TRANSFORMATIONS OF OBSERVATION DISSIMILARITIES

DIVISIVE CLUSTERING

BEGINS WITH THE ENTIRE DATA SET AS A SINGLE CLUSTER, AND RECURSIVELY DIVIDE ONE OF THE EXISTING CL. INTO TWO DAUGHTER CLUSTERS.

ALGORITHM PROPOSED BY MACNAUGHTON SMITH ET AL. : FIRST PLACE ALL OBSERVATIONS IN A SINGLE CLUSTER G. THEN CHOOSE THE OBSERVATION WHOSE AVERAGE DISSIMILARITY FROM ALL THE OTHER OBSERVATIONS IS LARGEST. THIS OBSERVATION FORMS THE FIRST MEMBER OF A SECOND CLUSTER H. WHILE THERE ARE OBSERVATIONS IN G THAT ARE, ON AVERAGE, CLOSER TO H, TRANSFER TO H SUCH OBSERVATION FOR WHICH THE CORRESPONDING DIFFERENCE IN AVERAGES IS THE LARGEST ONE. THE RESULT IS A SPLIT OF THE ORIGINAL CLUSTER INTO TWO DAUGHTER CLUSTERS. ⇝ SECOND LEVEL ... EACH SUCCESSIVE LEVEL IS PRODUCED BY APPLYING THIS SPLITTING PROCEDURE TO ONE OF THE CLUSTERS AT THE PREVIOUS LEVEL. KAUFMANN AND ROUSSEEUW (1990) SUGGEST CHOOSING THE CLUSTER WITH THE LARGEST DIAMETER. AN ALTERNATIVE WOULD BE TO CHOOSE THE ONE WITH THE LARGEST AVERAGE DISSIMILARITY $\bar{d}_G = \frac{1}{N_G} \sum_{i \in G} \sum_{i' \in G} d_{ii'}$.

25.01.2010 Mo
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

SELECTED TOPICS

MINIMUM DESCRIPTION LENGTH

MDL APPROACH GIVES A SELECTION CRITERION FORMALLY IDENTICAL TO THE BIC APPROACH, BUT IS MOTIVATED FROM AN OPTIMAL CODING VIEWPOINT. SUPPOSE FIRST THAT WE WANT TO TRANSMIT POSSIBLE MESSAGES $z_1, \dots, z_m$. OUR CODE USES A FINITE ALPHABET, I.E. $A = \{0, 1\}$.

EXAMPLE:

MESSAGE	z_1	z_2	z_3	z_4
CODE	0	10	110	111

A SO-CALLED INSTANTENOUS PREFIX CODE: NO CODE IS PREFIX OF ANY OTHER.

STRATEGY - SHORTER CODES FOR MORE FREQUENT MESSAGES

FAMOUS THEOREM DUE TO SHANNON: IF MESSAGES ARE SENT WITH PROBABILITIES $Pr(z_i)$, WE SHOULD USE CODE LENGTHS $\boxed{l_i = -\log_2 Pr(z_i)}$, AND THE AVERAGE MESSAGE LENGTH SATISFIES:

$$\boxed{E(\text{length}) \geq -\sum Pr(z_i) \log_2 (Pr(z_i))}$$

RIGHT-HAND SIDE ≡ THE ENTROPY OF THE DISTR. $Pr(z_i)$.

IN GENERAL, THE LOWER BOUND CANNOT BE ACHIEVED, HUFFMANN CODING SCHEME CAN GET CLOSE TO THE BOUND.

NOW WE APPLY THIS RESULT TO THE PROBLEM OF MODEL SELECTION. WE HAVE A MODEL M WITH PARAMETERS θ, AND DATA $Z = (X, Y)$ CONSISTING OF BOTH INPUTS AND OUTPUTS. LET THE CONDITIONAL PROBABILITY OF THE OUTPUTS UNDER THE MODEL BE $Pr(Y | \theta, M, X)$. ASSUME THE RECEIVER KNOWS ALL OF THE INPUTS, AND WE WISH TO TRANSMIT THE OUTPUTS. THEN THE MESSAGE LENGTH REQUIRED TO TRANSMIT THE OUTPUTS IS:

$$\text{length} = -\log Pr(Y | \theta, M, X) - \log Pr(\theta | M), \qquad (*)$$

THE LOG-PROBABILITY OF THE TARGET VALUES GIVEN THE INPUTS. THE SECOND TERM IS THE AVERAGE CODE LENGTH FOR TRANSMITTING THE MODEL PARAMETERS θ.

THE MDL PRINCIPLE SAYS THAT WE SHOULD CHOOSE THE MODEL THAT MINIMIZES $(*)$.

WE RECOGNIZE $(*)$ AS THE NEGATIVE LOG-POSTERIOR DISTRIBUTION, AND HENCE MINIMIZING DESCRIPTION LENGTH IS EQUIVALENT TO MAXIMIZING POSTERIOR PROBABILITY. HENCE THE BIC CRITERION, DERIVED AS APPROXIMATION TO LOG-POSTERIOR PROBABILITY, CAN ALSO BE VIEWED AS A DEVICE FOR MODEL CHOICE BY MDL.

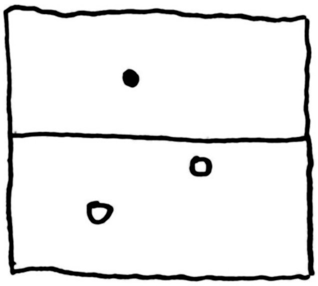
EXAM: NAIL029 STROJOVÉ UČENÍ

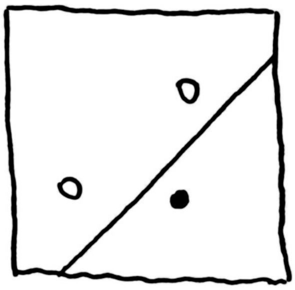
VAPNIK - CHERVONENKIS DIMENSION

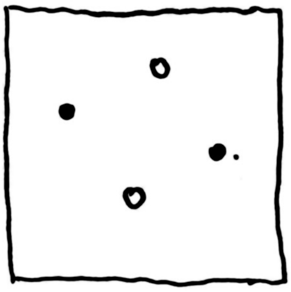
A DIFFICULTY IN USING ESTIMATES OF IN-SAMPLE ERROR IS THE NEED TO SPECIFY THE NUMBER OF PARAMETERS (OR THE COMPLEXITY) d USED IN THE FIT. THE VAPNIK-CHERVONENKIS (VC) THEORY PROVIDES SUCH A GENERAL MEASURE OF COMPLEXITY, AND GIVES ASSOCIATED BOUNDS ON THE OPTIMISM.

SUPPOSE WE HAVE A CLASS OF FUNCTIONS $\{f(x, \alpha)\}$ INDEXED BY A PARAMETER VECTOR α, WITH $x \in \mathbb{R}^p$. FOR INSTANCE, IF $\alpha = (\alpha_0, \alpha_1)$ AND f IS AN INDICATOR FUNCTION $I(\alpha_0 + \alpha_1^T x > 0)$, THEN IT SEEMS REASONABLE THAT THE COMPLEXITY IS THE NUMBER OF PARAMETERS $p+1$.

THE VC DIMENSION OF THE CLASS $\{f(x, \alpha)\}$ IS DEFINED TO BE THE LARGEST NUMBER OF POINTS (IN SOME CONFIGURATION) THAT CAN BE SHATTERED BY MEMBERS OF $\{f(x, \alpha)\}$.







A SET OF POINTS IS SAID TO BE SHATTERED BY A CLASS OF FUNCTIONS IF, NO MATTER HOW WE ASSIGN A BINARY LABEL TO EACH POINT, A MEMBER OF THE CLASS CAN PERFECTLY SEPARATE THEM.

THE VC DIMENSION OF A CLASS OF REAL-VALUED FUNCTIONS $\{g(x,\alpha)\}$ IS DEFINED TO BE THE VC DIMENSION OF THE INDICATOR CLASS $\{I(g(x,\alpha)-\beta>0)\}$, WHERE β TAKES VALUES OVER THE RANGE OF g.

IF WE FIT N TRAINING POINTS USING A CLASS OF FUNCTIONS $\{f(x,\alpha)\}$ HAVING VC DIMENSION h, THEN WITH PROBABILITY AT LEAST $1-\eta$ OVER TRAINING SETS:

$$\mathrm{Err}_{\mathcal{T}} \le \overline{\mathrm{err}} + \frac{\varepsilon}{2}\left(1+\sqrt{1+\frac{4\,\overline{\mathrm{err}}}{\varepsilon}}\right) \quad \left(\begin{array}{c}\text{BINARY}\\ \text{CLASSIFICATION}\end{array}\right)$$

$$\mathrm{Err}_{\mathcal{T}} \le \overline{\mathrm{err}} \Big/ \left(1-c\sqrt{\varepsilon}\right)_+ \quad (\text{REGRESSION})$$

WHERE $\varepsilon = a_1 \dfrac{h\left[\log(a_2 N/h)+1\right]-\log(\eta/4)}{N}$

$0 < a_1 \le 4, \; 0 < a_2 \le 2$

$a_1 = 4, \; a_2 = 2 \; \equiv$ WORST-CASE

EXAM: NAIL029 STROJOVÉ UČENÍ

VERSION SPACE

A VERSION SPACE IN CONCEPT LEARNING OR INDUCTION IS THE SUBSET OF ALL HYPOTHESES THAT ARE CONSISTENT WITH THE OBSERVED TRAINING EXAMPLES.

IN SETTINGS WHERE THERE IS A GENERALITY-ORDERING ON HYPOTHESES, IT IS POSSIBLE TO REPRESENT THE VERSION SPACE BY TWO SETS OF HYPOTHESES:

(1) THE MOST SPECIFIC CONSISTENT HYPOTHESES (I.E. THE SPECIFIC BOUNDARY SB) - COVER THE OBSERVED POSITIVE TRAINING EXAMPLES, AND AS LITTLE OF THE REMAINING FEATURE SPACE AS POSSIBLE.

(2) THE MOST GENERAL HYPOTHESES (I.E. THE GENERAL BOUNDARY GB) - COVER THE OBSERVED POSITIVE TRAINING EXAMPLES, BUT ALSO COVER AS MUCH OF THE REMAINING FEATURE SPACE WITHOUT INCLUDING ANY NEGATIVE TR. EXAMPLES.

HYPOTHESIS - CONJUNCTION OF TESTS ON INPUT ATTRIBUTES $\langle ?, \text{COLD}, \text{HIGH}, ?, ?, ? \rangle$

PARTIAL ORDERING - $\underset{\text{MORE GENERAL HYPOTHESIS}}{h_1} >_g \underset{\text{MORE SPECIFIC HYPOTHESIS}}{h_2}$

MOST GENERAL HYP. = $\langle ?, ?, \ldots, ? \rangle$
MOST SPECIFIC HYP. = $\langle 0, 0, \ldots, 0 \rangle$

H ... THE SPACE OF ALL HYPOTHESES
D ... TRAINING DATA
$VS_{H,D}$ ... VERSION SPACE

$VS_{H,D} = \{ h \in H \mid \text{Consistent}(h, D) \}$

GENERAL BOUNDARY : $G = \{ g \in VS_{H,D} \mid \neg \exists g' \in VS_{H,D} : g' >_g g \}$

SPECIFIC BOUNDARY : $S = \{ g \in VS_{H,D} \mid \neg \exists s' \in VS_{H,D} : s >_g s' \}$

<u>FIND-S</u> ... FINDS ONE MAX. SPECIFIC HYPOTHESIS

```
1. $h \leftarrow \langle 0, \ldots, 0 \rangle$
2. FOR EACH POSITIVE TRAINING SAMPLE $x \in D$
   RELAX h TO CLOSEST MORE GENERAL
   HYPOTHESIS CONSISTENT WITH x
```

<u>CANDIDATE - ELIMINATION</u> ... FINDS G AND S

```
1. $G \leftarrow$ MAXIMAL GENERAL HYPOTHESES IN H
2. $S \leftarrow$ MAXIMAL SPECIFIC HYPOTHESES IN H
3. FOR EACH TRAINING EXAMPLE $d \in D$ DO
4.    IF d IS A POSITIVE EXAMPLE :
        a) REMOVE FROM G ANY HYPOTHESIS
           INCONSISTENT WITH d
        b) FOR EACH HYP. s IN S NOT CONSISTENT WITH d
```

- REMOVE s FROM S
- ADD TO S ALL MINIMAL GENERALIZATIONS h OF s SUCH THAT h IS CONSISTENT WITH d, AND SOME MEMBER OF G IS MORE GENERAL THAN h
- REMOVE FROM S ANY HYPOTHESIS THAT IS MORE GENERAL THAN ANOTHER HYPOTHESIS IN S

5. IF d IS A NEGATIVE EXAMPLE:
   a) REMOVE FROM S ANY HYPOTHESIS INCONSISTENT WITH d
   b) FOR EACH HYP. g IN G NOT CONSISTENT WITH d
   - REMOVE g FROM G
   - ADD TO G ALL MINIMAL SPECIALIZATIONS h OF g SUCH THAT h IS CONSISTENT WITH d, AND SOME MEMBER OF S IS MORE SPECIFIC THAN h
   - REMOVE FROM G ANY HYPOTHESIS THAT IS LESS GENERAL THAN ANOTHER HYPOTHESIS IN G

PAC - LEARNING

PROBABLY APPROXIMATELY CORRECT LEARNING (PAC LEARNING) IS A FRAMEWORK FOR MATHEMATICAL ANALYSIS OF MACHINE LEARNING. (LESLIE VALIANT '84)

THE LEARNER RECEIVES SAMPLES AND MUST SELECT A GENERALIZATION FUNCTION (HYPOTHESIS) FROM A CERTAIN CLASS OF FUNCTIONS.

GOAL: WITH HIGH PROBABILITY (PROBABLE PART) SELECT A FUNCTION WITH LOW GENERALIZATION ERROR (APPROXIMATELY CORRECT PART), GIVEN ANY ARBITRARY APPROXIMATION RATIO ε, PROBABILITY OF SUCCESS δ, OR DISTRIBUTION OF SAMP. D

X ... INSTANT SPACE, ENCODING OF ALL THE SAMPLES

$c \subseteq X$ ... CONCEPT, $C \subseteq \mathcal{P}(X)$ ... CONCEPT CLASS

$EX(c, D)$ ... PROCEDURE THAT DRAWS AN EXAMPLE x USING A PROBABILITY DISTRIBUTION D, AND GIVES THE CORRECT LABEL $c(x) = \begin{cases} 1 & x \in c \\ 0 & x \notin c \end{cases}$

SUPPOSE THERE IS AN ALGORITHM A THAT GIVEN ACCESS TO $EX(c, D)$ AND INPUTS ε AND δ THAT, WITH PROBABILITY AT LEAST $1-\delta$, A OUTPUTS A HYPOTHESIS $h \in C$ THAT HAS ERROR $\leq \varepsilon$ WITH EXAMPLES DRAWN FROM X WITH DISTR. D.

IF THERE IS SUCH ALGORITHM FOR EVERY CONCEPT $c \in C$, EVERY DISTRIBUTION D OVER X, AND FOR ALL $0 < \varepsilon < 1/2$, $0 < \delta < 1/2$, THEN C IS PAC LEARNABLE.

CONSIDER A CONCEPT CLASS C AND A FIXED GOAL CONCEPT $c \in C$.

FOR ANY HYPOTHESIS $h \in C$ LET US DEFINE

$$\text{True Error}(h) = \Pr\left[h(x) \neq c(x) \mid x \text{ FROM } EX(c, D) \right]$$

5/10

25.01.2010 Mo
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

HYPOTHESIS h IS APPROXIMATELY CORRECT, IF
$\text{True Error}(h) \leq \varepsilon$

SUPPOSE WE HAVE A TRAINING SET $\mathcal{X}$, $|\mathcal{X}| = m$

BAD HYPOTHESES :

$$H_{BAD} = \{ h_b \in C \mid h_b \text{ IS CONSISTENT WITH } \mathcal{X} \text{ AND } \text{True Error}(h_b) > \varepsilon \}$$

SUPPOSE $h \in C$, $\text{True Error}(h) > \varepsilon$. IF WE DRAW x FROM $EX(c, D)$, THEN :

$$\Pr[h(x) = c(x)] \leq 1 - \varepsilon$$

SINCE $\mathcal{X}$ IS DRAWN FROM $EX(c, D)$ m-TIMES, THE PROBABILITY THAT h IS CONSISTENT WITH WHOLE $\mathcal{X}$ IS $\leq (1-\varepsilon)^m$, I.E.

$$\Pr[h \in H_{BAD} \mid \text{True Error}(h) \overset{h \in C}{>} \varepsilon, |\mathcal{X}| = m] \leq (1-\varepsilon)^m$$

THUS THE EXPECTED NUMBER OF UNFILTERED BAD HYPOTHESES WITH OUR TRAINING SET $\mathcal{X}$ IS:

$$|H_{BAD}| \leq |C| \cdot (1-\varepsilon)^m \leq |C| \cdot e^{-\varepsilon m}$$

WE WANT $|H_{BAD}| \leq |C| \cdot \delta$.

THUS $m \geq -\frac{1}{\varepsilon} \ln \delta = \frac{1}{\varepsilon} \ln \frac{1}{\delta}$

INSTANCE BASED LEARNING (IBL)

CONSISTS OF SIMPLY STORING THE PRESENTED TRAINING DATA. TO CLASSIFY A NEW QUERY INSTANCE FIND A SET OF SIMILAR RELATED INSTANCES.

k-NEAREST NEIGHBORHOOD LEARNING

METRIC:

- EUCLIDEAN: $d(x_i, x_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$
- HAMMING (MANHATTAN): $d(x_i, x_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$
- OVERLAP: $d(x_i, x_j) = \sum_{k=1}^{p} (1 - \delta(x_{ik}, x_{jk}))$
- COSINE: $d(x_i, x_j) = \dfrac{\sum_{k=1}^{p} x_{ik} \, x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik} \cdot x_{ik}} \sqrt{\sum_{k=1}^{p} x_{jk} \, x_{jk}}}$

THE PRIMARY OUTPUT OF IBL ALGORITHMS IS A <u>CONCEPT DESCRIPTION</u> (OR CONCEPT). THIS IS A FUNCTION THAT MAPS INSTANCES TO CATEGORIES → CLASSIFICATION.

AN INSTANCE-BASED CONCEPT DESCRIPTION INCLUDES A SET OF STORED INSTANCES AND, POSSIBLY, SOME INFORMATION CONCERNING THEIR PAST PERFORMANCE DURING CLASSIFICATION (E.G., THEIR NUMBER OF CORRECT AND INCORRECT CLASSIFICATION PREDICTIONS). THIS SET OF INSTANCES CAN CHANGE AFTER EACH

6/10

25.01.2010 Po
PETER ČERNO

EXAM: NAIL 029 STROJOVÉ UČENÍ

TRAINING INSTANCE IS PROCESSED. HOWEVER, IBL ALGORITHMS DO NOT CONSTRUCT EXTENSIONAL CONCEPT DESCRIPTIONS. INSTEAD, CONCEPT DESCRIPTIONS ARE DETERMINED BY HOW THE IBL ALGORITHM'S SELECTED SIMILARITY AND CLASSIFICATION FUNCTIONS USE THE CURRENT SET OF SAVED INSTANCES.

1. SIMILARITY FUNCTION : SIMILARITY BETWEEN A TRAINING INSTANCE i AND THE INSTANCES IN THE CONCEPT DESCRIPTION (CD) → NUMERIC-VALUED.
2. CLASSIFICATION FUNCTION : IT YIELDS A CLASSIFICATION BASED ON THE CLASSIFICATION PERFORMANCE RECORDS OF THE INSTANCES IN THE CD.
3. CONCEPT DESCRIPTION UPDATER : THIS MAINTAINS RECORDS ON CLASSIFICATION PERFORMANCE AND DECIDES WHICH INSTANCES TO INCLUDE IN THE CD.

ALGORITHMS : IB1, IB2, IB3

IB1 IS THE SIMPLEST. IB2 CAN DRASTICALLY REDUCE IB1'S STORAGE REQUIREMENTS, BUT IS SENSITIVE TO THE AMOUNT OF NOISE PRESENT IN THE TRAINING SET.

WE DESCRIBE IB3, A NOISE-TOLERANT EXTENSION OF IB2 THAT EMPLOYS A SIMPLE SELECTIVE UTILIZATION FILTER TO DETERMINE WHICH OF THE SAVED INSTANCES SHOULD BE USED TO MAKE CLASSIFICATION DECISIONS.

IB3 ALGORITHM

```
CD ← $\emptyset$
FOR EACH x IN TRAINING SET DO
1.      FOR EACH $y \in CD$ DO
            Sim[y] ← Similarity(x, y)
2.      IF $\exists y \in CD$ : Acceptable(y) THEN
            $y_{MAX}$ ← SOME ACCEPTABLE $y \in CD$ WITH
                MAXIMAL Sim[y]
        ELSE   i ← RANDOMLY-SELECTED $\in \{1, \dots, |CD|\}$
            $y_{MAX}$ ← SOME $y \in CD$ ITH MOST SIMILAR TO x
3.      IF class(x) = class($y_{MAX}$) THEN
            classification ← correct
        ELSE   classification ← incorrect
            CD ← CD $\cup$ {x}
4.      FOR EACH $y \in CD$ DO
            IF Sim[y] $\geq$ Sim[$y_{MAX}$] THEN
                UPDATE y'S CLASSIFICATION RECORD
            IF y'S RECORD IS SIGNIFICANTLY POOR
                THEN CD ← CD − {y}
```

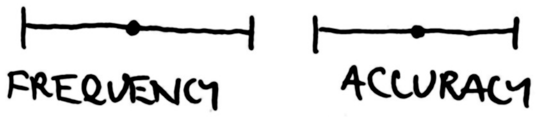
FOR EACH TRAINING INSTANCE t, CLASSIFICATION RECORDS ARE UPDATED FOR ALL SAVED INSTANCES THAT ARE AT LEAST AS SIMILAR AS t'S MOST SIMILAR ACCEPTABLE NEIGHBOR.

7/10

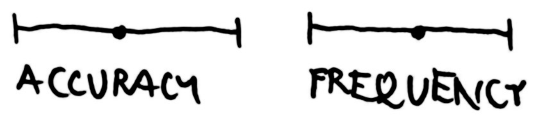
25.01.2010 Po
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

IB3 ACCEPTS AN INSTANCE (Acceptable(.)) IF ITS CLASSIFICATION ACCURACY IS SIGNIFICANTLY GREATER THAN ITS CLASS'S OBSERVED FREQUENCY AND REMOVES THE INSTANCE FROM THE CONCEPT DESCRIPTION (CD) IF ITS ACCURACY IS SIGNIFICANTLY LESS. CONFIDENCE INTERVALS ARE CONSTRUCTED AROUND BOTH THE INSTANCE'S CURRENT CLASSIFICATION ACCURACY (I.E., ITS PERCENTAGE OF CORRECT CLASSIFICATION ATTEMPTS) AND ITS CLASS'S CURRENT OBSERVED RELATIVE FREQUENCY (I.E., THE PERCENTAGE OF PROCESSED TRAINING INSTANCES THAT ARE MEMBERS OF THIS CLASS).



⇒ ACCEPTABLE INSTANCE



⇒ REMOVE THE INSTANCE

LET US DENOTE:

p ... TRUE CLASSIFICATION ACCURACY

s ... NUMBER OF CORRECT CLASSIFICATION ATTEMPTS

N ... TOTAL NUMBER OF CLASSIFICATION ATTEMPTS

APPARENTLY $s \sim \text{Binom}(N, p)$

CONFIDENCE INTERVALS

FOR LARGE N : $s \longrightarrow \mathcal{N}\left(Np,\ Np(1-p)\right)$

I.E. $\dfrac{\frac{s}{N} - p}{\sqrt{\frac{p(1-p)}{N}}} \sim N(0,1)$

OUR ESTIMATES : $\hat{p} = \frac{s}{N}$, $\hat{\sigma}^2 = \dfrac{\frac{s}{N}\left(1-\frac{s}{N}\right)}{N}$

CONFIDENCE INTERVAL FOR p WITH CONFIDENCE $1-\alpha$:

$$\hat{p} - z \cdot \hat{\sigma} \leq p \leq \hat{p} + z\hat{\sigma}, \quad z = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

WE USE 90% CONFIDENCE FOR ACCEPTANCE,

I.E. $z_{Acc} = \Phi^{-1}\left(1 - \frac{0{,}1}{2}\right) = \Phi^{-1}(0{,}95) \approx 1{,}645$

AND 75% CONFIDENCE FOR DROPPING,

I.E. $z_{Drop} = \Phi^{-1}\left(1 - \frac{0{,}25}{2}\right) \approx 1{,}15$.

FOR EACH CLASS j WE HAVE ESTIMATES

$$\hat{p}_j = \frac{N_j}{N}, \quad \hat{\sigma}_j^2 = \hat{p}_j(1-\hat{p}_j)/N$$

FOR EACH INSTANCE inst WE HAVE ESTIMATES

$$\hat{p}_{inst} = \frac{s_{inst}}{N_{inst}} \quad \hat{\sigma}_{inst}^2 = \hat{p}_{inst}(1-\hat{p}_{inst})/N$$

SUPPOSE THAT $class(inst) = j$.

WE ACCEPT inst IF $\overset{u_{apri}}{\overbrace{\hat{p}_j + z_{Acc}\hat{\sigma}_j}} < \overset{l_{inst}}{\overbrace{\hat{p}_{inst} - z_{Acc}\hat{\sigma}_{inst}}}$

WE DROP inst IF $\underset{u_{inst}}{\underbrace{\hat{p}_{inst} + z_{Drop}\hat{\sigma}_{inst}}} < \underset{l_{apri}}{\underbrace{\hat{p}_j - z_{Drop}\hat{\sigma}_j}}$

8/10

25.01.2010 r.
PETER ČERNO

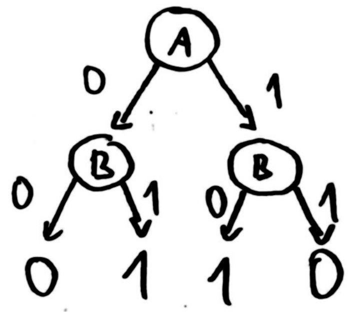
EXAM: NAIL029 STROJOVÉ UČENÍ

DECISION TREES

DECISION TREE LEARNING IS A METHOD FOR APPROXIMATING DISCRETE VALUE FUNCTIONS, IN WHICH THE LEARNED FUNCTION IS REPRESENTED BY A DECISION TREE. IT IS ONE OF THE MOST WIDELY USED APPROACH FOR INDUCTIVE INFERENCE.

INTERMEDIATE NODES : ATTRIBUTES
EDGES : ATTRIBUTE VALUES
LEAVE NODES : OUTPUT VALUES

F = A XOR B



CAN BE REWRITTEN AS IF-THEN-ELSE RULES

BASIC DECISION TREE LEARNING ALGORITHM:
<u>ID3 ALGORITHM</u> (QUINLAN 1986) AND IT'S SUCCESSORS C4.5 AND C5.0

GREEDY SEARCH THE SPACE OF POSSIBLE DECISION TREES

ID3 (Examples, Target-attribute, Attributes)

```
1. CREATE A Root NODE FOR THE TREE
2. IF ALL Examples ARE POSITIVE, RETURN
   THE SINGLE-NODED TREE Root, WITH LABEL = +
3. IF ALL Examples ARE NEGATIVE, RETURN
   THE SINGLE-NODED TREE Root, WITH LABEL = −
4. IF Attributes = $\emptyset$, RETURN THE SINGLE-NODE
   TREE Root, WITH LABEL = MOST COMMON VALUE
   OF Target-attribute IN Examples.
5. OTHERWISE BEGIN:
6. A ← THE ATTRIBUTE FROM Attributes THAT
          BEST CLASSIFIES Examples
7. THE DECISION ATTRIBUTE FOR Root ← A
8. FOR EACH POSSIBLE VALUE $v_i$ OF A:
     9. ADD A NEW TREE BRANCH BELOW Root
        CORRESPONDING TO THE TEST $A = v_i$
    10. LET $\text{Examples}_{v_i}$ BE THE SUBSET OF Examples
        THAT HAVE VALUE $v_i$ FOR A
    11. IF $\text{Examples}_{v_i} = \emptyset$ THEN
          BELOW THIS NEW BRANCH ADD A LEAF NODE
          WITH LABEL = MOST COMMON VALUE OF
          Target_attribute IN Examples
    12. ELSE BELOW THIS NEW ADD THE SUBTREE
          ID3 ($\text{Examples}_{v_i}$, Target-attribute, Attributes − A)
13. RETURN Root.
```

9/10

25.01.2010 Mo
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

WHICH ATTRIBUTE TO SELECT ?
ID3 USES INFORMATION GAIN MEASURE TO SELECT AMONG THE CANDIDATE ATTRIBUTES.

INFORMATION GAIN IS BASED ON ENTROPY.
SHANNON ENTROPY (INFORMATION ENTROPY) IS A MEASURE OF THE UNCERTAINTY ASSOCIATED WITH A RANDOM VARIABLE. IT QUANTIFIES THE INFORMATION CONTAINED IN A MESSAGE, USUALLY IN BITS OR BITS / SYMBOL. IT IS THE MINIMUM MESSAGE LENGTH NECESSARY TO COMMUNICATE INFORMATION.

THE INFORMATION ENTROPY OF A DISCRETE RANDOM VARIABLE $X \in \{1, \dots, x_n\}$ IS

$$H(X) = E(I(X)) = - \sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$I(X)$ ... INFORMATION CONTENT OF X

GIVEN A COLLECTION S, CONTAINING POSITIVE AND NEGATIVE SAMPLES :

$$\text{Entropy}(S) = - p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

THE INFORMATION GAIN IS THE EXPECTED REDUCTION IN ENTROPY CAUSED BY PARTITIONING THE EXAMPLES TO THE ATTRIBUTE A.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \underbrace{\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)}_{\text{ENTROPY OF } S \text{ AFTER PARTITION}}$$

$$S_v = \{ s \in S \mid A(s) = v \}$$

THE VALUE Gain(S,A) IS THE NUMBER OF BITS SAVED WHEN ENCODING THE TARGET VALUE OF AN ARBITRARY MEMBER OF S, BY KNOWING THE VALUE OF ATTRIBUTE A.

- ID3'S ALGORITHM SEARCHES COMPLETE HYPOTHESIS SPACE. ☺
- ID3 MAINTAIN ONLY A SINGLE CURRENT HYPOTHESIS AS IT SEARCHES THROUGH THE SPACE OF DECISION TREES. ☹
- ID3 CAN CONVERGE TO SUBOPTIMAL SOLUTIONS. ☹
- ID3 USES ALL TRAINING EXAMPLES AT EACH STEP IN THE SEARCH TO MAKE STATISTICALLY BASED DECISIONS REGARDING HOW TO REFINE ITS CURRENT HYPOTHESIS. ☺

CANDIDATE-ELIMINATION ALGORITHM IS LANGUAGE BIASED - HYPOTHESIS WAS ASSUMED TO BE CONJUNCTION OF ATTRIBUTES.

ID3 ALGORITHM HAS PREFERENCE / SEARCH BIAS - SELECTS TREES THAT PLACE THE ATTRIBUTES WITH HIGHEST INFORMATION GAIN CLOSEST TO THE ROOT.

<u>ISSUES</u> IN DECISION TREE LEARNING:

HOW DEEPLY TO GROW?
CONTINUOUS ATTRIBUTES?
CHOOSING AN ATTRIBUTE?
SELECTION MEASURE?
MISSING ATTRIBUTE VALUES?
DIFFERING ATTRIBUTE COSTS?

10/10

25.01.2010 Mo
PETER ČERNO

EXAM: NAIL 029 STROJOVÉ UČENÍ

OCCAM'S RAZOR

EXPLANATION OF ANY PHENOMENON SHOULD MAKE AS FEW ASSUMPTIONS AS POSSIBLE ...
"ALL OTHER THINGS BEING EQUAL, THE SIMPLEST SOLUTION IS THE BEST."

⇒ PREFER THE SIMPLEST HYPOTHESIS THAT FITS THE DATA
RAZOR - THE ACT OF SHAVING AWAY UNNECESSARY ASSUMPTIONS TO GET THE SIMPLEST EXPLANATION.

HOW TO AVOID OVERFITTING?

(1) STOP GROWING THE TREE EARLIER
(2) POST-PRUNING - MORE SUCCESSFUL IN PRACTICE

HOW TO DETERMINE CORRECT FINAL TREE SIZE?

(1) TRAINING AND VALIDATION
(2) APPLY STATISTICAL TESTS
(3) MINIMUM DESCRIPTION LENGTH PRINCIPLE (MDL)

PRUNING METHODS:

(1) REDUCED-ERROR PRUNING (QUINLAN 1987)
(2) RULE POST-PRUNING (QUINLAN 1993)

REDUCED ERROR PRUNING

PRUNING A DECISION NODE CONSISTS OF REMOVING THE SUBTREE ROOTED AT THAT NODE, MAKING IT A LEAF NODE, AND ASSIGNING IT THE MOST COMMON CLASSIFICATION OF THE TRAINING EXAMPLES AFFILIATED WITH THAT NODE.

NODES ARE REMOVED ONLY IF THE RESULTING PRUNED TREE PERFORMS NO WORSE THAN THE ORIGINAL OVER THE VALIDATION SET.

DRAWBACK : WHEN DATA IS LIMITED

RULE POST-PRUNING

1. INFER THE DECISION TREE FROM THE TRAINING SET, GROWING THE TREE UNTIL THE TRAINING DATA IS FIT AS WELL AS POSSIBLE, ALLOWING OVERFITTING TO OCCUR.
2. CONVERT THE LEARNED TREE INTO AN EQUIVALENT SET OF RULES.
3. PRUNE (GENERALIZE) EACH RULE BY REMOVING ANY PRECONDITIONS THAT RESULT IN IMPROVING ITS ESTIMATED ACCURACY.
4. SORT THE PRUNED RULES BY THEIR ESTIMATED ACCURACY.

ALTERNATIVE MEASURES FOR SELECTING ATTRIBUTES

GAIN RATIO - THE GAIN RATIO MEASURE PENALIZES ATTRIBUTES SUCH AS DATE BY INCORPORATING A TERM, CALLED SPLIT INFORMATION, THAT IS SENSITIVE TO HOW BROADLY AND UNIFORMLY THE ATTRIBUTE SPLITS THE DATA.

$$\text{Split Information}(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$\text{Gain Ratio}(S,A) = \frac{\text{Gain}(S,A)}{\text{Split Information}(S,A)}$$

1/1

26.01.2010 Tu
PETER ČERNO

EXAM: NAIL029 STROJOVÉ UČENÍ

HANDLING MISSING ATTRIBUTES

(1) ASSIGN IT A VALUE THAT IS MOST COMMON AMONG TRAINING EXAMPLES AT NODE n

(2) —||— THAT HAVE THE CLASSIFICATION c(x)

(3) ASSIGN A PROBABILITY TO EACH OF THE POSSIBLE VALUES OF A → USED IN C4.5

HANDLING ATTRIBUTES WITH DIFFERENT COST

WE WOULD PREFER DECISION TREES THAT USE LOW-COST ATTRIBUTES WHERE POSSIBLE, RELYING ON HIGH-COST ATTRIBUTES ONLY WHEN NEEDED TO PRODUCE RELIABLE CLASSIFICATIONS.

COST-SENSITIVE MEASURE : $Gain(S,A) / Cost(A)$

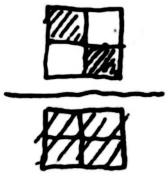
$$Gain^2(S,A) / Cost(A)$$

$$(2^{Gain(S,A)} - 1) / (Cost(A)+1)^w$$

MEASURING CREDIBILITY

		PREDICTED CLASS	
		CLASS 1	CLASS 2
ACTUAL CLASS	CLASS 1	TRUE POSITIVE	FALSE NEGATIVE
	CLASS 2	FALSE POSITIVE	TRUE NEGATIVE

$$\text{ACCURACY} = \frac{TP + TN}{TP + TN + FP + FN}$$



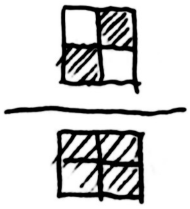
$\text{SENSITIVITY} = TP \,/\, (TP + FN)$

$\text{SPECIFICITY} = TN \,/\, (TN + FP)$

$\text{PRECISION} = TP \,/\, (TP + FP)$

RECALL = SENSITIVITY

$$\text{ERROR} = \frac{FP + FN}{TP + TN + FP + FN}$$



LEARNING CURVE :

X = N , Y = ERROR

LIFT CHART :

$Y = (TP + FP) / (TP + FP + TN + FN) \qquad X = TP$

ROC CURVE :

$$X = \frac{TP}{TP + FN} \qquad Y = \frac{FP}{FP + TN}$$

FP-RATE = 1- SPECIFICITY

TP-RATE = PRECISION

PRECISION RECALL CURVE :

$$X = \frac{TP}{TP + FN} \qquad Y = \frac{TP}{TP + FP}$$