MACHINE LEARNING AND PATTERN RECOGNITION

STIVEN HIDRI

1. DIMENSIONALITY REDUCTION

DEF: COMPUTE A MAPPING FROM A n-DIMENSIONAL SPACE TO A m DIMENSIONAL SPACE WITH $m \ll n$

GOALS: COMPRESS INFORMATION, REMOVE NOISE, DATA VISUALIZATION

SIMPLIFY CLASSIFICATION (REDUCE OVERFITTING, REDUCE EFFECTS OF HIGH DIMENSIONALITY)

For data visualization we want 2-D or 3-D representations which maintain the best the relationship between the samples, for improved classification we want to retain discriminant information, for compressing information we want to retain the maximum amount of information for a given output size.

2 METHODS
- PRINCIPAL COMPONENT ANALYSIS (PCA)
- LINEAR DISCRIMINANT ANALYSIS (LDA)

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA IS A DIMENSIONALITY REDUCTION TECHNIQUE WHICH CAN BE INTERPRETED AS A LINEAR MAPPING THAT PRESERVES THE DIRECTIONS WITH HIGHEST VARIANCE

IT IS UNSUPERVISED: NO GUARANTEES OF OBTAINING DISCRIMINANT DIRECTIONS

A DATASET $X = \{x_1, \dots x_k\}$ IS GIVEN WITH $x_i \in \mathbb{R}^n$

We want to find P, a subspace of $\mathbb{R}^n$ which preserves most of the information

$P \in \mathbb{R}^{n \times m}$ WHOSE COLUMNS ARE ORTHONORMAL

ITS COLUMNS FORM A BASIS OF A SUBSPACE OF $\mathbb{R}^n$ WITH DIMENSION m

IN ORDER TO DEFINE P WE TRY TO MINIMIZE THE AVERAGE RECONSTRUCTION ERROR:

$$\frac{1}{k}\sum_{1}^{k} \|x_i - \hat{x}_i\|^2 = \frac{1}{k}\sum_{1}^{k} \|x_i - PP^T x_i\|^2$$

P WILL REPRESENT A SUBSPACE WHOSE AXES PASS THROUGH THE ORIGIN.

IF THE DATASET IS NOT ZERO MEAN THE FIRST PCA DIRECTION WILL APPROXIMATELY CONNECT THE ORIGIN WITH THE DATA SET MEAN, NOT INTERESTING IN MANY CASES

SO WE WILL CENTER THE DATASET. NOW THE AVERAGE RECONSTRUCTION ERROR WILL BE:

$$\frac{1}{k}\sum_{1}^{k} \| z_i - PP^T z_i \|^2 \qquad z_i = x_i - \bar{x}$$

THEREFORE WE TRY TO SOLVE:

$$P^* = \underset{P}{\text{ARGMIN}} \ \frac{1}{k}\sum_{1}^{k} \| z_i - PP^T z_i \|^2$$

LET'S REWRITE THE OBJECTIVE FUNCTION:

$$L(P) = \frac{1}{k}\sum_{1}^{k} \| z_i - PP^T z_i \|^2 = \frac{1}{k}\sum_{1}^{k} z_i z_i^T - 2 z_i^T PP^T z_i + z_i^T PP^T PP^T z_i \quad (PP^T PP^T = I)$$

$$= \frac{1}{k}\sum_{1}^{k} z_i z_i^T - z_i^T PP^T z_i$$

$$= \frac{1}{k}\sum_{1}^{k} z_i z_i^T - T_R\left(P^T z_i z_i^T P\right)$$

$v^T w = T_R(v^T w)$

$T_R(AB) = T_R(BA)$

SINCE T_R IS A LINEAR OPERATOR, MINIMIZING L IS EQUIVALENT TO MAXIMIZE $\hat{L}$

$$\hat{L}(P) = T_R\left(P^T \left[\frac{1}{k}\sum_{1}^{k} z_i z_i^T\right] P\right)$$

IT CAN BE SHOWN THAT THE CORRECT SOLUTION IS GIVEN BY P WHOSE COLUMNS ARE THE m EIGENVECTORS OF $\frac{1}{k}\sum_{1}^{k} z_i z_i^T$ CORRESPONDING TO THE LARGEST m EIGENVALUES

THEN WE PROJECT OUR DATA OVER P^T. TO KEEP THE m DIRECTIONS WITH THE HIGHEST COVARIANCE WE KEEP THE FIRST V_m TRANFORMED DIRECTIONS

PROJECTION OVER U^T TRANSFORMS OUR DATA SO THAT THE DIFFERENT DIRECTIONS ARE UNCORRELATED

AN OPTIMAL m CAN BE FOUND THROUGH CROSSVALIDATION USING A VALIDATION SET

WE CAN ALSO SELECT m AS TO RETAIN A GIVEN PERCENTAGE f OF THE VARIANCE OF THE DATA

$$\min_m \ m \quad \text{S.T.} \quad \frac{\sum_{i}^{m} \sigma_i}{\sum_{i}^{m} \sigma_i} \geq f$$

LINEAR DISCRIMINANT ANALYSIS (LDA)

IT IS A DIMENSIONALITY REDUCTION TECHNIQUE THAT MAXIMIZES THE BETWEEN CLASS VARIABILITY OVER WITHIN-CLASS VARIABILITY RATIO FOR THE TRANSFORMED SAMPLES

$$\max_w \frac{w^T S_B w}{w^T S_W w}$$

"FIND A DIRECTION THAT HAS LARGE SEPARATION BETWEEN THE CLASSES AND SMALL SPREAD INSIDE EACH CLASS"

$$S_B = \frac{1}{N} \sum_{c=1}^{K} m_c (\mu_c - \mu)(\mu_c - \mu)^T$$

(m_c: # SAMPLES FOR EACH CLASS; μ: DATASET MEAN; μ_c: CLASS MEAN)

BETWEEN CLASS COVARIANCE: covariance matrix for the class means where each class is weighted by the corresponding sample size

$$S_W = \frac{1}{N} \sum_{c=1}^{K} \sum_{i=1}^{m_c} (x_{c,i} - \mu_c)(x_{c,i} - \mu_c)^T$$

(N: TOTAL SAMPLES)

WITHIN CLASS COVARIANCE: weighted average of the covariance matrix of each class

$$S_B + S_W = \frac{1}{N} \sum_{c=1}^{k} \sum_{i=1}^{m_c} (x_{c,i} - \mu)(x_{c,i} - \mu)^T$$

covariance matrix of the dataset as a whole

SINCE WE ARE LOOKING FOR A DISCRIMINANT DIRECTION w WE NOW CONSIDER THE BETWEEN AND WITHIN CLASS VARIANCE OF THE PROJECTED SAMPLES $w^T x$

$$\Lambda_B = w^T S_B w$$
$$\Lambda_W = w^T S_W w$$

THE CRITERION OF OPTIMALITY IS THE MAXIMIZATION OF THE RATIO OF BETWEEN AND WITHIN CLASS VARIANCE FOR THE PROJECTED POINTS (WE ASSUME S_W AS FULL RANK)

$$L(w) = \frac{\Lambda_B}{\Lambda_W} = \frac{w^T S_B w}{w^T S_W w}$$

OBJECTIVE FUNCTION

WE CAN SELECT A MAXIMIZER WITH UNIT FORM SINCE IF w IS A MAXIMIZER ALSO λw WILL BE A MAXIMIZER (THE CRITERION DOES NOT DEPEND ON THE SCALE OF w)

WE CAN GET AN OPTIMUM BY SOLVING $\nabla_w L(w) = 0$ (GRADIENT OF $L(w)$). WE GET:

$$S_W^{-1} S_B w = L(w) w$$

THE OPTIMAL SOLUTION IS AN EIGENVECTOR OF $S_W^{-1} S_B$ CORRESPONDING TO THE EIGENVALUE $L(w)$ WHICH IS THE RATIO THAT WE WANT TO MAXIMIZE

LDA AS LINEAR CLASSIFICATOR

THIS METHOD WAS ORIGINALLY USED TO SOLVE BINARY PROBLEMS

ONCE w IS ESTIMATED WE CAN PROJECT THE TEST SAMPLES OVER IT AND ASSIGN THE CLASS CHECKING WEATHER THE PROJECTED VALUES (SCORES) ARE $\geq$ OF A THRESHOLD t

$$C(x_t) = \begin{cases} C_1 & \text{if } w^T x_t \geq t \\ C_2 & \text{if } w^T x_t < t \end{cases}$$

BINARY CASE

FOR THE BINARY CASE THE EIGENVECTOR OF $S_W^{-1} S_B$ ASSOCIATED TO THE NON-ZERO EIGENVALUE IS:

$$w \propto S_W^{-1} (\mu_1 - \mu_0)$$

MULTIPLE DIRECTIONS

WE ARE INTERESTED IN THE m MOST DISCRIMINANT DIRECTIONS. WE DEFINE THESE AS W

$$\hat{S}_B = W^T S_B W$$
$$\hat{S}_W = W^T S_W W$$

DOES NOT HAVE TO BE ORTHOGONAL SINCE WE ARE NOT INTERESTED IN HOW THE DATA IS SCALED, BUT ONLY IN HOW IS DISTRIBUTED ALONG THE AXIS

ONE CRITERIA USED TO GENERALIZE THE 1-DIMENSIONAL CASE IS LOOKING FOR THE MAXIMIZER OF

$$L = Tr\left(S_W^{-1} S_B\right)$$

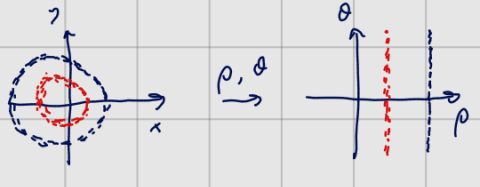
THE SOLUTION IS GIVEN BY THE m EIGENVECTORS CORRESPONDING TO THE m LARGEST EIGENVALUES OF $S_W^{-1} S_B$

$C \rightarrow$ NUMBER OF CLASSES!

LDA ALLOWS ESTIMATING AT MOST $C-1$ DIRECTION SINCE S_B, BY ITS DEFINITION, ALLOWS AT MOST $C-1$ NON-ZERO EIGENVALUES

COMMENTS

- LDA ASSUMES GAUSSIAN DISTRIBUTED NOISE
- IS OFTEN HELPFUL TO PRE PROCESS OUR DATA USING PCA BEFORE APPLYING LDA
- LINEAR TRANSFORMATIONS ARE NOT ALWAYS SUITED FOR THE DATA:



IN THIS CASE WHERE THE DATA IS SPREAD IN CIRCULAR SHAPE WE CAN USE POLAR COORDINATES TO REMAP THE DATA.

THEN WE CAN APPLY LDA

TO SOLVE THE GENERALIZED EIGENVALUE PROBLEM WE CAN ALSO USE A METHOD CALLED JOINT DIAGONALIZATION OF S_W AND S_B

① WHITEN S_W

1) EIGENVALUE DECOMPOSITION OF S_W → $S_W = U_W \Sigma_W U_W^T$ (RIGHT EIGENVECTORS OF S_W; DIAGONAL MATRIX WITH EIGENVALUES OF S_W)

2) APPLY WHITENING TRANSFORMATION TO THE DATASET → $P_W = U_W \Sigma_W^{-\frac{1}{2}} U_W^T$. $\tilde{x} = P_x$

3) $S_W \to I$, $S_B \to P S_B P^T$ → NOW ALL DATA IS EQUIDISTANT TO THE MEAN

② DIAGONALIZE S_B

1) COMPUTE EIGENVALUE DECOMPOSITION $S_B = P S_B P^T = U_B \Sigma_B U_B^T$

2) DIAGONALIZE BY PROJECTING OVER U_B^T

3) $S_W \to I$, $S_B \to \Sigma_B$

$S_W \to I$: IDENTITY MATRIX

$S_B \to \Sigma_B$: DIAGONAL MATRIX

THE FIRST m DIRECTIONS OF THE TRANSFORMED SAMPLES CORRESPOND TO THE LDA SUBSPACE

THE TRANSFORMATION APPLIED TO THE DATA CORRESPOND TO $W = P_W^T U_B$

GENERATIVE AND LINEAR QUADRATIC CLASSIFIERS I - GENERATIVE MODELS

CLOSED SET CLASSIFICATION PROBLEM

WE HAVE A SAMPLE x_t THAT WE WANT TO CLASSIFY AS BELONGING TO ONE OF K CLASSES

WE ASSUME x_t TO BE A REALIZATION OF A R.V. X_t

WE ASSUME ITS CLASS LABEL ALSO TO BE A REALIZATION OF A R.V. $C_t \in \{1 \dots K\}$

OPTIMAL BAYES DECISION: ASSUMING UNIFORM COSTS WE ASSIGN THE CLASS WITH HIGHEST POSTERIOR PROBABILITIES

$$c_t^* = \operatorname{argmax}_c P(C_t = c \mid X_t = x)$$

SO GIVEN A SAMPLE WE COMPUTE THE POSTERIOR PROBABILITIES FOR EACH CLASS AND ASSIGN THE LARGEST ONE

ASSUMPTION: $X_t, C_t \sim X, C \rightarrow f_{X_t, C_t}(x_t, c) = f_{X,C}(x_t, c)$

$$P(C_t = c, X_t = x_t) = \frac{f_{X,C}(x_t, c)}{\sum_{c' \in C} f_{X,C}(x_t, c')}$$

JOINT DISTRIBUTION

$$f_{X,C}(x_t, c) = f_{X|C}(x_t \mid c) P_C(c)$$

CLASS CONDITIONAL PROBABILITY

PRIOR PROBABILITY CLASS C, APPLICATION DEPENDENT

OBJECTIVE: MODEL CLASS CONDITIONAL DISTRIBUTION $f_{X|C}(x_t \mid c)$

ASSUMPTION: CLASS DATA CAN BE MODELED BY A MULTIVARIATE GAUSSIAN DISTRIBUTION **MVG**

$$(x_t \mid C_t = c) \sim (X \mid C = c) \sim \mathcal{N}(\mu_c, \Sigma_c)$$

WE DO NOT KNOW THE VALUES OF THE MODEL PARAMETERS $\theta = [(\mu_1, \Sigma_1) \dots (\mu_K, \Sigma_K)]$

BUT WE HAVE A LABELED SET OF TRAINING DATA $D = [(x_1, c_1) \dots (x_m, c_m)]$

WE WANT TO LEARN THE PARAMETERS FROM THE DATA

$$(X \mid C = c, \theta) \sim \mathcal{N}(\mu_c, \Sigma_c)$$

ASSUMPTION: GIVEN MODEL PARAMETERS OBSERVATIONS ARE IID, TRAINING AND EVALUATION SAMPLES ARE IID

CLASS CONDITIONAL DISTRIBUTION FOR ALL OBSERVATION IS A GAUSSIAN WITH BOTH CLASS DEPENDENT MEAN μ_c AND COVARIANCE MATRIX Σ_c

WE FOLLOW A FREQUENTIST APPROACH: COMPUTE AN ESTIMATOR θ^* OF THE MODEL PARAMETERS

$$f_{X_t|C_t}(x_t \mid c) \approx N(x_t \mid \mu_c^*, \Sigma_c^*)$$

ONCE WE HAVE θ^* WE CAN ESTIMATE $f_{X|C}$

ONE WAY TO ESTIMATE THE MODEL PARAMETERS IS TO MAXIMIZE THE DATA LOG-LIKELIHOOD

$$L(\theta) = f_{X_1 \dots X_m, C_1 \dots C_m}(x_1 \dots x_m, c_1 \dots c_m) =$$

$$= \prod_{i=1}^{m} f_{X,C|\theta}(x_i, c_i \mid \theta) = \prod_{i=1}^{m} f_{X|C,\theta}(x_i \mid c_i, \theta) P(c_i)$$

$$= \prod_{i=1}^{m} \mathcal{N}(x_i \mid \mu_{c_i}, \Sigma_{c_i}) P(c_i)$$

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_i^m \log \mathcal{N}(x_i \mid \mu_{c_i}, \Sigma_{c_i}) + \log P(c_i)$$

$$= \sum_{c=1}^{k} \sum_{i \mid c_i = c} \log \mathcal{N}(x_i \mid \mu_c, \Sigma_c) + \xi$$

ξ → COLLECTS TERMS NOT DEPENDING ON θ AND SO NOT RELEVANT TO THE MAXIMIZATION WRT θ

$$= \sum_{c=1}^{k} \ell_c(\mu_c, \Sigma_c) + \xi$$

↳ LOG LIKELIHOOD OF A GAUSSIAN MODEL FOR DATA OF CLASS c

IN ORDER TO MAXIMIZE $\ell(\theta)$ WE CAN MAXIMIZE $\ell_c(\mu_c, \Sigma_c)$

SOLVING THE FOLLOWING WILL ALLOW US TO FIND THE MAXIMUM:

$$\begin{cases} \nabla_{\Lambda_c, \ell_c}(\mu_c, \Lambda_c) \\ \nabla_{\mu_c, \ell_c}(\mu_c, \Lambda_c) \end{cases} \rightarrow \begin{cases} \Sigma_c^* = \Lambda_c^{-1} = \frac{1}{N_c} \sum_{i \mid c_i = c} (x_i - \mu_c)(x_i - \mu_c)^T \\ \mu_c^* = \frac{1}{N_c} \sum_{i \mid c_i = c} x_{i,c_i} \end{cases}$$

CLASS COVARIANCE

CLASS MEAN

NOW WE CAN SAY $f_{x \mid c}(x_t, c) \approx \mathcal{N}(x_t \mid \mu_c^*, \Sigma_c^*)$

CONSIDERING THE BINARY CLASS PROBLEM WE CAN EXPRESS THE COMPARISON RATIO IN TERMS OF CLASS POSTERIOR RATIO

$$\log r(x_t) = \log \frac{P(c=h_1 \mid x_t)}{P(c=h_0 \mid x_t)} = \log \frac{f_{x \mid c}(x_t \mid h_1) \cdot P(c=h_1)^{\pi}}{f_{x \mid c}(x_t \mid h_0) \cdot P(c=h_0)^{1-\pi}}$$

$$= \log \frac{f_{x \mid c}(x_t \mid h_1)}{f_{x \mid c}(x_t \mid h_0)} + \log \frac{\pi}{1-\pi}$$

$\log \frac{\pi}{1-\pi}$ → PRIOR LOG-ODDS

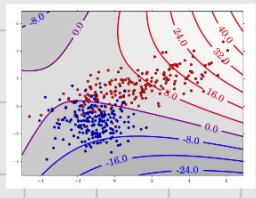
$llr(x_t)$: LOG-LIKELIHOOD RATIO

$llr(x_t)$: REPRESENTS THE RATIO BETWEEN THE LIKELIHOOD OF OBSERVING THE SAMPLE GIVEN THAT IT BELONGS TO h_1 OR h_0

WE ASSIGN CLASSES BASED ON: SCORE WITH PROBABILISTIC INTERPRETATION

$$llr(x_t) = \log \frac{f_{x \mid c}(x_t \mid h_1)}{f_{x \mid c}(x_t \mid h_0)} \gtrless -\log \frac{\pi}{1-\pi}$$

$$llr(x_t) = x^T A x + x^T b + c \quad \rightarrow \text{THIS DECISION FUNCTION IS QUADRATIC}$$



MULTICLASS

FOR MULTICLASS PROBLEMS $C \in \{h_1, \dots h_K\}$ WE CAN COMPUTE CLOSED-SET POSTERIOR PROBABILITIES AS

$$P(C=h|x_t) = \frac{f_{x|C}(x_t|h)P(h)}{\sum_{h' \in \{h\}} f_{x|C}(x_t|h')P(h')}$$

OPTIMAL DECISION REQUIRES CHOOSING THE CLASS WITH HIGHEST POSTERIOR PROBABILITY

$$P(C=h|x_t) \propto f_{x|C}(x_t|h)P(h) \longrightarrow$$

$$C_t^* = \arg\max_h P(C=h|x_t) = \arg\max_h \log f_{x|C}(x_t|h) + \log P(h)$$

SAMPLES MUST BE ENOUGH COMPARED TO THEIR DIMENSIONALITY, OTHERWISE THE ESTIMATIONS WILL BE INACCURATE

NAIVE BAYES

IF WE KNOW THAT, FOR EACH CLASS, THE DIFFERENT FEATURES ARE APPROXIMATELY INDEPENDENT WE CAN SIMPLIFY THE ESTIMATE ASSUMING THAT

$$f_{x|C}(x|C) \approx \prod_{j=1}^{D} f_{x_{[j]}|C}(x_{[j]}|C)$$

THE NAIVE BAYES ASSUMPTION COMBINED WITH GAUSSIAN ASSUMPTIONS MODELS

$$f_{x_{[j]}|C}(x_{[j]}|c) = \mathcal{N}(x_{[j]}|\mu_{c,[j]}, \sigma^2_{c,[j]})$$

WE CAN AGAIN COMPUTE THE ML ESTIMATES

$$\mathcal{L}(\theta) \propto \prod_{i=1}^{n} \prod_{j}^{D} \mathcal{N}(x_{i,[j]}|\mu_{c_i,[j]}, \sigma^2_{c_i,[j]})$$

$$l(\theta) = \xi + \sum_{j=1}^{D} \sum_{c=1}^{k} \sum_{i|c_i=c} \log \mathcal{N}(x_{i,[j]}|\mu_{c,[j]}, \sigma^2_{c,[j]})$$

WE CAN OPTIMIZE THE LOG-LIKELIHOOD INDIPENDETLY FOR EACH COMPONENT

THE ML SOLUTION IS

$$\begin{cases} \mu^*_{c,[j]} = \frac{1}{N_c} \sum_{i|c_i=c} x_{i,[j]} \\ \sigma^2_{c,[j]} = \frac{1}{N_c} \sum_{i|c_i=c} (x_{i,[j]} - \mu_{c,[j]})^2 \end{cases}$$

$$f_{x|C}(x|c) = \prod_{j=1}^{D} \mathcal{N}(x_{[j]}|\mu^*_{c,[j]}, \sigma^2_{c,[j]}) = \mathcal{N}(x|\mu_c, \Sigma_c)$$

$$\begin{cases} \mu_c = \begin{bmatrix} \mu_{c,[1]} \\ \vdots \\ \mu_{c,[D]} \end{bmatrix} \\ \Sigma_c = \begin{bmatrix} \sigma^2_{c,[1]} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{c,[D]} \end{bmatrix} \end{cases}$$

THE NAIVE BAYES GAUSSIAN CLASSIFIER CORRESPONDS TO A MVG CLASSIFIER WITH DIAGONAL COVARIANCE MATRIX (WHICH DOES NOT HOLD FOR A GENERIC DENSITY)

TIED

ASSUMES THAT THE COVARIANCE MATRIX OF DIFFERENT CLASSES AR TIED

- CLASS INDEPENDENT NOISE : $x_{c,i} = \mu_c + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0, \Lambda^{-1})$
- BADLY-CONDITIONED PROBLEMS (LARGE DIMENSIONAL DATA, SMALL NUMBER OF SAMPLES)

TIED COVARIANCE MODEL ASSUMES THAT:

$$f_{X|C}(x|c) = \mathcal{N}(x \mid \mu_c, \Sigma)$$

EACH CLASS HAS ITS OWN MEAN BUT THE COVARIANCE IS THE SAME FOR ALL CLASSES

THE ML SOLUTION IS:

$$\begin{cases} \mu_c^* = \frac{1}{N_c} \sum_{i|c_i=c} x_i \\ \Sigma^* = \frac{1}{N} \sum_c \sum_{i|c_i=c} (x_i - \mu_c)(x_i - \mu_c)^T \end{cases}$$

$$llr(x) = x^T b + c$$ → LINEAR SEPARATION RULE

$$b = \Sigma^{-1}(\mu_1 - \mu_0)$$

$$c = -\frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0)$$

RELATION WITH LDA

IN THE BINARY LDA CASE WE LOOK FOR THE DIRECTION WHICH MAXIMIZES:

$$\frac{w^T S_B w}{w^T S_W w}$$

WHERE $S_B = k(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ AND $S_W = \Sigma$

AT THE END WE GET THAT THE PROJECTION OVER SUCH DIRECTION IS:

$$w^T x = k x^T \Sigma^{-1}(\mu_1 - \mu_0)$$ → CORRESPONDS TO THE CLASSIFICATION RULE OF TIED MVG

INDEED LDA ASSUMES SAME WITHIN CLASS COVARIANCE S_W

HERE THE DIFFERENCE WITH LDA IS THAT THE OPTIMAL THRESHOLD IS $t = \log \frac{\pi}{1-\pi}$

WHICH IS APPLICATION DEPENDENT.

CONSIDERATIONS:

- IF DATA IS HIGH-DIMENSIONAL PCA CAN SIMPLIFY THE ESTIMATION
- PCA ALLOWS REMOVING DIMENSIONS WITH VERY SMAL VARIANCE
- MULTIVARIATE MODELS PERFORM BESIER IF WE HAVE ENOGH DATA TO RELIABLY ESTIMATE THE COVARIANCE MATRIX
- NAIVE BAYES MAY SIMPLIFY THE ESTIMATION BUT IF THE DATA IS HIGHLY CORRELATED IT CAN PERFORM BADLY
- TIED COVARIANCE MODELS CAN CAPTURE CORRELATIONS BUT MAY PERFORM POORLY WHEN CLASSES HAVE DIFFERENT DISTRIBUTION
- TIED + LDA PERFORMS AS TIED ALONE SINCE THE LDA SUBSPACE CONTAINS ALL THE INFORMATION USED BY THE TIED MODEL

DISCRETE MODEL

GATTO 1 GATTO 2

$\{(\text{RED}, \text{FEMALE}), (\text{BROWN}, \text{MALE}) \dots\}$

LET'S ASSUME TO HAVE A CATEGORICAL FEATURE $x \in \{1 \dots m\}$ AND A SET OF LABELED TRAINING SAMPLES $D = \{(x_1, C_1), \dots (x_m, C_m)\}$. WE ASSUME ALSO THE SAMPLES TO BE IID

WE WANT TO COMPUTE $P(X_r = x_r \mid C_r = c) = \pi_{c,x_r}$ FOR DIFFERENT HYPOTHESIS c FOR THE SAMPLE x_r. $\pi_c = (\pi_{c,1} \dots \pi_{c,m})$ WILL BE THE MODEL PARAMETERS FOR CLASS c WITH THE CONSTRAINT $\sum_{i=1}^{m} \pi_{c,i} = 1$

WE WILL IMPLEMENT THE FREQUENTIST APPROACH AND ESTIMATE THE ML SOLUTION FOR $\pi(\pi_1 \dots \pi_k)$

THE LIKELIHOOD OF THE TRAINING SET IS

$$\mathcal{L}(\pi) = \prod_{i=1}^{n} P(X_i = x_i \mid C_i = c_i)\, P(C_i = c_i)$$

$$l(\pi) = \sum^{n} \log \overbrace{P(X_i = x_i \mid C_i = c_i)}^{\pi_{c_i, x_i}} + \{$$

$$= \sum_{c=1}^{K} \sum_{i \mid C_i = c} \log \pi_{c_i, x_i} + \{$$

$$= \sum_{c=1}^{K} l_c(\pi_c) + \{$$

WE CAN THUS ESTIMATE THE PARAMETERS BY INDEPENDENTLY OPTIMIZING $l_c(\pi_c)$

$$l_c(\pi_c) = \sum_{i \mid C_i = c} \log \pi_{c,x_i}$$

$$= \sum_{j=1}^{m} N_{c,j} \log \pi_{c,j}$$

$N_{c,j}$: NUMBER OF TIME WE OBSERVED $x_i = j$

WE NOW LOOK FOR THE MAXIMIZER REMEMBERING THAT $\sum_{i}^{m} \pi_{c,i} = 1$

A SOLUTION CAN BE FOUND USING LAGRANGE MULTIPLIERS, WHICH GIVES AS RESULT

$$\pi^{*}_{c,i} = \frac{N_{c,i}}{N_c}$$

i.e. THE FREQUENCY VALUE i IN CLASS c

NOW WE CAN SAY THAT: $P(X_r = x_r \mid C_r = c) = \pi^{*}_{c,x_r}$

NAIVE BAYES WE MAY HAVE MORE THAN ONE CATEGORICAL ATTRIBUTE FOR EACH SAMPLE

IF WE HAVE MORE THAN AN ATTRIBUTE WE CAN AGAIN ADOPT NAIVE BAYES ASSUMING THAT FEATURES ARE INDEPENDENT. WE CAN OBTAIN ML ESTIMATES FOR EACH FEATURE AND THEN COMBINE THEM

$$P(X_r = x_r \mid C_r = c) = \prod \pi^{\gamma}_{c,x_r,[\gamma]}$$

OCCURRENCES

$\{$ (10, 2, 4, ROMANZO) (LIBRO 1), (10, 8, 0, ROMANZO) (LIBRO 2), (1, 30, 2, MANGA) (LIBRO 3) $\}$

NOW WE CONSIDER THE CASE WHEN FEATURES REPRESENT OCCURRENCES OF EVENTS, FOR EXAMPLE:

$x = (x_1 \dots x_m)$ WHERE x_i REPRESENTS THE NUMBER OF TIMES WE OBSERVED WORD i

THUS FOR EACH CLASS WE HAVE A SET OF PARAMETERS $\pi_c = (\pi_{c,1} \dots \pi_{c,m})$ THAT REPRESENT THE PROBABILITY OF OBSERVING A SINGLE INSTANCE OF WORD i

THE PROBABILITY OF FEATURE VECTOR x IS GIVEN BY THE MULTINOMIAL DENSITY

$$p(X=x \mid C=c) = \frac{\left(\sum_{j=1}^{m} x_{[j]}\right)!}{\prod_{j=1}^{m} x_{[j]}!} \prod_{j=1}^{m} \pi_{c,j}^{x_{[j]}} \propto \prod_{j=1}^{m} \pi_{c,j}^{x_{[j]}}$$

We did not write the multinomial coefficient since it's constant with respect to π and so can be written inside ξ

$$l(\pi) = \sum_c l_c(\pi_c) + \xi$$

$$l_c(\pi_c) = \sum_{i \mid c_i = c} \sum_{j=1}^{m} x_{i,[j]} \log \pi_{c,j} = \sum_{j=1}^{m} N_{c,j} \log \pi_{c,j}$$

$N_{c,j}$: TOTAL NUMBER OF OCCURRENCES OF EVENT j IN THE SAMPLES OF CLASS c $\rightarrow$ $N_{c,j} = \sum_{i \mid c_i = c} x_{i,[j]}$

WE NOTICE THAT THE FORM IS EXACTLY THE SAME AS THE ONE SOLVED FOR THE CATEGORICAL CASE.

THE ML SOLUTION IS AGAIN: $\pi_{c,j} = \frac{N_{c,j}}{N_c}$

- $\frac{N_{c,j}}{N_c}$: RELATIVE FREQUENCY OF WORD j IN CLASS c
- N_c: TOTAL NUMBER OF WORDS FOR CLASS c

WE WRITE THE LLR FOR A 2 CLASS PROBLEM (BINOMIAL COEFFICIENT CAN BE SIMPLIFIED AND THUS DISAPPEARS FROM THE EXPRESSION)

$$llr(x) = \log \frac{P(X=x \mid C=L_1)}{P(X=x \mid C=L_0)} = \sum_{j=1}^{m} x_{[j]} \log \pi_{L_0,j} - \sum_{j=1}^{m} x_{[j]} \log \pi_{L_0,j} = x^T b$$

$\rightarrow$ LINEAR DECISION FUNCTION

CONSIDERATIONS:

- RARE WORDS MAY CAUSE PROBLEMS. IF A WORD DOES NOT APPEAR IN A TOPIC WE WILL ESTIMATE $\pi_{c,j} = 0$. ANY TEST SAMPLE THAT CONTAINS THE WORD WILL HAVE 0 PROBABILITY OF BEING OF CLASS c
  - $\rightarrow$ WE CAN MITIGATE THIS ISSUE INTRODUCING PSEUDO-COUNTS. i.e. ASSUMING THAT EACH TOPIC CONTAINS A SAMPLE WHERE ALL WORDS APPEAR A FIXED NUMBER OF TIMES. IN PRACTICE WE CAN ADD A FIXED VALUES TO THE CLASS OCCURRENCES N_c BEFORE COMPUTING THE ML SOLUTION
- WE CAN CONSIDER PAIR OF WORDS, TRIPLETS AND SO ON IF WE WANT PARTIALLY ACCOUNT FOR CORRELATIONS
- WE CAN COMBINE DIFFERENT MODELS THROUGH A NAIVE BAYES ASSUMPTION

COMMENTS ON DISCRETE MODELS FOR CATEGORICAL EVENTS

WE CAN MODEL A DATASET OF CATEGORICAL SAMPLES AS n INDEPENDENT CATEGORICAL R.V. $X_i \in \{1 \dots m\}$

THE DISTRIBUTION IS DESCRIBED BY A VECTOR OF PROBABILITIES π THAT ALLOW COMPUTING $P(X=j) = \pi_j$

ALTERNATIVELY WE CAN MODEL THE DATASET IN TERMS OF OCCURRENCES OF EVENTS.

WE HAVE A RANDOM VECTOR $Y = (Y_1 \dots Y_m)$ WHOSE COMPONENTS ARE R.V. Y_i CORRESPONDING TO THE NUMBER OF OCCURRENCES OF EVENT i IN THE DATASET

THE DISTRIBUTION IS DESCRIBED BY A VECTOR OF PROBABILITIES $\pi = (\pi_1 \dots \pi_m)$ THAT REPRESENT THE PROBABILITIES OF SINGLE EVENTS

IN BOTH CASES THE ML SOLUTION FOR π CORRESPONDS TO THE RELATIVE FREQUENCIES OF OCCURRENCES

LOG-LIKELIHOOD FOR A GIVEN CLASS

FIRST CASE:

$$l_x(\pi) = \sum_{i=1}^{n} \log \pi_{x_i} = \sum_{j=1}^{m} N_j \log \pi_j$$

SECOND CASE:

$$l_y(\pi) = \log \frac{\left(\sum_{j=1}^{m} y_j\right)!}{\prod_{j=1}^{m} y_j!} + \sum_{j=1}^{m} y_j \log \pi_j = \xi + \sum_{j=1}^{m} N_j \log \pi_j$$

$$L_x(\pi) \approx L_y(\pi)$$

THEREFORE:

- SAME ML ESTIMATES
- SAME INFERENCE (SINCE EQUAL lk AND SO EQUAL CLASS POSTERIORI PROBABILITIES)

LOGISTIC REGRESSION

DISCRIMINATIVE APPROACH FOR CLASSIFICATION. RATHER THAN MODELING THE DISTRIBUTION OF OBSERVED SAMPLES $X|C$ WE DIRECTLY MODEL THE CLASS POSTERIOR DISTRIBUTION $C|X$

NEED TO DEFINE A MODEL FOR $P(C=c \mid X=x)$

WE START FROM THE (CLASS LOG-POSTERIOR PROBABILITY) OF A 2 CLASS PROBLEM USING A TIED GAUSSIAN MODEL

$$\log \frac{P(C=h_1|x)}{P(C=h_0|x)} = \log \frac{f_{X|C}(x|h_1)}{f_{X|C}(x|h_0)} + \log \frac{\pi}{1-\pi} = w^T x + b$$

(PRIOR INFO ABSORBED IN b)

GIVEN w, b WE CAN COMPUTE THE EXPRESSION FOR THE POSTERIOR CLASS PROBABILITY AS:

$$P(C=h_1|x,w,b) = e^{(w^T x+b)}\,\underbrace{(1-P(C=h_1|x,w,b))}_{\hookrightarrow P(C=h_0|x,w,b)}$$

WE SOLVE FOR $P(C=h_1|x,w,b)$

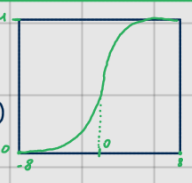
$$P(C=h_1|x,w,b) = \frac{1}{1+e^{-(w^T x+b)}} = \sigma(w^T x+b)$$

SIGMOID FUNCTION

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$1-\sigma(x) = \sigma(-x)$$

$$\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$$



THE LAST EQUATION PROVIDES A MODEL THAT ALLOWS COMPUTING THE POSTERIOR PROBABILITY FOR h_1 and h_0. THE DECISION RULES ARE LINEAR SURFACES ORTHOGONAL TO w.

MODEL PARAMETERS ARE (w, b). WE WILL FOLLOW A FREQUENTIST APPROACH i.e. COMPUTING AN ESTIMATION FOR (w, b) FROM A SET OF TRAINING SAMPLES

WE HAVE $D = [(x_1, c_1), \dots (x_m, c_m)]$ WITH CLASSES THAT ARE INDEPENDENTLY DISTRIBUTED

$C_i | x_i, w, b \sim C | x_i, w, b$

THE CLASS POSTERIOR MODEL ALLOWS EXPRESSING THE LIKELIHOOD FOR THE OBSERVED LABELS AS:

$$P(C_1 = c_1, \dots C_m = c_m \mid x_1 \dots x_m, w, b) = \prod_{i=1}^{m} P(C_i = c_i \mid x_i, w, b)$$

WE CAN APPLY A ML APPROACH TO ESTIMATE THE MODEL PARAMETERS THAT BEST DESCRIBE THE OBSERVED LABELS $(c_1 \dots c_m)$

WE ESTIMATE THE VALUE OF w THAT MAXIMIZES THE LIKELIHOOD OF OUR TRAINING LABELS

$$\begin{cases} h_1 = 1 \\ h_0 = 0 \end{cases}$$

$$y_i = P(C_i = 1 \mid x_i, w, b) = \sigma(w^T x + b)$$

$$P(C_i = 0 \mid x_i, w, b) = 1 - y_i = 1 - \sigma(w^T x + b) = \sigma(-w^T x - b)$$

$$C_i \mid x_i, w, b \sim Ber(\sigma(w^T x + b)) = Ber(y_i)$$

$$\mathcal{L}(w, b) = \prod_{i=1}^{n} P(C = c_i \mid x_i, w, b)$$

$$= \prod_{i=1}^{n} y_i^{c_i} (1 - y_i)^{(1 - c_i)}$$

$$l(w, b) = \sum_{i=1}^{n} \left[c_i \log y_i + (1 - c_i) \log (1 - y_i) \right]$$

NOW WE FIND w^*, b^* THAT MAXIMIZE $l(w, b)$

$$w^*, b^* = \underset{w,b}{\operatorname{argmax}}\ l(w, b)$$

BUT RATHER THAT DOING IT WE WILL MINIMIZE

$$J(w, b) = -l(w, b) = \sum_{i=1}^{n} -\left[c_i \log y_i + (1 - c_i) \log (1 - y_i) \right] = \sum_{i=1}^{n} H(c_i, y_i)$$

$H(c_i, y_i)$ IS THE BINARY CROSS-ENTROPY BETWEEN THE DISTRIBUTION OF OBSERVED DATA AND PREDICTED LABELS FOR THE i-th SAMPLE

CROSS ENTROPY

GIVEN P, Q 2 DISTRIBUTIONS OVER THE SAME DOMAIN

THE CROSS ENTROPY BETWEEN THE 2 DISTRIBUTIONS IS

$$H(P, Q) = -E_{P(x)} [\log Q(x)]$$

IN OUR CASE P IS THE DISTRIBUTION OF THE CLASS LABEL FROM THE POINT OF VIEW OF AN OBSERVER $\mathcal{E}$ WHO KNOWS THE ACTUAL LABELS.

$$P(C_i = 1 \mid X_i = x_i, \mathcal{E}) = \begin{cases} 1 & \text{IF } C_i = 1 \\ 0 & \text{IF } C_i = 0 \end{cases} = C_i \rightarrow \sim Ber(C_i)$$

$$P(C_i = 0 \mid X_i = x_i, \mathcal{E}) = \begin{cases} 0 & \text{IF } C_i = 1 \\ 1 & \text{IF } C_i = 0 \end{cases} = 1 - C_i \nearrow$$

DISTRIBUTION Q IS THE DISTRIBUTION FOR THE PREDICTED LABELS ACCORDING TO OUR RECOGNIZER R

$$Q(c) = P(c_i = c \mid x_i = x_i, R(w,b))$$

$$Q(1) = P(c_i = 1 \mid x_i = x_i, R(w,b)) = y_i = \sigma(w^T x + b)$$
$$Q(0) = P(c_i = 0 \mid x_i = x_i, R(w,b)) = 1 - y_i = 1 - \sigma(w^T x + b)$$

LOGISTIC REGRESSION LOOKS FOR THE MINIMIZER OF THE AVERAGE CROSS ENTROPY BETWEEN THE DISTRIBUTIONS FOR THE TRAINING SET LABELS OF AN EVALUATOR E WHO KNOWS THE REAL LABEL AND THE DISTRIBUTIONS FOR THE TRAINING SET LABELS AS PREDICTED BY THE MODEL $R(w,b)$ ITSELF

CROSS-ENTROPY IS A MEASURE OF GOODNESS OF THE PREDICTIONS AND THE EVALUATION IS PERFORMED OVER THE TRAINING DATA ITSELF. IN OUR CASE IT MEASURES THE DIFFERENCE BETWEEN THE PREDICTED DISTRIBUTION $Ber(y_i)$ AND THE EMPIRICAL LABEL DISTRIBUTION $Ber(c_i)$ (THE DISTRIBUTION OF THE EVALUATOR E)

THE CROSS-ENTROPY, AS A FUNCTION OF Q, IS MINIMIZED WHEN $Q = P$, SO IT MEANS THAT WE ARE LOOKING FOR A LABEL DISTRIBUTION THAT IS ON AVERAGE AS SIMILAR AS POSSIBLE TO THE EMPIRICAL ONE

AN INTERESTING INTERPRETATION OF LOGISTIC REGRESSION OBJECTIVE CAN BE OBTAINED BY REWRITING THE CROSS ENTROPY IN TERMS OF $z_i = 2c_i - 1$

$$z_i = 2c_i - 1 = \begin{cases} 1 & \text{IF } c_i = 1 \\ -1 & \text{IF } c_i = 0 \end{cases}$$

STILL REPRESENTS CLASS LABELS, BUT FOR SAMPLES BELONGING TO h_0 WE HAVE $c_i = -1$ AND FOR SAMPLES BELONGING TO h_1 WE HAVE $z_i = 1$

GIVEN $s_i = w^T x_i + b$ WE CAN REWRITE $H(c_i, y_i)$ USING z_i

$$H(c_i, y_i) = -\log \sigma(z_i s_i) = \log\left(1 + e^{-z_i (w^T x_i + b)}\right)$$

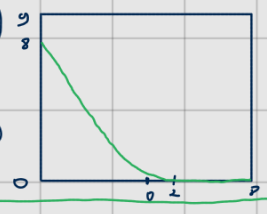
OBJECTIVE FUNCTION

$$J(w,b) = \sum_{i=1}^{m} H(c_i, y_i) = \sum_{i=1}^{m} \log\left(1 + e^{-z_i (w^T x_i + b)}\right) = \sum_{i=1}^{m} l\left(z_i (w^T x_i + b)\right)$$

→ IS A FUNCTION OF c_i, w, b, x_i SINCE $y_i = \sigma(w^T x_i + b)$

LOGISTIC LOSS FUNCTION

$$l(x) = \log\left(1 + e^{-x}\right)$$



COST OF PREDICTION MADE WITH MODEL (w,b) FOR EACH SAMPLE

RECALL CLASS LOG-POSTERIOR PROBABILITY RATIO

$$\log \frac{P(h_1 | x_i)}{P(h_0 | x_i)} = w^T x_i + b = s_i$$

→ LINEAR HYPERPLANE ORTHOGONAL TO w

→ DECISION RULES TAKE THE FORM $s_i \lessgtr t$

s_i RELATED TO DISTANCE OF x_i FROM THE SEPARATING SURFACE

WHEN s_i IS POSITIVE THE CLASSIFIER FAVORS CLASS h_1, WHEN s_i IT'S NEGATIVE THE CLASSIFIER FAVOURS h_0. THE COST WE PAY FOR EACH SAMPLE IS $l(z_i \Lambda_i)$

- PREDICTION AND ACTUAL CLASS AGREE: $z_i = 1$, $s_i > 0$ OR $z_i = -1$ AND $s_i < 0$ THEN $z_i \cdot s_i > 0$ AND WE PAY LOW COST. THE COST BECOMES EXPONENTIALLY SMALLER AS $|s_i|$ INCREASES (WE ARE FAR FROM THE SEPARATION SURFACE)

- PREDICTION AND ACTUAL CLASS DISAGREE: $z_i = 1$, $s_i < 0$ OR $z_i = -1$, $s_i > 0$. THEN $z_i s_i < 0$ AND WE PAY A COST THAT INCREASES LINEARLY WITH $|s_i|$

THUS WE CAN INTERPRET THE LOGISTIC REGRESSION OBJECTIVE AS A MEASURE OF EMPIRICAL RISK. OUR GOAL IS TO MINIMIZE IT.

LOGISTIC REGRESSION SOLUTION CANNOT BE COMPUTED IN CLOSED FORM. WE WILL RESORT TO NUMERICAL SOLVERS WHICH ITERATIVELY LOOKS FOR THE MINIMIZER OF A FUNCTION. WE WILL USE THE L-BFGS ALGORITHM. THE ALGORITHM REQUIRES A FUNCTION THAT COMPUTES THE LOSS AND ITS GRADIENT WITH RESPECT TO w AND b

IF CLASSES ARE LINEARLY SEPARABLE, THE LOGISTIC REGRESSION SOLUTION IS NOT DEFINED IN THIS CASE WE CAN MAKE THE VALUES OF s_i ARBITRARILY HIGH BY JUST INCREASING THE NORM OF w (AND CHANGING ACCORDINGLY THE VALUE OF b)

AS WE INCREASE $\|w\|$ THE LOSS BECOMES LOWER, THUS WE ARE DECREASING THE OBJECTIVE FUNCTION

THE FUNCTION DOES NOT HAVE A MINIMUM BUT IT HAS AN INFINUM $\inf(J(w,b))$

CORRESPONDING TO $\|w\| \to \infty$

TO MAKE THE PROBLEM SOLVABLE WE CAN LOOK FOR SOLUTION WITH SMALL NORM

BY INTRODUCING A NORM PENALTY (REGULARIZATION TERM) TO THE OBJECTIVE FUNCTION

THE OBJECTIVE FUNCTION THAT WE MINIMIZE IS

$$\tilde{R}(w,b) = \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{m} \log\left(1 + e^{-z_i(w^T x_i + b)}\right)$$

λ: HYPER-PARAMETER THAT ALLOWS SPECIFYING THE RELATIVE WEIGHT OF THE REGULARIZATION TERM

ALTERNATIVELY WE LOOK FOR THE MINIMIZER OF:

$$R(w,b) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\log\left(1 + e^{-z_i(w^T x_i + b)}\right)$$

λ: REGULARIZATION COEFFICIENT

$\frac{1}{m}$: RISK AVERAGED OVER ALL SAMPLES

λ BEING AN HYPER-PARAMETER SHOLD BE SELECTED AS TO OPTIMIZE THE PERFORMANCE OF THE CLASSIFIER. (WE CANNOT COMPUTING BY MINIMIZING R WRT λ BECAUSE WE WOULD OBTAIN THE TRIVIAL SOLUTION $\lambda = 0$). THE SELECTION OF λ SHOULD BE DONE USING OTHER APPROACHES, SUCH AS CROSS VALIDATION

THE MODEL IS CALLED REGULIRIZED LOGISTIC REGRESSION AND IS AN EXAMPLE OF REGULARIZED RISK MINIMIZATION PROBLEM.

$$R(w,b) = \Omega(w,b) + \frac{1}{m}\sum_{i=1}^{m} l(z_i, x_i, w, b)$$

REGULARIZATION ALLOWS REDUCING THE RISK OF OVER-FITTING THE TRAINING DATA

IF λ IS TOO LARGE, WE WILL OBTAIN A SOLUTION WITH SMALL NORM BUT NOT ABLE TO WELL SEPARATE CLASSES.

IF λ IS TOO SMALL WE WILL GET A SOLUTION WITH GOOD SEPARATION ON THE TRAINING SET, BUT MAY HAVE POOR CLASSIFICATION ACCURACY FOR UNSEEN DATA (OVERFITTING)

CONSIDERATIONS:

- NON-REGULARIZED MODEL IS INVARIANT TO LINEAR TRANSFORMATIONS OF THE FEATURE VECTORS. REGULARIZED MODE, ON THE OTHER HAND, IS NOT INVARIANT. SO IT COULD BE USEFUL TO

PRE-PROCESS DATA SO THAT DYNAMIC RANGES OF DIFFERENT FEATURES ARE SIMILAR

COMMON PREPROCESSING STRATEGIES:

- CENTER DATA
- STANDARDIZE VARIANCES (DIVIDE EACHE FEATURE BY THE STANDARD DEVIATION COMPUTED OVER THE TRAINING SET)
- WHITEN COVARIANCE MATRIX: NORMALIZE VARIANCES WHILE MAKING FEATURES UNCORRELATED

$x_i = A x_i \quad , A = \Sigma^{\frac{1}{2}}$ Σ IS THE TRAINING SET COVARIANCE

- L2 (LENGTH NORMALIZATION): $x_i' = \frac{x_i}{\|x_i\|}$ (OFTEN AFTER CENTERING AND WHITENING)

LOGISTIC REGRESSION SCORE CAN BE SEEN AS THE log OF THE RATIO BETWEEN CLASS POSTERIOR PROBABILITIES

IT REFLECTS THE EMPIRICAL CLASS PRIOR OF THE DATA.

WE CAN SIMULATE DIFFERENT CLASS PRIORS USING A DIFFERENT WEIGHTED VERSION OF THE MODEL:

$$R(w) = \frac{\lambda}{2}\|w\|^2 + \frac{\pi_T}{m_T}\sum_{i|z_i=+1} \ell(z_i s_i) + \frac{1-\pi_T}{m_F}\sum_{i|z_i=-1} \ell(z_i s_i)$$

PROBLEM: WE ARE USING PRIORS DURING TRAINING. IN CASE WE HAVE A NEW APPLICATION WE SHOULD RETRAIN OUR MODEL!

ANOTHER SOLUTION MAY BE TO "MOVE" THE SEPARATION LINE TOWARDS ONE OF THE TWO CLASSES BY SUBTRACTING A CONSTANT TERM TO THE SCORE

$$s_i \to s_i - \log \frac{m_T}{m_F}$$

$m_T \rightsquigarrow$ # SAMPLES OF h_1

$m_F \rightsquigarrow$ # SAMPLES OF h_0

THIS SOLUTION DOES NOT REQUIRE KNOWING PRIORS BEFORE TRAINING

MULTICLASS LOGISTIC REGRESSION

WE CONSIDER NOW A PROBLEM WITH k CLASSES $\{1 \dots k\}$

WE START FROM THE POSTERIOR LIKELIHOOD RATIOS OF THE LINEAR GAUSSIAN CLASSIFIER WITH UNIFORM PRIORS

$$\log \frac{P(C=j|x)}{P(C=R|x)} = (w_j - w_R)^T x + (b_j - b_R)$$

CORRESPONDING TO PAIR-WISE LINEAR CLASSIFICATION SURFACE BETWEEN CLASS j AND R

WE WILL USE R AS REFERENCE CLASS FOR OUR MATHEMATICAL DERIVATION.

NOTE THAT THIS IS ALSO CONSISTENT WITH $\log \frac{P(C=R|x)}{P(C=R|x)} = 0$

SO FOR ALL CLASSES $j \in 1 \dots k$

$$P(C=j|x) = P(C=R|x)\, e^{(w_j - w_R)^T x + (b_j - b_R)}$$

$$\sum_{j=1}^{k} P(C=j|x) = 1 \quad \text{REMEMBER!}$$

$$P(C=R|x) = 1 - \sum_{j \neq R} P(C=j|x) = 1 - \sum_{j \neq R}^{k} P(C=R|x)\, e^{(w_j - w_R)^T x + (b_j - b_R)}$$

$$= \frac{e^{w_R^T x + b_R}}{\sum_j e^{w_j^T x + b_j}} = f_R(w_R^T x + b_R) = f_R(s)$$

SOFTMAX FUNCTION

$$f_i(s) = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

SO GIVEN THE MODEL PARAMETERS $W = [w_1 \dots w_k]$, $b = \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$ THE LOGISTIC REGRESSION MODEL ALLOWS TO COMPUTE THE PROBABILITY OF EACH CLASS

$$P(C=k|W,b,x) = \frac{e^{w_k^T x + b_k}}{\sum_j e^{w_j^T x + b_j}}$$

CONSIDERING x_i ITS CLASS POSTERIOR DISTRIBUTION IS A CATEGORICAL DISTRIBUTION

$$c_i | W, b, x_i = x_i \sim \text{Cat}(y_i) \quad \text{WHERE} \quad y_{ik} = \frac{e^{w_k^T x_i + b_k}}{\sum_j e^{w_j^T x_i + b_k}}$$

REPRESENTS THE DISTRIBUTION OF THE CLASS LABEL ACCORDING TO LR MODEL

LET'S EXPRESS THE LOG-LIKELIHOOD

$$\ell(W,b) = \sum_i^m \log P(C_i = c_i | X_i = x_i, W, b)$$

$$\log P(C_i = c_i | X_i = x_i, W, b) = \sum_k^k z_{ik} \log y_{ik}$$

$$z_{ik} = \begin{cases} 1 & \text{if } c_i = k \\ 0 & \text{otherwise} \end{cases}$$

↘ 1 OF k OF CATEGORICAL DISTRIBUTIONS

$$l(W,b) = \sum_{i=1}^{m} \sum_{k=1}^{k} z_{ik} \log y_{ik}$$

$$H(z_i, y_i) = -\sum_{k=1}^{k} z_{ik} \log y_{ik}$$

MULTICLASS ENTROPY BETWEEN OBSERVED AND PREDICTED LABEL DISTRIBUTION FOR SAMPLE k.

WE ESTIMATE W, b AS TO MAXIMIZE THE LIKELIHOOD OF THE TRAINING LABEL WHICH CORRESPONDS TO MINIMIZING THE CROSS ENTROPY

$$\underset{W,b}{\operatorname{argmax}}\, l(W,b) = \underset{W,b}{\operatorname{argmin}} \sum_{i=1}^{m} H(z_i, y_i)$$

WE CAN CAST THE PROBLEM AS A MINIMIZATION OF A LOSS FUNCTION

$$J(W,b) = -\sum_{i}^{m} \sum_{k}^{k} z_{ik} \log y_{ik}$$

$$= -\sum_{i}^{m} \log \frac{e^{w_k^T x_i + b_k}}{\sum_{c_i}^{k} e^{w_{c_i}^T x_i + b_{c_i}}}$$

$$= \sum_{i}^{m} l(x_i, c_i, W, b)$$ ← SOFTMAX LOSS

AGAIN WE ADD A REGULARIZATION TERM TO REDUCE OVERFITTING:

$$R(W,b) = \Omega(W) + \frac{1}{m} J(W,b)$$

$$\Omega(W) = \frac{1}{2} \sum_{i} \|w_i\|^2$$

AGAIN WE CAN STILL REPLACE THE AVERAGE CROSS ENTROPY WITH THE PRIOR WEIGHTED VERSION TO ACCOUNT FOR PRIORS THAT ARE DIFFERENT FROM THE EMPIRICAL TRAINING SET PRIOR

QUADRATIC SURFACES RL

FROM GAUSSIAN CLASSIFIER WITH NON TIED COVARIANCES WE HAVE:

$$\log \frac{P(C=k_1|x)}{P(C=k_0|x)} = x^T A x + b^T x + c = \Lambda(x, A, b, c)$$ → QUADRATIC SEPARATION RULE

WE DON'T KNOW A, b, c WE SHOULD ESTIMATE THEM AS WE DID BEFORE.

HOWEVER WE COULD TRY TO FIND A TRANSFORMED SPACE IN WHICH LINEAR

SURFACES CORRESPOND TO QUADRATIC SURFACES IN THE x SPACE

$\Lambda(x, A, b, c)$ IS QUADRATIC IN x BUT LINEAR IN A, b

WE REWRITE

$$x^T A x = \langle xx^T, A \rangle = \text{vec}(xx^T)^T \text{vec}(A)$$

(INNER PRODUCT)

$\text{vec}(M)$: STACKS COLUMNS OF M

THEN

$$\phi(x) = \begin{bmatrix} \text{vec}(xx^T) \\ x \end{bmatrix} \quad w = \begin{bmatrix} \text{vec}(A) \\ b \end{bmatrix}$$

WE CAN EXPRESS CLASS-LOG POSTERIOR RATIO AS

$$\Lambda(x, w, c) = w^T \phi(x) + c$$

THUS WE CAN TRAIN AN RL MODEL USING THE FEATURES OF $\phi(x)$ RATHER THAN x. THIS SPACE IS CALLED EXPANDED FEATURE SPACE

THE DIMENSION OF THE EXPANDED FEATURE SPACE CAN GROW VERY QUICKLY!

MODEL EVALUATION FOR CLASSIFICATION

HOW GOOD IS OUR MODEL ON THE TEST SET?

TO ASSESS THIS WE COULD CHECK ACCURACY AND ERROR RATE

$$ACC = \frac{\#\,CORRECTLY\_CLASSIFIED}{\#\,SAMPLES} \qquad ERR = 1 - ACC$$

BUT ACCURACY CAN BE MISLEADING WHEN CLASSES ARE UNBALANCED

	CLASS F	CLASS T
PREDICTION F	TF	FF
PREDICTION T	FT	TT

WE CAN COMPUTE SOME ACCURACY MEASURES

FALSE NEGATIVE RATE:

$$FNR = \frac{FN}{FN+TP}$$

TRUE POSITIVE RATE

$$TPR = \frac{TP}{TP+FN} = 1 - FNR$$

FALSE POSITIVE RATE:

$$FPR = \frac{FP}{FP+TN}$$

TRUE NEGATIVE RATE:

$$TNR = \frac{TN}{TN+FP} = 1 - FPR$$

CLASSIFIER OUTPUT SCORES

GENERATIVE MODELS $\quad s = \log \frac{f(x \mid H_T)}{f(x \mid H_F)}$

DISCRIMINATIVE MODELS $\quad s = \log \frac{P(H_T \mid x)}{P(H_F \mid x)}$

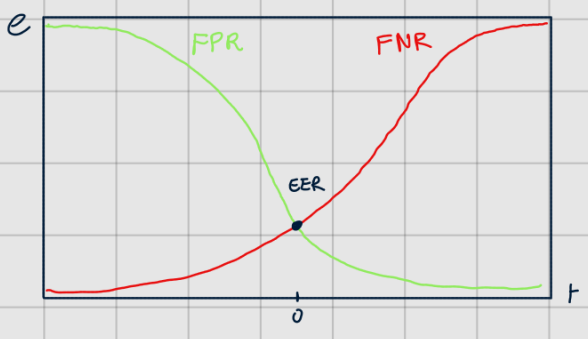
NON PROBABILISTIC MODELS $\quad s = w^T x$

CLASS ASSIGNMENT IS performed by comparing the score with a threshold t

$s \geq t \rightarrow H_T$ $\quad s < t \rightarrow H_F$

Different thresholds correspond to different error rates

We can visualize the error rates over the threshold

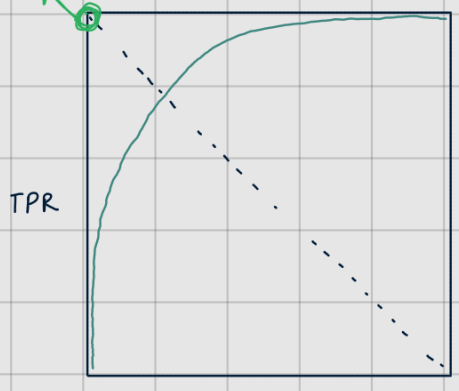


EER: EQUAL ERROR RATE

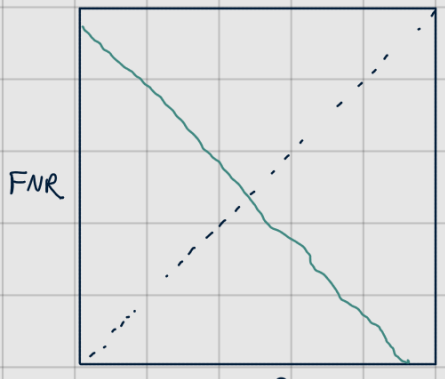
When we increase the threshold FPR will decrease (less FP) but FNR will increase (more FN)

HERE IS BEST

RECEIVE OPERATING CHARACTERISTICS
ROC CURVE



DETECTION ERROR TRADE OFF
DET CURVE



The goal of the clasifier is to choose an action a over a set of actions A

We can assign a cost $C(a|k)$ that we pay when we take action a and the sample belongs to class k

FROM NOW ON $C(a|k)$ MEANS THE COST OF ASSIGNING CLASS a WHEN IT BELONGS TO CLASS k

$$P(C=k|x,R)$$

POSTERIOR PROBABILITY OF SAMPLE BEING OF CLASS k GIVEN THE SAMPLE AND A RECOGNIZER R

$$C_{x,R}(a) = E[C(a|k)|x,R] = \sum_{k=1}^{K} C(a|k)P(C=k|x,R)$$

EXPECTED COST OF ACTION A WHEN THE POSTERIOR PROBABILITY FOR EACH CLASS IS $P(C=k|x,R)$

It measures the cost that we expect to pay given our knowledge of the class distribution $P(C=k|x,R)$

BAYES DECISION aims to take the action $a^*(x,R)$ that minimizes the expected cost $\boxed{a^*(x,R) = \arg\min_a C_{x,R}(a)}$

Different recognizers may have different posterior belief and provide different decisions.

EX:

COST MATRIX

$$C = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

PRIORS

$$\pi = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix}$$

POSTERIORS

$$q_r = \begin{bmatrix} P(C=1|x_r,R) \\ P(C=1|x_r,R) \\ P(C=1|x_r,R) \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.25 \\ 0.35 \end{bmatrix}$$

$$C \cdot q_r = \begin{bmatrix} 0\cdot 0.4 + 1\cdot 0.25 + 2\cdot 0.35 \\ 1\cdot 0.4 + 0.025 + 1\cdot 0.35 \\ 2\cdot 0.4 + 1\cdot 0.25 + 0.035 \end{bmatrix} = \begin{bmatrix} 0.95 \\ 0.75 \\ 1.05 \end{bmatrix}$$

WE SHOULD ASSIGN CLASS 2 EVEN THOUGH HAS A SMALLER POSTERIOR PROBABILITY

	CLASS H_F	CLASS H_T
PREDICTION H_F	0	$C(H_F\|H_T) = C_{fm}$
PREDICTION H_T	$C(H_T\|H_F) = C_{fp}$	0 → ASSUME 0 COST WHEN GUESSING CORRECTLY

CONSIDER A BINARY PROBLEM

$$C_{x,R}(H_T) = 0 \cdot P(H_T|x,R) + P(H_F|x,R) \cdot C_{fp} = P(H_F|x,R) \cdot C_{fp}$$

$$C_{x,R}(H_F) = 0 \cdot P(H_F|x,R) + P(H_T|x,R) \cdot C_{fm} = P(H_T|x,R) \cdot C_{fm}$$

OPTIMAL DECISION IS THE LABELING WITH LOWEST COST

$$\boxed{r(x) = \log \frac{C_{fm} P(H_T|x,R)}{C_{fp} P(H_F|x,R)} \qquad a^*(x,R) = \begin{cases} H_T & \text{if } r(x) > 0 \\ H_F & \text{if } r(x) < 0 \end{cases}}$$

IF R IS A GENERATIVE MODEL WE CAN EXPRESS $\mathcal{R}$ IN TERMS OF COSTS, PRIOR PROBABILITIES AN LIKELIHOODS

$$\mathcal{R}(x) = \log \frac{\pi_T C_{fn}}{(1-\pi_T) C_{fp}} \cdot \frac{f_{x|R}(x|H_T)}{f_{x|R}(x|H_F)}$$

$\mathcal{R}(x) > 0$

$$\log \frac{f_{x|R}(x|H_T)}{f_{x|R}(x|H_F)} \lessgtr - \log \frac{\pi_T C_{fn}}{(1-\pi_T) C_{fp}}$$

THE TRIPLET (π_T, C_{fn}, C_{fp}) IS A WORKING POINT OF AN APPLICATION FOR A BINARY CLASSIFICATION TASK

IT IS REDUNDANT SINCE IT IS POSSIBLE TO BUILD EQUIVALENT APPLICATIONS WHICH HAVE SAME DECISION RULES BUT DIFFERENT COSTS AND PRIORS

FOR EXAMPLE THE APPLICATION $(\tilde{\pi}, 1, 1)$ WHERE $\tilde{\pi} = \frac{\pi_T C_{fn}}{\pi_T C_{fn} + (1-\pi_T) C_{fp}}$

IT IS EQUIVALENT TO (C_{fn}, C_{fp}, π_T)

$\tilde{\pi}$ IS THE EFFECTIVE PRIOR: IF THE CLASS PRIOR FOR H_T WAS $\tilde{\pi}$ AND WE ASSUMED UNIFORM COSTS WE WOULD OBTAIN THE SAME DECISION RULES AS FOR OUR ORIGINAL APPLICATION (SIMPLER TO WORK WITH, WE DO NOT CARE ABOUT C_{fn} AND C_{fp})

SO WE HAVE A RECOGNIZER R WHICH MAKES DECISIONS $a(x, R)$ FOR SAMPLE x WITH CORRECT CLASS c

THEREFORE EACH DECISION HAS COST $C(a(x,R)|c)$

We can evaluate the expected cost (BAYES RISK) of decisions made by our classifier for the evaluation population

$$B = E_{x,c|\mathcal{E}}[C(a(x,R)|c)]$$

EVALUATOR WHITH COMPLETE KNOWLEDGE OF THE DGP

$\mathcal{E}$ DENOTES THE EVALUATION POPULATION WHICH IS ASSUMED TO BE DISTRIBUTED ACCORDING TO $x, c \mid \mathcal{E}$

WE EXPRESS THE BAYES RISK AS:

$$B = E_{x,c|\mathcal{E}}\left[C(a(x,R)|c)\right] = \int f_{x|\mathcal{E}}(x) \sum_{c=1}^{k} C(a(x,R)|c)\, P(c|x=x,\mathcal{E})\, dx$$

AS LONG AS $c|x,R \sim c|x,\mathcal{E}$ THE RISK IS MINIMIZED BY MINIMUM BAYES COST DECISIONS. THE BAYES RISK IN THIS CASE REPRESENTS THE BEST POSSIBLE COST WE WOULD PAY FOR CLASSIFYING TEST DATA, ACCORDING TO $c|x,\mathcal{E} \sim c|x,R$

IN REALITY HOWEVER $c|x,R$ WILL NOT CORRESPOND TO $c|x,\mathcal{E}$ SINCE DISTRIBUTION MISMATCHES MAY EXIST BETWEEN EVALUATION AND TRAIN DATA

WE CAN DEFINE THE BAYES RISK for decisions made by R OVER EVALUATION DATA SAMPLED FOR $x, c|\mathcal{E}$

$$B = E_{x,c|\mathcal{E}}\left[C(a(x,R)|c)\right] = \sum_{c=1}^{k} \pi_c \int C(a(x,R)|c)\, f_{x|c,\mathcal{E}}(x|c)\, dx$$

CONDITIONAL DISTRIBUTION OF EVALUATION POPULATION | THIS DISTRIBUTION REFLECTS THE KNOWLEDGE OF THE ESTIMATOR

WE DON'T HAVE IT!

$\mathcal{E}$ IS MEASURING HOW WELL ARE THE DECISIONS MADE BY R (DATA SAMPLED FROM $x|c,\mathcal{E}$)

HOWEVER IF WE HAVE A SET OF LABELED DATA THEN WE CAN APPROXIMATE THE EXPECTATIONS BY AVERAGING THE COST OVER THE SAMPLE) (SAMPLES GENERATED BY $x|c,\mathcal{E}$, LARGE NUMBER OF SAMPLES PER CLASS)

↓

$$\int C(a(x,R)|c)\, f_{x|c,\mathcal{E}}(x|c)\, dx \approx \frac{1}{N_c} \sum_{i,\, c_i = c} C(a(x_i,R)|c)$$

THE INTEGRAL CAN BE APPROXIMATED BY THE AVERAGE COST COMPUTED OVER THE SAMPLES OF EACH CLASS

WE CAN DEFINE THE EMPIRICAL BAYES RISK AS

$$B_{EMP} = \sum_{c=1}^{K} \frac{\pi_c}{N_c} \sum_{i|c_i=c} C(a(x_i, R)|c)$$

- IT MEASURES THE COST OF OUR DECISIONS OVER THE EVALUATION SAMPLES
- WE CAN USE IT TO COMPARE RECOGNIZERS
- LOWER COST, MORE ACCURACY

LET c_i^* BE THE PREDICTED LABEL OF SAMPLE x_i WHOSE LABEL IS c_i (BINARY CASE)

uDCF (UN-NORMALIZE COST DETECTION FUNCTION)

$$B_{EMP} = \frac{\pi_T}{N_T} \sum_{i|c_i=H_T} C(c_i^*|H_T) + \frac{1-\pi_T}{N_F} \sum_{i|c_i=H_F} C(c_i^*|H_F) = \pi_T C_{fm} \underbrace{P_{fm}}_{\text{FNR}} + (1-\pi_T) C_{fp} \underbrace{P_{fp}}_{\text{FPR}}$$

C_{fm}, C_{fp}, π_T ARE APPLICATION DEPENDENT

A DUMMY SYSTEM THAT ALWAYS ACCEPTS A TEST SEGMENT $(c_r = H_T)$

$P_{fp} = 1 \quad P_{fm} = 0 \rightarrow {}_uDCF = (1-\pi_T) C_{fp}$

A DUMMY SYSTEM THAT ALWAYS REJECTS A TEST SEGMENT $(c_r = H_F)$

$P_{fp} = 0 \quad P_{fm} = 1 \rightarrow {}_uDCF = \pi_T C_{fm}$

NORMALIZED DCF: COMPARE SYSTEM DCF WITH BEST DUMMY SYSTEM

INVARIANT TO SCALING

$$DCF(\pi_T, C_{fm}, C_{fp}) \text{ (NORMALIZED)} = \frac{{}_uDCF(\pi_T, C_{fm}, C_{fp})}{\min(\pi_T C_{fm}, (1-\pi_T) C_{fp})}$$

THE BEST DUMMY SYSTEM CORRESPONDS TO OPTIMAL BAYES DECISION BASED ON PRIOR INFO ALONE

IF WE RESCALE THE uDCF BY $\frac{1}{\pi_T C_{fm} + (1-\pi_T) C_{fp}}$

AND GIVEN $\tilde{\pi} = \frac{\pi_T C_{fm}}{\pi_T C_{fm} + (1-\pi_T) C_{fp}}$ AND $1 - \tilde{\pi}_T = \frac{(1-\pi_T) C_{fp}}{\pi_T C_{fm} + (1-\pi_T) C_{fp}}$

$${}_uDCF(\tilde{\pi}) = \tilde{\pi} P_{fm} + (1-\tilde{\pi}) P_{fp}$$

IN TERMS OF NORMALIZED DCF THE APPLICATION (π_T, C_{fp}, C_{fm}) AND $(\tilde{\pi}, 1, 1)$ ARE EQUIVALENT

LET'S LOOK TO THE ERROR RATE

$$e = \frac{\#\ \text{INCORRECTLY CLASSIFIED}}{\#\ \text{SAMPLES}} = \frac{N_T \cdot P_{fn} + N_F \cdot P_{fp}}{N} = \frac{N_T}{N} P_{fn} + \frac{N_F}{N} P_{fp}$$

CORRESPONDS TO THE DCF OF AN APPLICATION $\left(\frac{N_T}{N}, 1, 1\right)$ WHERE $\frac{N_T}{N}$ IS THE EMPIRICAL PRIOR OF THE EVALUATION SET

FOR SYSTEMS PRODUCING WELL CALIBRATED LOG-LIKELIHOOD RATIOS

$$s = \log \frac{f_{s|c}(s|H_T)}{f_{s|c}(s|H_F)}$$

THE OPTIMAL THRESHOLD (OPTIMAL BAYES DECISION) IS GIVEN BY

$$t = -\log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

LLR ALLOW DISENTANGLING THE CLASSIFIER FROM THE APPLICATION. IN GENERAL SYSTEMS DO NOT PRODUCE WELL CALIBRATED LLRS,

- NON-PROBABILISTIC SCORES
- MISMATCH BETWEEN TRAIN AND TEST POPULATIONS
- NON ACCURATE MODEL ASSUMPTIONS

IN THIS CASES WE SAY THAT THE SCORES ARE MISCALIBRATED, THE THRESHOLD $-\log \frac{\tilde{\pi}}{1-\tilde{\pi}}$ IS NOT OPTIMAL ANYMORE

WE CAN DEFINE THE MINIMUM COST DCF_{min} CORRESPONDING TO THE USE OF THE OPTIMAL THRESHOLD FOR THE EVALUATION SET. WE VARY THE THRESHOLD t TO OBTAIN ALL POSSIBLE COMBINATIONS OF P_{fn} AND P_{fp} FOR THE EVALUATION SET. WE TAKE t CORRESPONDING TO THE LOWEST DCF. DCF_{min} IS THE COST THAT WE WOULD PAY IF WE KNEW BEFOREHAND THE OPTIMAL THRESHOLD FOR THE EVALUATION

WE CAN THINK OF THIS AS A MEASURE OF THE QUALITY OF THE CLASSIFIER

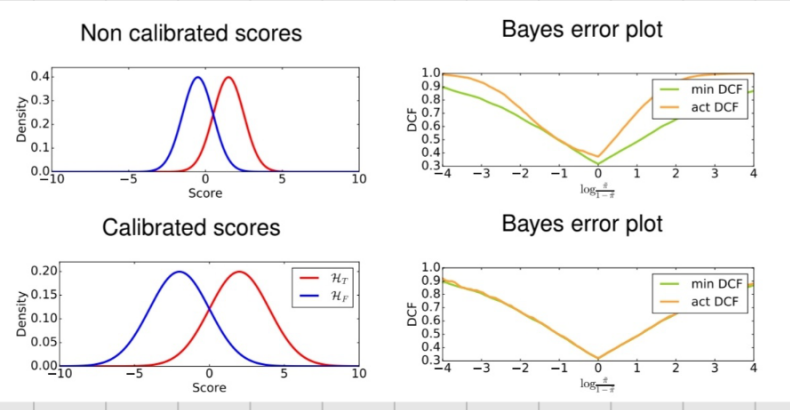
WE CAN ALSO COMPUTE THE ACTUAL DCF OBTAINED USING THE THRESHOLD CORRESPONDING TO THE EFFECTIVE PRIOR $\tilde{\pi}$

DIFFERENCE BETWEEN ACTUAL AND MINIMUM DCF REPRESENTS LOSS DUE TO SCORE MIS. CALIBRATION

WE CAN ALSO COMPARE DIFFERENT SYSTEMS OVER DIFFERENT APPLICATIONS THROUGH BAYES ERROR PLOTS

THESE PLOTS CAN BE USED TO REPORT ACTUAL AND MINIMUM DCF FOR DIFFERENT APPLICATIONS

WE CAN PLOT THE DCF AS A FUNCTION OF PRIOR LOG-ODDS $\log \frac{\tilde{\pi}}{1-\tilde{\pi}}$



TO REDUCE MIS-CALIBRATION WE CAN ADOPT DIFFERENT STRATEGIES

- WE CAN USE A VALIDATION SET TO FIND A (CLOSE-TO) OPTIMAL THRESHOLD FOR A GIVEN APPLICATION
- MORE GENERAL APPROACH LOOK FOR FUNCTIONS THAT TRANSFORM THE CLASSIFIER SCORES s INTO APPROXIMATELY WELL CALIBRATED LLR, IN A WAY THAT IS AS MUCH AS POSSIBLE INDEPENDENT FROM TARGET APPLICATION i.e. COMPUTE A TRANSFORMATION FUNCTION f THAT MAPS THE CLASSIFIER SCORES s TO WELL CALIBRATED SCORES $s_{CAL} = f(s)$

SCORE MODEL: PRIOR WEIGHTED LOGISTIC REGRESSION

WE CONSIDER NON CALIBRATED SCORES AS FEATURES

$$f(s) = as + \gamma$$

SINCE $f(s)$ SHOULD PRODUCE WELL CALIBRATED SCORES IT CAN BE INTERPRETED AS:

$$f(s) = \log \frac{f_{s|c}(s|H_T)}{f_{s|c}(s|H_F)}$$ LLR FOR THE 2 CLASS HYPOTHESIS

THE CLASS POSTERIOR PROBABILITIES FOR $\tilde{\pi}$ CORRESPOND TO

$$\log \frac{P(c=H_T|s)}{P(c=H_F|s)} = as + \gamma + \log \frac{\tilde{\pi}}{1-\tilde{\pi}} = as + \beta$$

WE CAN SET $\pi_T = \tilde{\pi}$ TO LEARN THE MODEL PARAMETERS a, β OVER OUR TRAINING SCORES (CALIBRATION SET)

TO RECOVER $f(s)$ WE WILL NEED TO COMPUTE

$$f(s) = as + \gamma - as + \gamma - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

EVEN THOUGH WE HAVE TO SPECIFY A PRIOR $\tilde{\pi}$ AND WE ARE EFFECTIVELY OPTIMIZING THE CALIBRATION FOR A SPECIFIC APPLICATION $\tilde{\pi}$, THE MODEL WILL OFTEN PROVIDE GOOD CALIBRATION FOR A WIDER RANGE OF DIFFERENT APPLICATION

SO WE NEED A CALIBRATION SET TO ESTIMATE THE TRANSFORMATION:

- MIS CALIBRATION DUE TO NON-PROBABILISTIC SCORES OR TO OVERFITTING/UNDERFITTING MODELS: THE CALIBRATION CAN BE TAKEN FROM THE TRAINING SET MATERIAL
- MIS CALIBRATION DUE TO MISMATCH BETWEEN TRAINING AND EVALUATION POPULATION: THE CALIBRATION SET SHOULD MIMIC THE EVALUATION POPULATION

SUPPORT VECTOR MACHINE (SVM)

HARD MARGIN

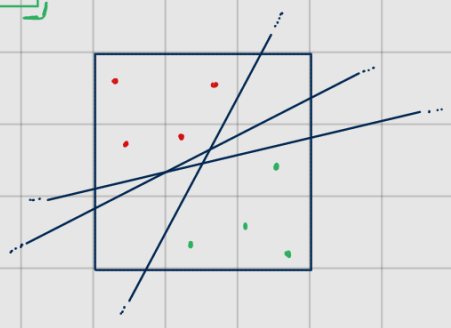
MAIN DIFFERENCE WRT LR

SVM IS A DISCRIMINATIVE NON PROBABILISTIC CLASSIFIER

SVM CAN BE CAST AS A GENERALIZED RISK MINIMIZATION PROBLEM

IT PROVIDES A NATURAL WAY TO ACHIEVE NON LINEAR SEPARATION WITHOUT THE NEED TO EXCPLICITELY EXPAND OUR FEATURES. HOWEVER ITS OUTPUT CANNOT BE INTERPRETED AS CLASS POSTERIOR PROBABILITY

GIVEN 2 LINEARLY SEPARABLE CLASSES WE CAN FIND AN INFINITE NUMBER OF SEPARATING HYPERPLANES



INTUITIVELY WE CAN SELECT THE HYPERPLANE THAT SEPARATES THE CLASSES WITH THE HIGHEST MARGIN (DISTANCE OF THE CLOSEST POINT WRT THE SEPARATION HYPERPLANE)

$f(x) = w^T x + b$ FUNCTION REPRESENTING THE SEPARATION SURFACE

$d(x_i) = \frac{|f(x_i)|}{\|w\|}$ DISTANCE OF x_i FROM HYPERPLANE

$z_i = \begin{cases} +1 & c_i = H_T \\ -1 & c_i = H_F \end{cases}$

SINCE CLASS ARE SEPARABLE WE CONSIDER SOLUTIONS THAT CORRECTLY CLASSIFY ALL POINTS

$f(x_i) > 0$ IF $c_i = H_T$

$f(x_i) < 0$ IF $c_i = H_F$

$$d(x) = \frac{|f(x)|}{\|w\|} = \frac{|z_i (w^T x + b)|}{\|w\|}$$

THE MAXIMUM MARGIN HYPERPLANE IS THE ONE WHICH MAXIMIZES THE MINIMUM DISTANCE OF ALL POINTS FROM THE HYPERPLANE

$$w^*, b^* = \arg \max_{w,b} \min_{i \in \{1 \dots m\}} d(x_i) = \arg \max_{w,b} \min_{i \in \{1 \dots m\}} \frac{|z_i (w^T x_i + b)|}{\|w\|}$$

FOR VALUES w, b WHICH CORRECTLY SEPARATES THE 2 CLASSES WE HAVE THAT

$z_i (w^T x + b) > 0 \quad \forall x_i$ AND $\min_i z_i (w^T x + b) \geq 0$

SAID THAT WE WILL NOW CONSIDER AN EQUIVALENT FORMULATION:

$$w^*, b^* = \arg\max_{w,b} \min_i \frac{|z_i (w^T x_i + b)|}{\|w\|}$$

$$= \arg\max_{w,b} \frac{1}{\|w\|} \min_i \left[z_i (w^T x_i + b) \right]$$

IT HOLDS SINCE ① WE DON'T NEED $\|w\|$ TO COMPUTE THE MINIMUM, SINCE IT DOESN'T CHANGE ② $z_i (w^T x_i + b)$ WILL ALWAYS GIVE A POSITIVE RESULT. IF IT DOESN'T IT MEANS THAT IT IS NOT AN OPTIMAL SOLUTION AND WE DON'T CONSIDER IT

WE FURTHER TRANSFORM THE PROBLEM

$$\frac{1}{\|w\|} \min_i \left[z_i (w^T x_i + b) \right] = \frac{1}{\|aw\|} \min_i \left[z_i (a w^T x_i + a b) \right] \quad \text{for } a > 0$$

THUS IF (w^*, b^*) IS AN OPTIMAL SOLUTION ALSO (aw^*, ab^*) WILL BE

THE COLLECTION OF VALUES (aw^*, ab^*) FORMS AN EQUIVALENCE CLASS OF EQUIVALENT SOLUTIONS, FOR EACH OF THESE EQUIVALENCE CLASSES WE ARE FREE TO SELECT ANY OF THE EQUIVALENT SOLUTIONS. IN PARTICULAR WE RESTRICT OUR PROBLEM TO SOLUTION FOR WHICH

$$\min_i z_i (w^T x + b) = 1$$

AND FOR ALL TRAINING POINTS WE'LL HAVE

$$z_i (w^T x_i + b) \geq 1$$

SO NOW THE PROBLEM BECOMES:

$$\arg\max_{w,b} \frac{1}{\|w\|} \quad \text{S.T.} \begin{cases} z_i (w^T x_i + b) \geq 1 \ * \\ \min_i z_i (w^T x_k + b) = 1 \end{cases} \quad i = 1 \dots m$$

WE CAN DROP THIS SINCE AN OPTIMAL SOLUTION OF * WOULD SATISFY IT

PRIMAL FORMULATION

PRIMAL OBJECTIVE

$$L_p(w,b) = \underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 \quad \text{S.T.} \quad z_i(w^T x_i + b) \geq 1 \quad i = 1 \dots m$$

IT IS A CONVEX QUADRATIC PROGRAMMING PROBLEM

TO SOLVE IT WE CONSIDER A LAGRANGIAN FORMULATION OF THE PROBLEM

$$L(w, b, a_i) = \frac{1}{2}\|w\|^2 - \sum_i^m a_i \left[z_i (w^T x_i + b) - 1 \right]$$

THE OPTIMAL SOLUTION IS OBTAINED BY MINIMIZING L WRT w, b REQUIRING THAT:

EITHER THE DERIVATIVES WRT a_i VANISH GIVEN $a_i \geq 0$ OR $a_i = 0$

SINCE THE PROBLEM IS CONVEX WE CAN EQUIVALENTLY MAXIMIZE L WRT a_i REQUIRING THAT THE DERIVATIVES WRT w, b VANISH GIVEN $a_i \geq 0$

SETTING THE DERIVATIVES OF L WRT w, b GIVES

$$w = \sum_i^m a_i z_i x_i \qquad 0 = \sum_i^m a_i z_i$$

BY REPLACING THIS CONSTRAINTS WE GET:

DUAL SVM PROBLEM:

$$L_D(a) = \max \sum_i^m a_i - \frac{1}{2} \sum_i^m \sum_j^m a_i a_j z_i z_j x_i^T x_j$$

$$\text{S.T.} \quad a_i \geq 0 \ \forall i, \qquad \sum_i^m a_i z_i = 0$$

→ MATRIX FORM

$$H_{ij} = z_i z_j x_i^T x_j$$

$$L_D(a) = a^T \mathbf{1} - \frac{1}{2} a^T H a$$

FOR ANY FEASIBLE SOLUTION OF THE PRIMAL PROBLEM w, b AND ANY FEASIBLE SOLUTION OF THE DUAL PROBLEM a WE HAVE

$$L_D(a^*) \leq L_P(w^*, b^*)$$

$$L_P(w^*, b^*) - L_D(a^*) \geq 0$$

→ IS CALLED DUALITY GAP AND IT IS 0 FOR OPTIMAL SOLUTION!

$$L_P(w^*, b^*) = L_D(a^*)$$

→ OPTIMAL SOLUTION

A SOLUTION IS OPTIMAL IFF IT SATISFIES THE KARUSH - KUHN - TUCKER CONDITIONS KKT (NECESSARY AND SUFFICIENT)

$$\nabla_w L(w,b,a) = w - \sum_i^n a_i z_i x_i = 0$$
$$\frac{\partial L(w,b,a)}{\partial b} = -\sum_i^n a_i z_i = 0$$

AT THE OPTIMAL PRIMAL SOLUTION THE GRADIENT OF THE LAGRANGIAN WRT w, b BECOMES $\emptyset$

$$z_i (w^T x_i + b) - 1 \geq 0 \quad \forall i$$
$$a_i \geq 0 \quad \forall i$$

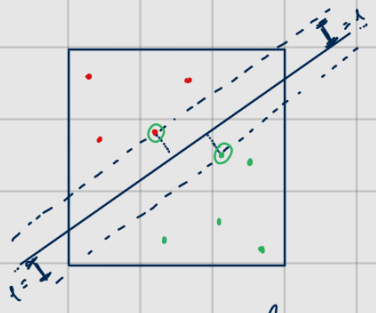
BOTH THE PRIMAL SOLUTION w, b AN THE CORRESPONDING DUAL SOLUTION ARE FEASIBLE

$$a_i [z_i (w^T x_i + b) - 1] = 0 \quad \forall i$$

THE OPTIMAL DUAL SOLUTION MAXIMIZES THE LAGRANGIAN SO EITHER $a_i = 0$ OR $\frac{\partial L}{\partial a_i} = 0 \rightarrow a_i \frac{\partial L}{\partial a_i} = 0$

ALL POINTS THAT DO NOT LIE ON THE MARGIN $(z_i (w^T x_i + b) > 1) \rightarrow a_i = 0$

$a_i \neq 0 \rightarrow$ THE POINT IS ONTHE MARGIN (IT IS A SUPPORT VECTOR)



ONCE WE GET a WE CAN THEN ESTIMATE b THROUGH KKT CONDITIONS

CONSIDERING $w = \sum_i^n a_i z_i x_i$ AND $z_i (w^T x_i + b) = 1$ FOR SUPPORT VECTORS $(\text{so } b = 1 - z_i w^T x_i)$

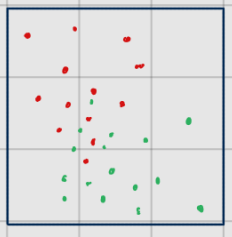
WE GET

$$s(x_r) = w^T x_r + b = \sum_i^n a_i z_i x_i^T x_r + b$$

SINCE FOR VECTORS THAT ARE NOT SUPPORT VECTORS ALPHA IS EQUAL TO ZERO AND SO THEY DO NOT INFLUENCE THE CLASSIFICATION

NOW WE CONSIDER 2 NON SEPARABLE CLASSES

SOFT MARGIN



NO MATTER THE VALUE OF w, SOME POINTS WILL VIOLATE THE CONSTRAINT $z_i (w^T x_i + b) \geq 1$

WE CAN TRY TO MINIMIZE THE # POINTS THAT VIOLATE THIS CONSTRAINT

WE INTRODUCE THE SLACK VARIABLES ξ_i WHICH INDICATES HOW MUCH A POINT IS VIOLATING A CONSTRAINT

$$z_i (w^T x_i + b) \geq 1 - \xi_i \qquad \xi_i \geq 0$$

i.e. WE ALLOW A MARGIN

WE NOW CONSIDER THE FUNCTIONAL

$$\Phi = \sum_{i=1}^{m} \xi_i^{\sigma} \qquad \sigma > 0$$

FOR SUFFICIENTLY SMALL VALUES OF σ, Φ INDICATES THE NUMBER OF POINTS INSIDE THE MARGIN IF WE REMOVED THESE POINTS WE COULD TRAIN A MAXIMUM MARGIN HYPERPLANE OVER THE REMAINING POINTS. FORMALLY THIS CORRESPONDS TO

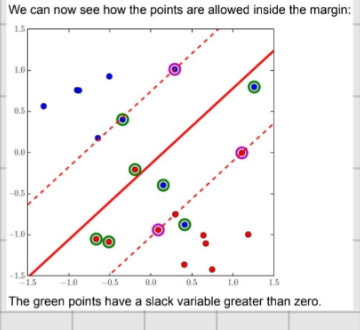
$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + CF\left(\sum_i^m \xi_i^{\sigma}\right) \quad \text{s.t.} \quad \begin{cases} z_i (w^T x_i + b) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0 & \forall i \end{cases}$$

C: CONSTANT; F: CONVEX MONOTONIC FUNCTION

LARGE VALUES OF C WILL ALLOW A LOWER NUMBER OF POINTS INSIDE THE MARGIN (IF LARGE ENOUGH WE GET THE SAME RESULTS OF PREVIOUS SVM MODEL). SMALL VALUES OF C ALLOW A LARGER NUMBER OF POINTS INSIDE THE MARGIN.

WE SIMPLIFY THE PROBLEM CONSIDERING $\sigma = 1$ WHICH GUARANTEES A UNIQUE SOLUTION FOR THE SEPARATION SURFACE AND $F(u) = u$ NOW WE GET

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i \quad \text{s.t.} \quad \begin{cases} z_i (w^T x_i + b) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0 & \forall i \end{cases}$$



ξ_i ACTS AS A PENALTY:

- POINTS INSIDE THE MARGIN HAVE $\xi_i > 0$, MISSCLASSIFIED POINTS HAVE $\xi_i > 1$
- THE FURTHER THE POINT FROM THE HYPERPLANE, THE LARGER IS ξ_i
- $\sum_i^m \xi_i$ UPPER BOUND FOR THE NUMBER OF MISSCLASSIFIED POINTS

HYPERPLANE WITH NO POINTS INSIDE THE MARGIN IS CALLED HARD MARGIN WHILE THIS SOLUTION IS CALLED SOFT MARGIN

THIS CAN BE SOLVED USING KKT CONDITIONS AND LAGRAGIAN FORMULATION

FOR OUR CONCERN THE SOLUTION REMAINS THE SAME BUT WITH AN ADDITIONAL CONSTRAINT:

$$\boxed{a_i \leq C}$$ (BEFORE IT COULD GROW INDEFINITELY)

AND WE CAN COMPUTE THE DUAL PROBLEM

$$\max_a L_D(a) = \max_a \left[\sum_i^m a_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m a_i a_j z_i z_j x_i^T x_j \right] \quad \text{s.t.} \begin{cases} 0 \leq a_i \leq C & \forall i \\ \sum_{i=1}^m a_i t_i = 0 \end{cases}$$

$$L_D(a) = a^T \mathbf{1} - \frac{1}{2} a^T H a$$

THE SCORES ARE STILL

$$s(x_t) = w^T x_t + b = \sum_i^m a_i z_i x_i^T x_t + b$$

THE FORMULA OF THE SUPPORT VECTOR HAS CHANGED:

$$a_i [z_i (w^T x_i + b) - 1 + \xi_i] = 0$$

IF A POINT IS MISSCLASSIFIED AND LIES IN THE MARGIN ξ_i WILL BE 2

SEVERAL METHODS TO SOLVE THE DUAL PROBLEM, IF WE ARE INTERESTED IN LINEAR SEPARATION WE CAN DIRECTLY SOLVE THE PRIMAL SVM PROBLEM

$$\min \frac{1}{2} \|w\|^2 + C \sum_i^m \xi_i \quad \text{s.t.} \begin{cases} z_i (w^T x_i + b) \geq 1 - \xi_i & \forall i \\ \xi_i \geq 0 & \forall i \end{cases}$$

SO DATA ON THE CORRECT SIDE OF THE MARGIN WILL HAVE $\xi_i = 0$ WHILE THE OTHERS WILL HAVE $\xi_i = 1 - z_i (w^T x_i + b)$. SO NOW WE HAVE

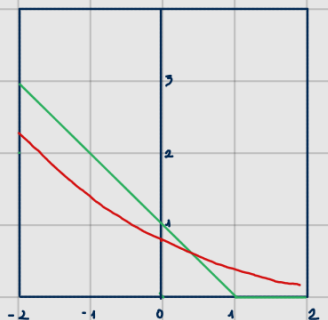
(ERROR OF A SAMPLE)

(CONSTRAINT HAVE BEEN ABSORBED HERE)

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_i^n \max \left[0, 1 - z_i (w^T x_i + b) \right]$$

$$f(s) = \max(0, 1 - s) \quad \text{HINGE LOSS}$$

- LOGISTIC LOSS
- HINGE LOSS



WE CAN RE WRITE THE OBJECTIVE FUNCTION AS:

(DEPENDS ON C)

$$\min_{w,b} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_i^m \max \left[0, 1 - z_i (w^T x_i + b) \right]$$

TO GET THE SEPARATION HYPER PLANE WE CAN SOLVE:

PRIMAL

- MINIMIZE WRT w, b

$$\mathcal{L}_p(w,b) = \frac{1}{2}\|w\|^2 + C\sum_1^m [1 - z_i(w^T x_i + b)]$$

$$\text{S.T.} \begin{cases} a_i \geq 0 \\ \sum_i^m a_i z_i = 0 \end{cases}$$

- PARAMETERS $w \in \mathbb{R}^D,\ b \in \mathbb{R}$
- SCORING COMPLEXITY IS $O(D)$
- $s(x_r) = w^T x_r + b$

- EMBEDDING A NON LINEAR TRANSFORMATION REQUIRES EXPLICITELY DEFINING THE MAPPING FROM R^D TO THE HILBERT SPACE H

$$\Phi : \mathbb{R}^D \to H$$

DUAL

- MAXIMIZE WRT a

$$\mathcal{L}_D(a) = a^T \mathbf{1} - \frac{1}{2} a^T H a$$

$$= \sum_i a_i - \frac{1}{2}\sum_i^{\sim}\sum_j^{\sim} a_i a_j z_i z_j x_i^T x_j$$

$$\text{S.T.} \begin{cases} 0 \leq a_i \leq C \quad \forall i \\ \sum_i^m a_i z_i = 0 \end{cases}$$

- PARAMETERS $a \in \mathbb{R}^N$
- SCORING COMPLEXITY $O(SV)$
- $s(x_r) = \sum_i^{\sim} a_i z_i x_i^T x_r + b$

$$= \sum_{i|a_i>0}^{\sim} a_i z_i x_i^T x_r + b$$

- EMBEDDING A NON LINEAR TRANSFORMATION REQUIRES ONLY DOT PRODUCTS IN THE EXPANDED SPACE

$$\Phi(x_i)^T \Phi(x_j)$$

IF WE HAVE A FUNCTION THAT EFFICIENTLY COMPUTES THE DOT PRODUCT IN THE EXPANDED SPACE:

$$k(x_1, x_2) = \Phi(x_1)^T \Phi(x_2)$$

IS CALLED KERNEL FUNCTION

THEN BOTH TRAINING AND SCORING CAN BE DONE USING JUST k

SCORING BECOMES

$$s(x_r) = \sum_{i=1|a_i>0}^{m} a_i z_i \Phi(x_i)^T \Phi(x_r) + b =$$

$$= \sum_{i=1|a_i>0}^{\sim} a_i z_i k(x_i, x_r) + b$$

↓ KERNEL FUNCTION

A KERNEL FUNCTION ALLOWS TO TRAIN A SVM IN A LARGE DIMENSIONAL HILBERT SPACE WITHOT HAVING TO EXPLICITELY COMPUTE THE MAPPING

EVEN IF THE EXPANSION WAS FINITE, THE COMPLEXITY OF THE PRIMAL MAY BE TOO LARGE FOR THE DUAL PROBLEM THE COMPLEXITY ONLY DEPENDS ON THE NUMBER OF TRAINING POINTS

IN PRACTICE WE ARE COMPUTING A LINEAR SEPARATION SURFACE IN THE EXPANDED SPACE WHICH CORRESPONDS TO A NON LINEAR SEPARATION SURFACE IN THE ORIGINAL FEATURE SPACE

WE CAN DEFINE POLYNOMIAL KERNELS OF DEGREE d AS

$$k(x_1, x_2) = (x_1^T, x_2 + 1)^d$$

GAUSSIAN RADIAL BASIS FUNCTION KERNEL

$$k(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}$$

IF WE KNOW Φ WE CAN DEFINE A CORRESPONDING KERNEL FUNCTION

IN GENERAL WE WANT TO KNOW FOR WHICH KERNEL FUNCTION WE CAN ACTUALLY FIND Φ AND H

MERCER'S CONDITION PROVIDES A SUFFICIENT CONDITION FOR k TO DEFINE A DOT PRODUCT IN AN EXPANDED SPACE

ANYWAY THIS CONDITION DOES NOT TELL US HOW TO CONSTRUCT SUITABLE KERNELS, BUT WE CAN MAKE USE OF WELL KNOWN ONES AND RULES THAT APPLY TO THEM

PRACTICAL CONSIDERATIONS:

- WE NEED TO SELECT A GOOD KERNEL
- KERNELS OFTEN NEED HYPERPARAMETERS (SUCH AS γ FOR RBF). CROSS VALIDATION CAN HELP CHOOSING
- SVM IS NOT INVARIANT UNDER AFFINE TRANSFORMATIONS: FEATURE PRE-PROCESSING CAN BE RELEVANT, OFTEN IS USEFUL TO CENTER AND WHITEN DATA
- SVM SCORES HAVE NO PROBABILISTIC INTERPRETATION

- WE HAVE TO TAKE CARE OF UN-BALANCED DATASETS (WHEN EMPIRICAL TRAINING SET PRIOR IS SIGNIFICANTLY DIFFERENT FROM THE TARGET APPLICATION ONE). IN THIS CASE WEIGTHING THE COST OF DIFFERENT ERRORS MAY BE BENEFICIAL

TO BALANCE THE CLASSES WE HAVE TO USE DIFFERENT VALUES OF C FOR DIFFERENT CLASSES (OR SAMPLES)

$$\underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + \sum_i^m C_i[1 - z_i(w^T x_i + b)]$$

$$\underset{a}{\operatorname{argmax}} \; a^T 1 - \frac{1}{2} a^T H a \quad \text{S.T.} \begin{cases} \sum_i a_i z_i = 0 \\ 0 \le a_i \le C_i \quad \forall i \end{cases}$$

CORRESPONDING DUAL PROBLEM

FOR CLASS BALANCING WE COULD SET $C_i = C_T$ FOR SAMPLES BELONGING TO H_T AND $C_i = C_F$ FOR SAMPLES BELONGING TO H_F

$$C_T = C \cdot \frac{\pi_T}{\pi_T^{EMP}}$$

$$C_F = C \cdot \frac{\pi_F}{\pi_F^{EMP}}$$

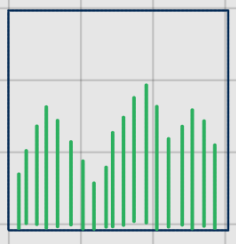
EMPIRICAL PRIORS (SAMPLE PROPORTIONS) COMPUTED OVER THE TRAINING SET

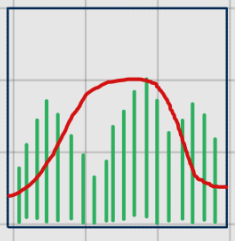
IN PRACTICE WE SIMULATE THE CLASS PROPORTIONS AT TRAINING TIME

- GOOD PERFORMANCE IN BINARY PROBLEMS, DIFFICULT TO EXTEND TO MULTICLASS ONES (BUT POSSIBLE)

GAUSSIAN MIXTURE MODELS (GMM)

THE GAUSSIAN CLASSIFIER IS AN EXAMPLE THAT ASSUMES THAT CLASS CONDITIONAL DISTRIBUTIONS ARE GAUSSIAN. IN MANY CASES HOWEVER THIS ASSUMPTION CAN BE INACCURATE





DIFFERENT DISTRIBUTIONS MAY BE USED IN SUCH CASES. DEPENDING ON THE TASK WE MAY BE ABLE TO IDENTIFY A REASONABLY GOOD FAMILY OF DISTRIBUTIONS. GMM ALLOW APPROXIMATING ANY SUFFICIENTLY REGULAR DISTRIBUTION TO A DESIRED DEGREE. SINCE WE ARE ESTIMATING THE DENSITY FROM THE DATA IT IS REQUIRED A SUFFICIENTLY LARGE AMOUNT OF IT TO GET A GOOD ESTIMATION.

GMM IT IS NOT RESTRICTED TO CLASSIFICATION, FOR EXAMPLE IT CAN PROVIDE AN ALTERNATIVE TO K-MEANS FOR CLUSTERING

A GMM IS A DENSITY MODEL OBTAINED AS A WEIGHTED COMBINATION OF GAUSSIANS

$$f_X(x) = \sum_{c=1}^{k} w_c \mathcal{N}(x; \mu_c, \Sigma_c)$$

THE DISTRIBUTION PARAMETERS ARE:

$M = [\mu_1 \dots \mu_k]$ COMPONENT MEANS

$S = [\Sigma_1 \dots \Sigma_k]$ COMPONENT COVARIANCES

$w = [w_1 \dots w_k]$ WEIGHTS

MOREOVER WE REQUIRE THE FOLLOWING FOR f_X IN ORDER TO BE A DENSITY

$$\int f_X(x) = \int \sum_{c=1}^{k} w_c \mathcal{N}(x; \mu_c, \Sigma_c) = \sum_{c=1}^{k} w_c = 1$$

(SAMPLES THAT WE WANT TO MODEL BY MEANS OF A GMM)

GIVEN A DATASET $D = [x_1 \dots x_m]$ (UNLABELED) WE CAN ASSUME THAT THE SAMPLES HAVE BEEN INDIPENDENTLY GENERATED BY A GMM.

WE ASSUME THAT THE R.V.s DESCRIBING THE SAMPLES X_i ARE I.I.D WITH

$X_i \sim X \sim GMM(M, S, w)$

AS WE DID FOR GAUSSIAN DENSITY WE CAN RESORT TO ML TO ESTIMATE THE MODEL PARAMETERS THAT BEST DESCRIBES THE DATASET D BUT ML FOR GMM IS AN ILL-POSED PROBLEM. THE COMBINATION OF HEURISTICS (TO AVOID DEGENERATE SOLUTIONS FOR WHICH THE LIKELIHOOD IS NOT BOUNDED ABOVE) PROVIDES GOOD DENSITY ESTIMATES

$$\mathcal{L}(\theta) = \prod_{i=1}^{m} f_{x_i}(x_i) = \prod_{i}^{m} GMM(x_i | M, S, w) = \prod_{i=1}^{m} \left(\sum_{c=1}^{k} w_c \mathcal{N}(x_i | \mu_c, \Sigma_c) \right)$$

$$\ell(\theta) = \sum_{i=1}^{m} \left(\log \sum_{c=1}^{k} w_c \mathcal{N}(x_i | \mu_c, \Sigma_c) \right)$$

GMM CAN BE SEEN AS THE MARGINAL OF A JOINT DISTRIBUTION OF DATA POINTS AND CORRESPONDING CLUSTERS:

$$f_{x_i}(x_i) = \sum_{c=1}^{k} w_c \mathcal{N}(x_i; \mu_c, \Sigma_c) = \sum_{c=1}^{k} p(C_i = c_i) f_{x_i | C_i}(x_i | c_i)$$

ALTHOUGH OUR TRAINING SET CANNOT BE WELL MODELED BY A GAUSSIAN DISTRIBUTION WE CAN IMAGINE THAT THE SET CAN BE PARTITIONED INTO SUBSETS IN SUCH A WAY THAT THE DISTRIBUTION OF THE POINTS OF EACH COMPONENT CAN BE MODELED BY A GAUSSIAN P.D.F.

IF WE KNEW THE COMPONENT RESPONSIBLE FOR EACH SAMPLE (ITS CLUSTER LABEL) WE COULD ESTIMATE THE PARAMETERS OF EACH GAUSSIAN BY ML FROM THE POINTS OF EACH CLUSTER

UNFORTUNATELY, IN GENERAL CLUSTERS ARE UNKNOWN. WE TREAT CLUSTER MEMBERSHIP AS AN UNOBSERVED RV. INTUITIVELY WE WANT TO ESTIMATE BOTH CLUSTER ASSIGNMENT AND MODEL PARAMETERS AS TO MAXIMIZE THE MARGINAL DISTRIBUTION OF THE DATA

LET'S CONSIDER A SET OF GMM PARAMETERS $\theta = (M, S, w)$

THE DENSITY FOR SAMPLE x_i AND COMPONENT c IS

$$f_{x_i, c}(x_i, c) = w_c \mathcal{N}(x_i; \mu_c, \Sigma_c)$$

WE CAN COMPUTE COMPONENT POSTERIOR PROBABILITY

$$\gamma_{c,i} = P(C=c_i \mid X=x_i) = \frac{f_{x_i,c_i}(x_i, c)}{f_{x_i}(x_i)} = \frac{w_c \mathcal{N}(x_i; \mu_c, \Sigma_c)}{\sum_{c'} w_{c'} \mathcal{N}(x_i; \mu_{c'}, \Sigma_{c'})}$$

RESPONSABILITY HOW MUCH THE SAMPLE INFLUENCES THE CLUSTER

WE MIGHT DECIDE TO ASSIGNT A POINT TO THE CLUSTER C WITH HIGHEST POSTERIOR PROBABILITY

$$c_i^* = \arg\max_c P(C_i = c \mid X_i = x_i)$$

(1) → WE ASSOCIATE IT TO EACH SAMPLE

(2) WE CAN NOW ESTIMATE BY ML THE NEW GMM PARAMETERS $\theta^{NEW} = (M^{NEW}, S^{NEW}, w^{NEW})$

WE TREAT CLUSTER ASSIGNEMENTS AS IF THEY WERE KNOWN CLASS LABELS

$$\begin{aligned} l(\theta) &= \sum_i^m [\log f_{x_i \mid C_i}(x_i \mid c_i^*) + \log P(c_i = c_i^*)] \\ &= \sum_i^m [\log \mathcal{N}(x_i \mid \mu_{c_i^*}, \Sigma_{c_i^*})] + \sum_i^m [\log w_{c_i^*}] \\ &= l_{\mathcal{N}}(M, S) + l_c(w) \end{aligned}$$

LOG-LIKELIHOOD OF AN MVG CLASSIFICATION MODEL WITH CLASS ASSUMED TO BE c_i^*

LOG LIKELIHOOD OF A CATEGORICAL MODEL WITH PARAMETER w_c

SOLUTION FOR μ_c AND Σ_c IS THUS

$$\begin{cases} \mu_c^* = \frac{1}{N_c} \sum_{i \mid c_i^* = c} x_i \\ \Sigma_c^* = \frac{1}{N_c} \sum_{i \mid c_i^* = c} (x_i - \mu_c^*)(x_i - \mu_c^*)^T \end{cases}$$

SOLUTION FOR w_c IS

$$w_c^* = \frac{N_c}{\sum_{c=1}^k N_c}$$

WE COULD OBTAIN THE UPDATED SET OF MODEL PARAMETERS $\theta^{NEW} = (M^*, S^*, w^*)$

WE COULD ITERATE THE PROCESS BY COMPUTING NEW CLUSTER ASSIGNMENTS USING θ^{NEW} AND USING THE UPDATED ASSIGNMENTS TO UPDATE ONCE AGAIN THE MODEL PARAMETERS

STOPPING WHEN SOME CRITERION IS MET

PROBLEM, WE FORM HARD CLUSTERS (A POINT IS ASSIGNED TO 1 AND ONLY 1 COMPONENT COMPONENT OF GMM)

IF $P(C=c_i \mid X=x_i) \simeq 1$ THAN IT IS CORRECT TO ASSUME THAT THE POINT BELONGS TO THAT COMPONENT, but if $P(C=c_1 \mid X_i=x_i) \simeq P(C=c_2 \mid X=x_i)$ we are doing a crude approximation. BOTH C_1 AND C_2 MIGHT HAVE BEEN RESPONSIBLE FOR THE GENERATION OF X_i

Shortly the algorithm discussed is not maximizing the LIKELIHOOD of the observed samples x_i

WE'LL BRIEFLY SEE A METHOD TO ESTIMATE LOCAL MAXIMUM OF THE LIKELIHOOD, STILL CONSIDERING HARD ASSIGNMENTS

WE FIX $\Sigma_c = I$, $w_c = \frac{1}{k}$. IN THIS CASE:

$$c_i^* = \arg\max_c P(C_i = c \mid X_i = x_i) = \arg\min_c \|x_i - \mu_c\|^2$$

WHICH SUMMARIZES TO: (K-MEANS CLUSTERING ALGORITHM)

- COMPUTE THE COMPONENT C_i^* WHOSE CENTROID $\mu_{C_i^*}$ IS CLOSEST TO OUR POINT AND ASSIGN x_i TO THAT CLUSTER.
- RE-ESTIMATE THE CLUSTER CENTROID FROM THE GIVEN POINTS AND ITERATE UNTIL COVERAGE

EXTEND SOLUTION TO HANDLE SOFT ASSIGNMENTS

CONSIDER THE LOG-LIKELIHOOD OF OUR DATA (EXPLICITATING THE MODEL PARAMETERS)

$$\ell(\theta) = \sum_i^m \log f_{x_i}(x_i \mid \theta) = \sum_i^N \log \left(\sum_{c=1}^k w_c \mathcal{N}(x_i \mid \mu_c, \Sigma_c) \right)$$

COMPUTING THE GRADIENT WRT μ_c WE GET:

$$\mu_c = \frac{\sum_i \gamma_{c,i} x_i}{\sum_i \gamma_{c,i}}$$

IF WE KNEW THE RESPONSABILITIES WE COULD GET μ_c ($\gamma_{c,i}$ DEPENDS ON μ_c)

IF WE INTERPRET IT AS A WEIGTHED EMPIRICAL MEAN, THE WEIGTH OF EACH SAMPLE IS THE CORRESPONDING RESPONSABILITY

$$N_c = \sum_{i=1}^N \gamma_{c,i}$$

0 ORDER STATISTIC

$$F_c = \sum_{i=1}^N \gamma_{c,i} x_i$$

1st ORDER STATISTIC

ADOPTING A SIMILAR STRATEGY FOR THE COVARIANCE MATRIX WE CAN GET THE TERMS

$$S_c = \sum_i \gamma_{c,i} \, x_i x_i^T$$

2nd ORDER STATISTIC

WEIGHTS CAN BE RE-ESTIMATED AS:

$$w_c = \frac{N_c}{N} \qquad N = \sum_c^k N_c$$

SINCE WE AR NOT GIVEN $\gamma_{c,i}$ WE FOLLOW THE SAME PROCEDURE OF THE HARD-ASSIGNMENT SOLUTION: (EXPECTATION-MAXIMIZATION EM)

- GIVEN θ WE ESTIMATE $\gamma_{c,i} \; \forall i$ (THE CLUSTER POSTERIOR PROBABILITIES)

$$\gamma_{c,i} = P(C_i = c \mid X_i = x_i)$$

- GIVEN THE RESPONSABILITIES WE RE-ESTIMATE THE GMM PARAMETERS ϑ

EXPECTATION MAXIMIZATION

DIRECT MAXIMIZATION OF THE GMM LOG-LIKELIHOOD PROVED DIFFICULT BECAUSE OF THE FORM OF THE MARGINAL LOG-DENSITY

$$\log f_X(x|\theta) = \log \left(\sum_{c=1}^k w_c \, N(x \mid \mu_c, \Sigma_c) \right)$$

ON THE CONTRARY WE HAVE SEEN THAT THE OPTIMIZATION OF JOINT CLUSTER-FEATURES IS STRAIGHT-FORWARD: THE JOINT LIKELIHOOD CONSISTS OF THE PRODUCT OF CLUSTER CONDITIONAL NORMAL LOG-DENSITIES AND CLUSTER PRIOR PROBABILITIES

$$\log f_{X,C}(x, c) = \log N(x \mid \mu_c, \Sigma_c) + \log w_c$$

SIMPLE EXPRESSION

EM IS AN ITERATIVE PROCEDURE SUITED FOR THE ML ESTIMATION OF THE PARAMETERS OF COMPLEX LIKELIHOODS $f_X(x|\theta)$ THAT CAN BE EXPRESSED THROUGH MARGINALIZATION OF JOINT LIKELIHOODS $f_{X,H}(x, h|\theta)$

$$f_x(x) = \int f_{x,H}(x,h)\,dh = \int f_{x|H}(x|h)\, f_H(h)\,dh$$

x is a RV with realization x

(LATENT)

H HIDDEN RANDOM VARIABLE

R.V. FOR WHICH WE DO NOT HAVE A REALIZATION

THE EM TRANSFORMS THE MAXIMIZATION OF A LOG-LIKELIHOOD $\log f_x(x|\theta)$ INTO A SEQUENCE OF OPTIMIZATIONS OF THE JOINT LOG LIKELIHOOD $\log f_{x,H}(x,h|\theta)$

GMM FOR CLASSIFICATION

WE ASSUME THAT SAMPLES OF CLASS c ARE GENERATED BY A GMM WITH PARAMETERS (μ_c, Σ_c, w_c)

FOR EACH CLASS WE WANT TO RECOGNIZE WE CAN COMPUTE THE ML ESTIMATE OF A GMM FOR THE SAMPLES OF THAT CLASS. WE CAN THEN USE THE ESTIMATED DENSITIES TO COMPUTE CLASS CONDITIONAL LOG-LIKELIHOODS AND CLASS POSTERIOR DISTRIBUTION(S) OR LOG-LIKELIHOOD RATIO(S)

- CHOOSE THE NUMBER OF COMPONENTS FOR EACH CLASS k_c

  FIT A GMM WITH k_c COMPONENTS TO SAMPLES OF EACH CLASS

$$x_i | C_i = c \sim GMM(\mu_c, \Sigma_c, w_c)$$

$$f_{x_t|C_t}(x_t|c) = \sum_{k=1}^{k_c} w_k \mathcal{N}(x_t|\mu_{c,k}, \Sigma_{c,k})$$

$$P(C_t = c | x_t = x_t) = \frac{P(C_t = c) \sum_{k=1}^{k_c} w_{c,k} \mathcal{N}(x_t|\mu_{c,k}, \Sigma_{c,k})}{\sum_{c'} P(C_t = c') \sum_{k=1}^{k_{c'}} w_{c',k} \mathcal{N}(x_t|\mu_{c',k}, \Sigma_{c',k})}$$

NON-OF-THE-OTHERS CLASS

WE CAN EXPLOIT GMMS FOR OPEN-SET MULTICLASS PROBLEMS

THE DIFFICULTY CONSISTS IN BUILDING A MODEL FOR THE "NOTO" CLASS WHICH IS GENERALLY VERY HETEROGENEOUS

THIS ISSUE CAN BE ALLEVIATED THANKS TO GMM

WE CAN ASSUME THAT SAMPLES OF KNOWN CLASSES ARE MVG DISTRIBUTED AND THEN WE COLLECT A LARGE SET OF UNLABELED SAMPLES OF THE "NOTA" CLASS WHICH WILL BE MODELED THROUGH A GMM.

DIAGONAL COVARIANCES

WE CAN TRAIN GMM WITH DIAGONAL COVARIANCE MATRICES TO REDUCE NUMBER OF PARAMETERS TO ESTIMATE. (DOES NOT CORRESPOND TO THE NAIVE BAYES ASSUMPTION)

THE NAIVE BAYES ASSUMPTION WOULD CORRESPOND TO TRAINING A DIFFERENT GMM FOR EACH SET OF FEATURES THAT ARE ASSUMED INDEPENDENT FROM THE OTHERS

WE CAN ALSO ASSUME THAT ALL COMPONENTS OF A GMM HAVE THE SAME COVARIANCE MATRIX (TIED GMM)

LBG ALGORITHM

(1) SPLIT THE COMPONENTS OF A G-COMPONENT GMM:

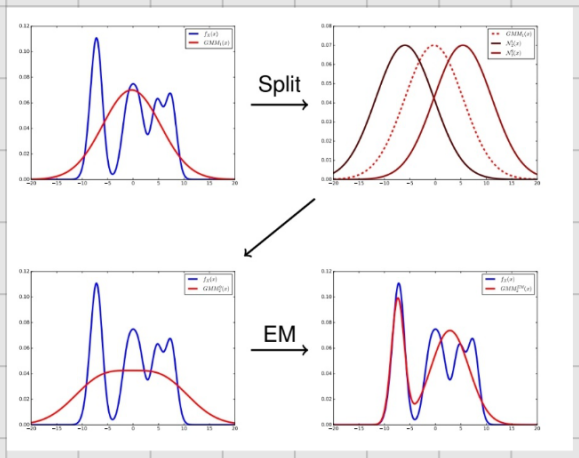
- $\mu_c^+ = \mu_c + \varepsilon$
- $\mu_c^- = \mu_c - \varepsilon$

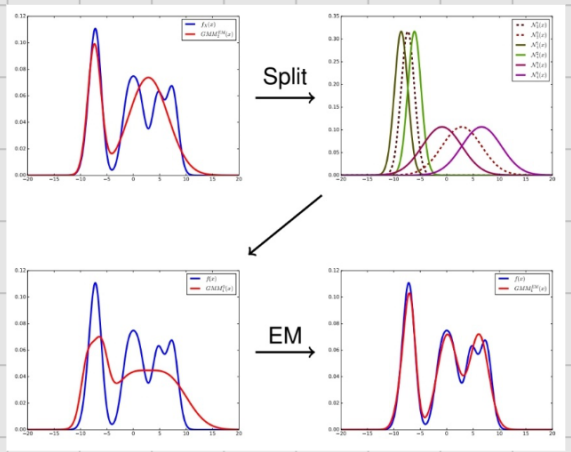
A GOOD VALUE FOR ε CAN BE A DISPLACEMENT ALONG THE PRINCIPAL EIGENVECTOR OF Σ_c

TO OBTAIN AN INITIAL 2G-COMPONENTS GMM

(2) RUN EM UNTIL CONVERGENCE FOR THE 2G COMPONENTS

(3) ITERATE UNTILL THE WANTED NUMBER OF GAUSSIANS IS NEEDED





PROBLEM: HOW MANY GAUSSIANS? MORE COMPONENTS WILL INCREASE THE LIKELIHOOD, BUT WE CANNOT CHOOSE JUST LOOKING ONLY AT IT. WE CAN RESORT TO CROSS VALIDATION