

1. DIMENSIONALITY REDUCTION

DEF: COMPUTE A MAPPING FROM A m -DIMENSIONAL SPACE TO A m' -DIMENSIONAL SPACE WITH $m' < m$

GOALS: COMPRESS INFORMATION, REMOVE NOISE, DATA VISUALIZATION

SIMPLIFY CLASSIFICATION (REDUCE OVERTFITTING, 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.



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

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

We want to find P , a subspace of \mathbb{R}^m which preserves most of the information $P \in \mathbb{R}^{m \times m}$ whose columns are ORTHONORMAL

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

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

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

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:

$$\frac{1}{k} \sum_i^k \|z_i - P P^T z_i\|^2$$

$$z_i = x_i - \bar{x}$$

THEREFORE WE TRY TO SOLVE:

$$P^* = \underset{P}{\operatorname{argmin}} \frac{1}{k} \sum_i^k \|z_i - P P^T z_i\|^2$$

LET'S REWRITE THE OBJECTIVE FUNCTION:

$$\begin{aligned} L(P) &= \frac{1}{k} \sum_i^k \|z_i - P P^T z_i\|^2 = \frac{1}{k} \sum_i^k z_i z_i^T - z_i^T P P^T z_i + z_i^T P P^T P z_i \\ &= \frac{1}{k} \sum_i^k z_i z_i^T - z_i^T P P^T z_i \quad \left. \begin{array}{l} \text{=} I \\ \text{Tr}(W) = \text{Tr}(z^T W) \end{array} \right\} \text{Tr}(AB) = \text{Tr}(BA) \\ &= \frac{1}{k} \sum_i^k z_i z_i^T - \text{Tr}(P^T z_i z_i^T P) \end{aligned}$$

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

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

IT CAN BE SHOWN THAT THE CORRECT SOLUTION IS GIVEN BY P WHOSE COLUMNS ARE THE m EIGENVECTORS OF $\frac{1}{k} \sum_i^k z_i z_i^T$ 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 \sqrt{m} FIRST TRANSFORMED 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 t OF THE VARIANCE OF THE DATA

$$\min_m m \text{ S.T.}$$

$$\frac{\sum_i^m \lambda_i}{\sum_i^m \lambda_i} \geq t$$

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

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

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

SAMPLES FOR
each class
DATASET
MEAN
CLASS MEAN
 μ

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

TOTAL
SAMPLES

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

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 or 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 NORM 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 $\nabla_w L(w) = 0$. WE GET:

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

THE OPTIMAL SOLUTION IS AN EIGENVECTOR OF $S_w^{-1} S_B$

THE EIGENVALUE CORRESPONDING TO THE SOLUTION w IS $\lambda(w) = L(w)$ (THE VALUE OF THE RATIO WE WANT TO MAXIMIZE)

THUS THE MAXIMUM OF L IS THE EIGENVECTOR OF $S_W^{-1}S_B$ WITH LARGEST EIGENVALUE
 FOR THE BINARY CASE WE HAVE $W \perp S_W^{-1}(M_1 - M_0) \rightarrow$ LINES CONNECTING 2 MEANS AND
 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 ON ITS
 DISTRIBUTION ALONG THE AXIS.

WE LOOK FOR THE MAXIMIZER OF $T_p(\hat{S}_W^{-1} \hat{S}_B)$

THE SOLUTION IS GIVEN BY THE m EIGENVECTORS OF $\hat{S}_W^{-1} \hat{S}_B$ CORRESPONDING TO
 THE m LARGEST EIGENVALUES.

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

LDA ASSUMES GAUSSIAN DISTRIBUTED NOISE

IS OFTEN HELPFUL TO PRE PROCESS OUR DATA USING PCA BEFORE APPLYING LDA

GENERATIVE AND LINEAR QUADRATIC CLASSIFIERS I - GENERATIVE MODELS

CLOSED SET CLASSIFICATION PROBLEM, OBSERVATIONS ARE INDEPENDENT AND IDENTICALLY DISTRIBUTED

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

WE ASSUME x_r TO BE A REALIZATION OF A R.V. X_r

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

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

$$P(C_r = c | X_r = x_r)$$

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

JOINT DISTRIBUTION

$$P(C=c | X=x_r) = \frac{f_{X,C}(x_r, c)}{\sum_{c' \in C} f_{X,C}(x_r, c')}$$

$$f_{X,C}(x_r, c) = f_{X|C}(x_r, c) P_C(c)$$

WE NEED TO COMPUTE IT

PRIOR PROBABILITY
CLASS C, APPLICATION DEPENDENT

WE WILL ASSUME THAT THE DATA OF EACH CLASS CAN BE MODELED BY A MULTIVARIATE GAUSSIAN DISTRIBUTION

MVG

$$(X | C=c) \sim N(\mu_c, \Sigma_c)$$

HOWEVER WE DO NOT KNOW THE VALUES OF THE MODEL PARAMETERS $\Theta = [(\mu_1, \Sigma_1) \dots (\mu_k, \Sigma_k)]$

ON THE OTHER HAND 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

$$(X | C=c, \Theta) \sim N(\mu_c, \Sigma_c)$$

i.e. the class conditional distribution for all observation in a Gaussian with class-dependent mean μ_c and class-dependent covariance matrix Σ_c

WE FOLLOW A FREQUENTIST APPROACH AND THUS WE COMPUTE AN ESTIMATOR $\hat{\Theta}^*$ OF THE MODEL PARAMETERS

$$f_{X|C}(x_r, c) = N(x_r | \mu_c^*, \Sigma_c^*)$$

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

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

$$= \prod_{i=1}^m f_{X_i, C_i}(\mathbf{x}_i, c_i | \Theta) = \prod_{i=1}^m f_{X_i | C_i, \Theta}(\mathbf{x}_i | c_i, \Theta) P(c_i)$$

$$= \prod_{i=1}^m N(\mathbf{x}_i | \mu_{c_i}, \Sigma_{c_i}) P(c_i)$$

$$l(\theta) = \log L(\theta) = \sum_i \log N(x_i | \mu_{c,i}, \Sigma_{c,i}) + \log P(c_i)$$

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

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

\hookrightarrow LOG LIKELIHOOD OF A GAUSSIAN MODEL FOR DATA OF CLASS C

IN ORDER TO MAXIMIZE $l(\theta)$ WE CAN MAXIMIZE $l_c(\mu_c, \Sigma_c)$

SOLVING THE FOLLOWING WILL ALLOW US TO FIND THE MAXIMUM:

$$\begin{cases} \nabla_{\mu_c, \Sigma_c} l_c(\mu_c, \Sigma_c) \\ \nabla_{\mu_c, \Sigma_c} l_c(\mu_c, \Sigma_c) \end{cases} \rightarrow \begin{cases} \Sigma_c = \lambda_c^{-1} = \frac{1}{N_c} \sum_{i|c=c} (x_i - \mu_c)(x_i - \mu_c)^T \\ \mu_c = \frac{1}{N_c} \sum_{i|c=c} x_i \end{cases}$$

CLASS COVARIANCE
CLASS MEAN

NOW WE CAN SAY

$$f_{x|c}(x_r, c) = N(x_r | \mu_c, \Sigma_c^*)$$

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

$$\begin{aligned} \log \frac{P(c=l_1|x_r)}{P(c=l_0|x_r)} &= \log \frac{f_{x|c}(x_r | l_1) \cdot P(c=l_1)}{f_{x|c}(x_r | l_0) \cdot P(c=l_0)} \\ &= \log \frac{f_{x|c}(x_r | l_1)}{f_{x|c}(x_r | l_0)} + \log \frac{\pi}{1-\pi} \\ &\quad llr(x_r) \end{aligned}$$

$llr(x_r)$: REPRESENTS THE RATIO BETWEEN THE LIKELIHOOD OF OBSERVING THE SAMPLE GIVEN THAT IT BELONGS TO l_1 OR l_0 .

WE ASSIGN CLASSES BASED ON:

$$llr(x_r) = \log \frac{f_{x|c}(x_r | l_1)}{f_{x|c}(x_r | l_0)} \gtrsim -\log \frac{\pi}{1-\pi}$$

$$ll(x_L) = \mathbf{x}^T A \mathbf{x} + \mathbf{x}^T b + c \rightarrow \text{THIS DECISION FUNCTION IS QUADRATIC}$$

MULTICLASS

FOR MULTI CLASS PROBLEMS $C \in \{1, \dots, k\}$ WE CAN COMPUTE CLOSED-SET POSTERIOR PROBABILITIES AS

$$P(C=L|x_r) = \frac{f_{x|C}(x_r|L)P(L)}{\sum_{L' \in \{1, \dots, k\}} f_{x|C}(x_r|L')P(L')}$$

OPTIMAL DECISION REQUIRES CHOOSING THE CLASS WITH HIGHEST POSTERIOR PROBABILITY

$$\begin{aligned} C_r^* &= \arg \max_L P(C=L|x_r) \propto f_{x|C}(x_r|L)P(L) \\ C_r^* &= \arg \max_L \log f_{x|C}(x_r|L) + \log P(L) \end{aligned}$$

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

$$f_{x_{(j)}|C}(x_{(j)}|C) = \mathcal{N}(x_{(j)} | \mu_{c,(j)}, \sigma_{c,(j)}^2)$$

WE CAN AGAIN COMPUTE THE ML ESTIMATE

$$L(\theta) \propto \prod_{i=1}^m \prod_{j=1}^D \mathcal{N}(x_{i,(j)} | \mu_{c,(j)}, \sigma_{c,(j)}^2)$$

$$L(\theta) = \int + \sum_{j=1}^D \sum_{c=1}^k \sum_{i|c_i=c} \log \mathcal{N}(x_{i,(j)} | \mu_{c,(j)}, \sigma_{c,(j)}^2)$$

WE CAN OPTIMIZE THE LOG-LIKELIHOOD INDEPENDENTLY FOR EACH COMPONENT

THE ML SOLUTION IS

$$\begin{cases} \mu_{c,(j)} = \frac{1}{N_c} \sum_{i|c_i=c} x_{i,(j)} \\ \sigma_{c,(j)}^2 = \frac{1}{N_c} \sum_{i|c_i=c} (x_{i,(j)} - \mu_{c,(j)})^2 \end{cases}$$

$$f_{x|C}(x|C) = \prod_{j=1}^D \mathcal{N}(x_{(j)} | \mu_{c,(j)}, \sigma_{c,(j)}^2) = \mathcal{N}(x | \mu_C, \Sigma_C)$$

$$\begin{cases} \mu_C = \begin{bmatrix} \mu_{1,(1)} \\ \vdots \\ \mu_{1,(D)} \end{bmatrix} \\ \Sigma_C = \begin{bmatrix} \sigma_{1,1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{D,D}^2 \end{bmatrix} \end{cases}$$

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 ARE TIED

- CLASS INDEPENDENT NOISE: $x_{c,i} = \mu_c + \epsilon_i$, $\epsilon_i \sim N(0, \Sigma)$
- BADLY-CONDITIONED PROBLEMS (LARGE DIMENSIONAL DATA, SMALL NUMBER OF SAMPLES)

TIED COVARIANCE MODEL ASSUMES THAT:

$$f_{x|C}(x|C) = N(x | \mu_c, \Sigma)$$

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

THE ML SOLUTION IS:

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

$$\ln(x) = x^T b + c \rightarrow \text{LINEAR SEPARATION RULE}$$

HIGHLY RELATED TO LDA!

CONSIDERATIONS:

- IF DATA IS HIGH-DIMENSIONAL PCA CAN SIMPLIFY THE ESTIMATION
- NAIVE BAYES MAY SIMPLIFY THE ESTIMATION BUT IF THE DATA IS HIGHLY CORRELATED IT CAN PERFORM BADLY
- PCA ALLOWS REMOVING DIMENSIONS WITH VERY SMALL VARIANCE
- TIED COVARIANCE MODELS CAN CAPTURE CORRELATION BUT MAY PERFORM POORLY WHEN CLASSES HAVE DIFFERENT DISTRIBUTION
- MULTIVARIATE MODELS PERFORM BETTER IF WE HAVE ENOUGH DATA TO RELIABLY ESTIMATE THE COVARIANCE MATRIX
- 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 $P(x_r = x_r | C_r = c) = \prod_{c,x} \pi_{c,x}$ FOR DIFFERENT HYPOTHESES 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

$$L(\pi) = \prod_{i=1}^m P(x_i = x_i | C_i = c_i) P(C_i = c_i) \\ = \prod_{i=1}^m \pi_{c_i, x_i}$$

$$\begin{aligned} l(\pi) &= \sum_{i=1}^m \log \underbrace{P(x_i = x_i | C_i = c_i)}_{\pi_{c_i, x_i}} + \{ \\ &= \sum_{\substack{i=1 \\ c_i=c}} \log \pi_{c_i, x_i} + \{ \\ &= \sum_{c=1}^k l_c(\pi_c) + \{ \end{aligned}$$

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

$$\begin{aligned} l_c(\pi_c) &= \sum_{i|C_i=c} \log \pi_{c, x_i} \\ &= \sum_{i=1}^m N_{c,i} \log \pi_{c, x_i} \\ &\quad \text{NUMBER OF TIMES WE OBSERVED } x_i = x_i \end{aligned}$$

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

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

$$\boxed{\pi_{c,i}^* = \frac{N_{c,i}}{N_c}}$$

i.e., THE FREQUENCY VALUE i IN CLASS c

NOW WE CAN SAY THAT:

$$P(x_r = x_r | C_r = c) = \pi_{c,x_r}^*$$

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 FEATURES AND THEN COMBINE THEM

$$\boxed{P(x_r = x_r | C_r = c) = \prod_{c,x_r} \pi_{c,x_r}^{N_{c,x_r}}}$$

OCCURRENCES

$$\left\{ \begin{array}{l} \text{LIBRO 1} \\ (10, 2, 1, \text{ROMANZO}) \end{array}, \begin{array}{l} \text{LIBRO 2} \\ (10, 8, 0, \text{ROMANZO}) \end{array}, \begin{array}{l} \text{LIBRO 3} \\ (1, 20, 2, \text{MANZ}) \end{array} \right\}$$

NOW WE CONSIDER THE CASE WHERE 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

$$P(x=x | c=c) = \frac{\left(\sum_{j=1}^m x_{ij}\right)!}{\prod_{j=1}^m x_{ij}!} \cdot \prod_{j=1}^m \pi_{c,j}^{x_{ij}}$$

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

$$\begin{aligned} l_c(\pi_c) &= \sum_{i|c=c} \sum_{j=1}^m x_{i,j} \log \pi_{c,j} \\ &= \sum_{j=1}^m N_{c,j} \log \pi_{c,j} \end{aligned}$$

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

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

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

THE ML SOLUTION IS AGAIN:

$$\pi_{c,j} = \frac{N_{c,j}}{N_c}$$

RELATIVE FREQUENCY
OF WORD j IN CLASS 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)

$$\begin{aligned} llr(x) &= \log \frac{P(x=x | c=L_1)}{P(x=x | c=L_0)} \\ &= \sum_{j=1}^m x_{1,j} \log \pi_{L_1,j} - \sum_{j=1}^m x_{0,j} \log \pi_{L_0,j} \\ &= x^T b \end{aligned}$$

↓ LINEAR DECISION FUNCTION

CONSIDERATIONS:

- RARE WORDS MAY CAUSE PROBLEMS.
IF A WORD DOES NOT APPEAR ^{WE TAKE} WE WILL ESTIMATE $\pi_{c,j} = 0$. ANY TEST SAMPLE THAT CONTAINS THE WORD WILL HAVE 0 PROBABILITY OF BEING OF CLASS c

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, TRIPLES AND SO ON IF WE WANT PARTIALLY ACCOUNT FOR CORRELATIONS

- WE CAN COMBINE DIFFERENT MODELS THROUGH A NAIVE BAYES ASSUMPTION

LOGISTIC REGRESSION

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

WE START FROM A 2 CLASS PROBLEM USING A TIED GAUSSIAN MODEL

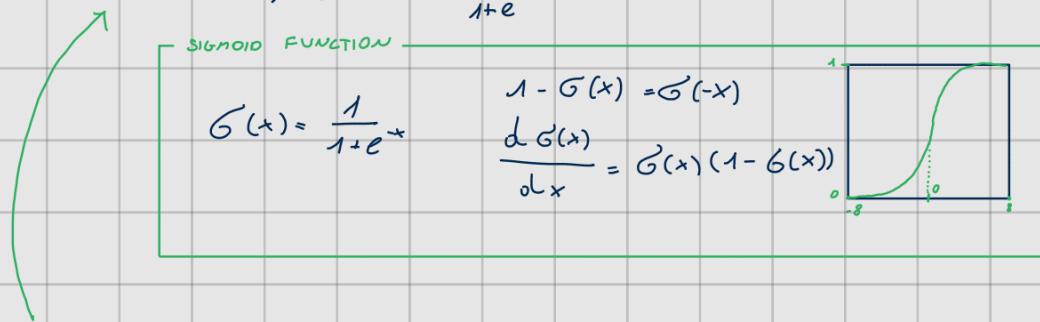
$$\log \frac{P(C=L_1|x)}{P(C=L_0|x)} = \log \frac{f_{x|C}(x|L_1)}{f_{x|C}(x|L_0)} + \log \frac{\pi}{1-\pi} = w^T x + b$$

GIVEN w, b we can compute the expression for the posterior class probability as:

$$P(C=L_1|x, w, b) = e^{(w^T x + b)} (1 - P(C=L_0|x, w, b))$$

we solve for $P(C=L_1|x, w, b)$

$$P(C=L_1|x, w, b) = \frac{1}{1+e^{-(w^T x + b)}} = G(w^T x + b)$$



THE LAST EQUATION PROVIDES A MODEL THAT ALLOWS COMPUTING THE POSTERIOR PROBABILITY FOR L_1 and L_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:

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

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

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

$$y_i = P(C_i=1|x_i, w, b) = G(w^T x + b)$$

$$P(C_i=0|x_i, w, b) = 1 - y_i = 1 - G(w^T x + b) = G(-w^T x - b)$$

$$C_i | x_i, w, b \sim \text{Ber}(G(w^T x + b)) = \text{Ber}(y_i)$$

$$\begin{aligned} L(w, b) &= \prod_{i=1}^n P(C_i=c_i|x_i, w, b) \\ &= \prod_{i=1}^n y_i^{c_i} (1-y_i)^{1-c_i} \end{aligned}$$

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

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

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

BUT RATHER THAN DOING IT WE WILL MINIMIZE

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

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

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 ϵ WHO KNOWS THE ACTUAL LABELS.

$$P(C_i=1|x_i=x_i, \epsilon) = \begin{cases} 1 & \text{IF } c_i=1 \\ 0 & \text{IF } c_i=0 \end{cases} = c_i \quad \sim \text{Ber}(c_i)$$

$$P(C_i=0|x_i=x_i, \epsilon) = \begin{cases} 0 & \text{IF } c_i=1 \\ 1 & \text{IF } c_i=0 \end{cases} = 1 - c_i$$

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

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

$$Q_1 = P(c_i = 1 \mid X_i = x_i, R(w, b)) = y_i = \sigma(w^T x_i + b)$$

$$Q_0 = P(c_i = 0 \mid X_i = x_i, R(w, b)) = 1 - y_i = 1 - \sigma(w^T x_i + b)$$

LOGISTIC REGRESSION LOOKS FOR THE MINIMIZER OF THE AVERAGE CROSS ENTROPY BETWEEN THE DISTRIBUTIONS FOR THE TRAINING SET LABELS OF AN EVALUATOR ε 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 $\text{Ber}(y_i)$ AND THE EMPIRICAL LABEL DISTRIBUTION $\text{Ber}(c_i)$ (THE DISTRIBUTION OF THE EVALUATOR ε)
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

A N 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 l_0 WE HAVE $c_i = -1$ AND FOR SAMPLES BELONGING TO l_1 WE HAVE $Z_i = 1$

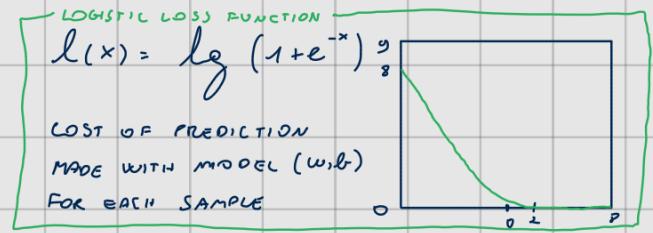
$$J(w, b) = \sum_{i=1}^n H(c_i, y_i)$$

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

GIVEN $s_i = w^T x_i + b$

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

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



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

\rightarrow DECISION RULES TAKE THE FORM $s_i \leq t$

RELATED TO DISTANCE OF x_i FROM THE SEPARATING SURFACE

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

- 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 $z_i \cdot s_i < 0$ 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 SIMPLY 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 INFIMUM $\inf(J(w, b))$ CORRESPONDING TO $\|w\| \rightarrow \infty$

TO MAKE THE PROBLEM SOLVABLE WE CAN LOOK FOR SOLUTION WITH SMALL NORM BY INTRODUCING AN ℓ_2 PENALTY (REGULARIZATION TERM) TO THE OBJECTIVE FUNCTION

THE OBJECTIVE FUNCTION THAT WE MINIMIZE IS

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

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

ALTERNATIVELY WE LOOK FOR THE MINIMIZER OF:

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

↓
REGULARIZATION COEFFICIENT RISK AVERAGED OVER ALL SAMPLES

λ BEING AN HYPER-PARAMETER SHOULD BE SELECTED AS TO OPTIMIZE THE PERFORMANCE OF THE CLASSIFIER. (WE CANNOT COMPUTE 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 REGULARIZED LOGISTIC REGRESSION AND IS AN EXAMPLE OF REGULARIZED RISK MINIMIZATION PROBLEM.

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

THE REGULARIZATION TERM $\mathcal{R}\left(\frac{\lambda}{2} \|w\|^2\right)$ CAN BE SEEN AS A TERM THAT FAVOURS SIMPLER SOLUTIONS. 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

CONSIDERATIONS:

- NON-REGULARIZED MODEL IS INVARIANT TO LINEAR TRANSFORMATIONS OF THE FEATURE VECTORS. REGULARIZED MODEL, 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 EACH 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): $x_i^t = \frac{x_i}{\|x_i\|}$ (OFTEN AFTER CENTERING AND WHITENING)

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 PAIRWISE LINEAR CLASSIFICATION SURFACE BETWEEN CLASSES j AND r

WE OVER-PARAMETERIZED THE MODEL BY INTRODUCING AN EXTRA SET OF PARAMETERS (w_n, b_n) SO THAT WE DO NOT HAVE TO EXPLICITLY ENFORCE $\log \frac{P(C=n|x)}{P(C=r|x)} = 0$

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

$$P(C=j|x) = P(C=r|x) e^{(w_j - w_r)^T x + (b_j - b_r)}$$

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

$$P(C=r|x) = 1 - P(C=j|x) = 1 - \sum_{j \neq r}^k P(C=j|x) e^{(w_j - w_r)^T x + (b_j - b_r)}$$

$$P(C=r|x) = \frac{e^{w_r^T x + b_r}}{\sum_j e^{w_j^T x + b_j}}$$

SO REPEATING THE OPERATION FOR ALL k CLASSES WE GET

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

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

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

SOFTMAX FUNCTION

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

CONSIDERING x_i , ITS CLASS POSTERIOR DISTRIBUTION IS A CATEGORICAL DISTRIBUTION

$$c_i | w, b, x_i = x_i \sim \text{Cat}(y_i) \text{ where } y_{ik} = \frac{e^{w_k^\top + b_k}}{\sum_j e^{w_j^\top + b_j}}$$

REPRESENTS THE DISTRIBUTION OF THE CLASS LABEL ACCORDING TO LR MODEL

LET'S EXPRESS THE LOG-LIKELIHOOD

$$l(w, b) = \sum_i \log P(c_i = c_i | x_i = x_i, w, b)$$

$$\log P(c_i = c_i | x_i = x_i, w, b) = \sum_{k=1}^K z_{ik} \log y_{ik}$$

z_i IS A VECTOR WITH ALL ZEROES BUT ONE 1 AT INDEX c_i

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

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

THE EXPRESSION

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

REPRESENTS THE MULTICLASS ENTROPY BETWEEN THE OBSERVED AND PREDICTED LABEL DISTRIBUTION FOR CLASS x_i

WE FIND w, b THAT MAXIMIZES THE LIKELIHOOD OF THE TRAINING LABELS, SO THAT MINIMIZES AVERAGE 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

$$J(w, b) = - \sum_{i=1}^m \sum_{k=1}^K z_{ik} \log y_{ik}$$

$$= \sum_{i=1}^m \left[\log \left(\sum_{c=1}^K e^{w_c^\top x_i + b_c} \right) - w_{c_i}^\top x_i - b_{c_i} \right]$$

$$= \sum_{i=1}^m \ell(x_i, c_i, w, b)$$

SOFTMAX LOSS

AGAIN WE ADD A REGULARIZATION TERM TO REDUCE OVERFITTING, WE THIS LOOK FOR THE MINIMUM OF

$$R(w, b) = J(w) + \frac{1}{m} \sum_{i=1}^m J(w, b)$$

$$J(w) \text{ COULD BE } J(w) = \frac{1}{2} \sum_i \|w_i\|^2$$

WE PUT THIS IN ORDER TO TAKE INTO ACCOUNT PRIORS THAT ARE DIFFERENT FROM THE EMPIRICAL TRAINING SET PRIOR

...

MODEL EVALUATION FOR CLASSIFICATION

HOW GOOD IS OUR MODEL ON THE TEST SET?

TO ASSESS THIS WE COULD CHECK ACCURACY AND ERROR RATE

$$acc = \frac{\# \text{ CORRECTLY CLASSIFIED}}{\# \text{ SAMPLES}}$$

$$err = 1 - acc$$

BUT ACCURACY CAN BE MISLEADING WHEN CLASSES ARE UNBALANCED

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

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 :

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

CLSIFIER OUTPUT SCORES

GENERATIVE MODELS

$$s = \log \frac{f(x | H_T)}{f(x | H_F)}$$

DISCRIMINATIVE MODELS

$$s = \log \frac{P(H_T | x)}{P(H_F | x)}$$

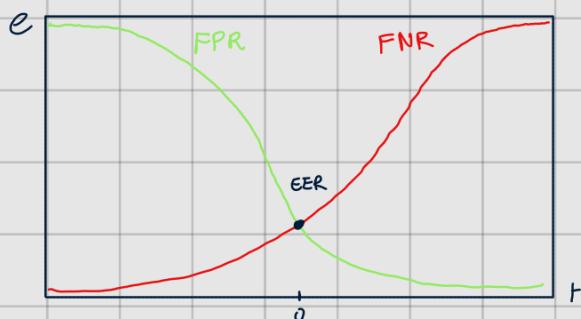
NON PROBABILISTIC MODELS

$$s = w^T x$$

CLASS ASSIGNMENT IS performed by comparing the score with a threshold t
 $s > t \rightarrow H_T$ $s < t \rightarrow H_F$

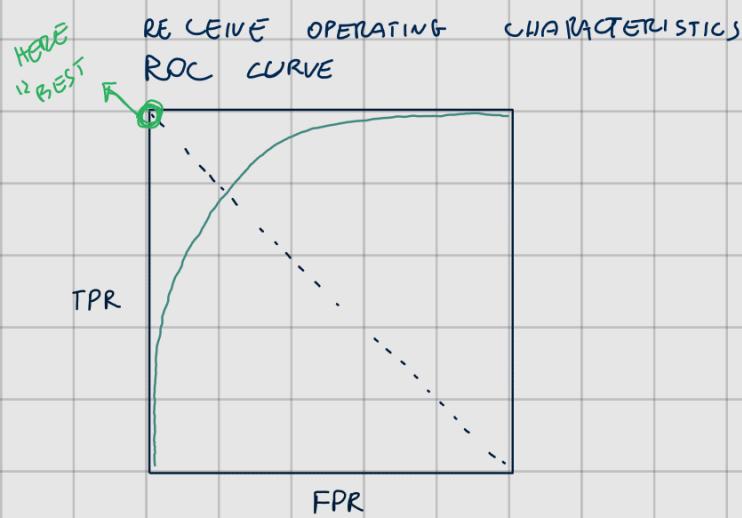
Different thresholds correspond to different error rates

We can visualize the error rates over the threshold

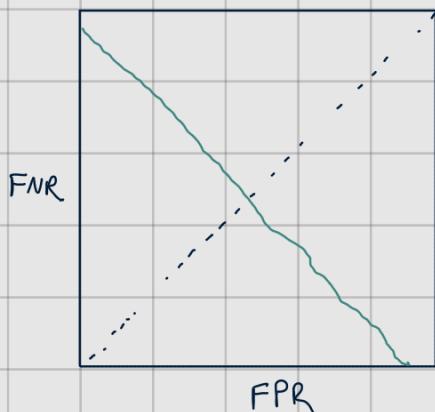


EER: EQUAL ERROR RATE

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



DETECTION ERROR TRADE OFF DET CURVE



The goal of the classifier 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 $a^*(x, R) = \arg \min_a C_{x,R}(a)$
 Different recognizers may have different posterior belief and provide different decisions.

EX:

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

Priors

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

Posterior

$$q_r = \begin{bmatrix} P(C=1|x_r, R) \\ P(C=2|x_r, R) \\ P(C=3|x_r, R) \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.25 \\ 0.35 \end{bmatrix}$$

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

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

PREDICTION H_F 0 $C(H_F|H_T) = C_{f_m}$

PREDICTION H_T $C(H_T|H_F) = C_{f_p}$ 0

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

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

OPTIMAL DECISION IS THE LABELING WITH LOWEST COST

$$a^*(x, R) = \begin{cases} H_T & \text{if } r(x) > 0 \\ H_F & \text{if } r(x) \leq 0 \end{cases}$$

$$r(x) = \log \frac{C_{f_m} P(H_T|x, R)}{C_{f_p} P(H_F|x, R)}$$

IF R IS A GENERATIVE MODEL WE CAN EXPRESS π IN TERMS OF COSTS, PRIOR PROBABILITIES AND LIKELIHOODS

$$\pi(x) = \log \frac{\pi_T c_{fm}}{(1-\pi_T) c_{fp}} \cdot \frac{f_{x|R}(x|H_T)}{f_{x|R}(x|H_F)}$$

$$\log \frac{f_{x|R}(x|H_T)}{f_{x|R}(x|H_F)} \Leftrightarrow -\log \frac{\pi_T c_{fm}}{(1-\pi_T) c_{fp}}$$

THE TRIPLET (π_T, c_{fm}, 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 $\tilde{\pi} = \frac{\pi_T c_{fm}}{\pi_T c_{fm} + (1-\pi_T) c_{fp}}$
IT IS EQUIVALENT TO (c_{fm}, 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_{fm} 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 decision made by our classifier for the evaluation population

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

EVALUATOR WITH COMPLETE KNOWLEDGE OF THE DATA

\mathcal{E} →
REPRESENTS THE EVALUATION POPULATION WHICH IS ASSUMED TO BE DISTRIBUTED
ACCORDING TO $x, c | \mathcal{E}$

WE EXPRESS THE BAYES RISK AS:

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

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 DECISION MADE BY R OVER EVALUATION DATA SAMPLED FOR $x, c | \mathcal{E}$

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

WE DON'T HAVE IT!

CONDITIONAL DISTRIBUTION OF EVALUATION POPULATION
THIS DISTRIBUTION REFLECTS THE KNOWNED OF THE ESTIMATOR

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

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)

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

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

WE CAN DEFINE THE EMPIRICAL BAYES RISK AS

$$B_{\text{Emp}} = \sum_{c=1}^k \frac{\pi_c}{N_c} \sum_{i|c_i=c} C(\alpha(x_i, R)|c)$$

- 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

$$uDCF = \underbrace{\pi_T \sum_{i|c_i=H_T} C(C_i^*|H_T)}_{uDCF} + \underbrace{(1-\pi_T) \sum_{i|c_i=H_F} C(C_i^*|H_F)}_{FNR} = \pi_T C_{fm} p_{fm} + (1-\pi_T) C_{fp} p_{fp}$$

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

A DUNNY 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 DUNNY SYSTEM THAT ALWAYS REJECTS A TEST SEGMENT ($C_r = H_F$)

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

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

↗ INVARIANT TO SCALING

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

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

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

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

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

LET'S LOOK TO THE ERROR RATE

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

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

FOR SYSTEMS PRODUCING WELL CALIBRATED LOG-LIKELIHOOD RATIOS

$$s = \log \frac{f_{x|c}(x|H_c)}{f_{x|c}(x|H_p)}$$

THE OPTIMAL THRESHOLD (OPTIMAL BAYES DECISION) IS GIVEN BY

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

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

- 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 $-\log \frac{\pi}{1-\pi}$ IS NOT OPTIMAL ANYMORE

WE CAN DEFINE THE MINIMUM COST DCF_{min} CORRESPONDING TO THE USE OF THE OPTIMAL THRESHOLD FOR THE EVALUATION SET. IT IS THE COST THAT WE WOULD PAY IF

WE KNEW BEFORE HAND 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 $\bar{\pi}$

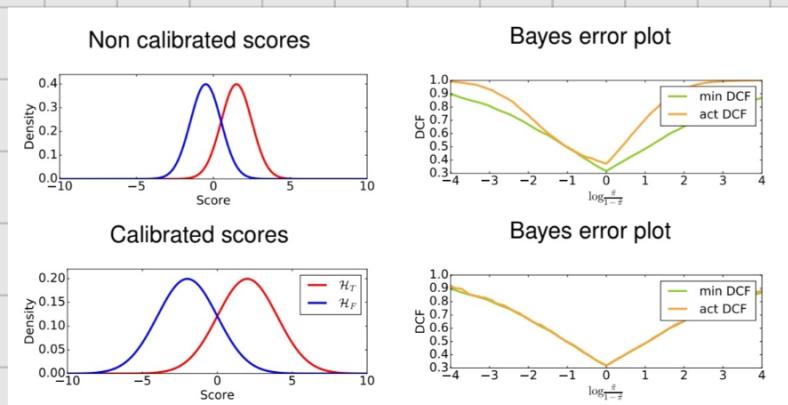
DIFFERENCE BETWEEN ACTUAL AND MINIMUM DCF REPRESENTS LOSS DUE TO SCORE MIS-CALIBRATION

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

$$\log \frac{\pi}{1-\pi}$$



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 $s_{cal} = f(s)$

SOME CALIBRATION APPROACHES

SCORE MODEL: PRIOR WEIGHTED LOGISTIC REGRESSION

WE CONSIDER NON CALIBRATED SCORES AS FEATURES

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

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

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

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

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

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

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

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

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 OFFER PROVIDE GOOD CALIBRATION FOR A WIDER RANGE OF DIFFERENT APPLICATION

SO WE NEED A CALIBRATION SET TO ESTIMATE THE TRANSFORMATION:

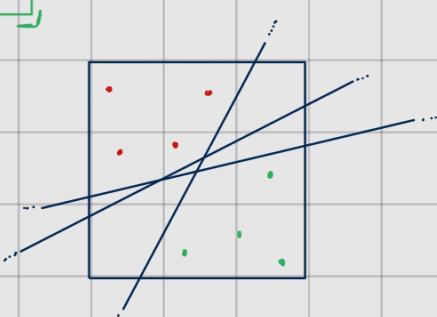
- MIS CALIBRATION DUE TO NON-PROBABILISTIC SCORES OR TO OVERFITTING/UNDERFITTING MODELS: THE CALIBRATION CAN BE TAKEN FROM THE TRAINING SET MATERIAL
- MIS CALIBRATION DUE TO MISMATCH BETWEEN TRAINING AND EVALUATION POPULATION: THE CALIBRATION SET SHOULD MIMIC THE EVALUATION POPULATION

SUPPORT VECTOR MACHINE (SVM)

SVM IS A DISCRIMINATIVE NON PROBABILISTIC CLASSIFIER

SVM CAN BE CAST AS A GENERALIZED RISK MINIMIZATION PROBLEM, FURTHERMORE IT PROVIDES A NATURAL WAY TO ACHIEVE NON LINEAR SEPARATION WITHOUT THE NEED FOR AN EXPLICIT EXPANSION OF OUR FEATURES. HOWEVER, IN CONTRAST WITH LR, THE OUTPUT OF SVM CANNOT BE DIRECTLY INTERPRETED AS (CLASS) POSTERIORS

GIVEN 2 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 WRIT THE SEPARATION HYPERPLANE)

$$f(x) = w^T x + b \quad \text{FUNCTION REPRESENTING THE SEPARATION SURFACE}$$

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

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

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

$$f(x_i) > 0 \quad \text{IF } c_i = H_T$$

$$f(x_i) < 0 \quad \text{IF } c_i = H_F$$

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

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

$$z_i(w^T x_i + b) > 0 \quad \forall x_i \text{ AND } \min_i z_i(w^T x_i + b) > 0$$

SAID THAT WE WILL NOW CONSIDER AN EQUIVALENT FORMULATION:

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

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

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 RESULTS.

IF IT DOESN'T IT MEANS THAT IT IS NOT AN OPTIMAL SOLUTION AND WE DON'T CONSIDER IT

WE FURTHER TRANSFORM THE PROBLEM SINCE ITS SOLUTION IS NOW TRIVIAL

WE CAN OBSERVE THAT

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

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

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

AND FOR ALL TRAINING POINTS WE'LL HAVE

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

SO NOW THE PROBLEM BECOMES

$$\arg \max_{w, b} \frac{1}{\|w\|}$$

OR

$$\text{s.t. } \begin{cases} z_i(w^T x_i + b) \geq 1 \\ \min_i z_i(w^T x_i + b) = 1 \end{cases}$$

EQUIVALENTLY

$$\arg \min_{w, b} \frac{1}{2} \|w\|^2$$

$$\text{s.t. } \begin{cases} z_i(w^T x_i + b) \geq 1 \\ \min_i z_i(w^T x_i + b) = 1 \end{cases}$$

WE CAN PROP

THIS
SINCE AN OPTIMAL

WOULD DIRECTLY
SATISFY ▲

AND FINALLY WE HAVE

$$\underset{w, b}{\text{argmin}} \quad \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad z_i(w^T x_i + b) \geq 1$$

IT IS A CONVEX QUADRATIC PROGRAMMING PROBLEM

TO SOLVE IT WE CONSIDER A LAGRANGIAN FORMULATION OF THE PROBLEM

PRIMAL FORMULATION

$$L_p(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i^\infty \alpha_i [z_i(w^T x_i + b) - 1]$$

BY COMPUTING THE DERIVATIVES OF L wrt w, b, α WE GET THE DUAL SRN PROBLEM:

$$L_D(\alpha) = \max \sum_i^\infty \alpha_i - \frac{1}{2} \sum_i^\infty \sum_j^\infty \alpha_i \alpha_j z_i z_j x_i^T x_j$$

$$\text{s.t. } \alpha_i \geq 0 \quad \forall i \quad \sum_i^\infty \alpha_i z_i = 0$$

MATRIX FORM

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

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

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

$$L_D(\alpha) \leq L_p(w, b)$$

↓

$$L_p(w, b) - L_D(\alpha) \geq 0 \quad \begin{matrix} \rightarrow \text{IS CALLED DUALITY GAP} \\ \text{IF } \alpha \text{ IS FOR OPTIMAL SOLUTION} \end{matrix}$$

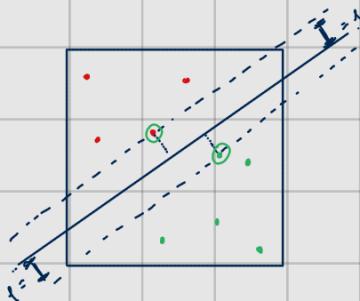
$$L_p(\alpha^*) = L_D(\alpha^*) \Rightarrow \text{OPTIMAL SOLUTION}$$

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

$$\left\{ \begin{array}{l} \nabla_w L(w, b, \alpha) = w - \sum_i \alpha_i z_i x_i = 0 \\ \frac{\partial L(w, b, \alpha)}{\partial b} = - \sum_i \alpha_i z_i = 0 \\ z_i (w^T x_i + b) - 1 \geq 0 \quad \forall i \\ \alpha_i \geq 0 \\ \alpha_i [z_i (w^T x_i + b) - 1] = 0 \quad \forall i \end{array} \right. \begin{array}{l} \text{AT THE OPTIMAL PRIMAL SOLUTION THE GRADIENT OF THE LAGRANGIAN} \\ \text{WRT } w, b \text{ BECOMES } \emptyset \\ \text{BOTH THE PRIMAL SOLUTION } w, b \text{ AND THE CORRESPONDING} \\ \text{DUAL SOLUTION ARE FEASIBLE} \\ \text{THE OPTIMAL DUAL SOLUTION MAXIMIZES THE LAGRANGIAN} \\ \text{SO EITHER } \alpha_i = 0 \text{ OR } \frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \alpha_i \frac{\partial L}{\partial \alpha_i} = 0 \end{array}$$

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

$\alpha_i \neq 0 \rightarrow$ THE POINT IS ON THE MARGIN (IT IS A SUPPORT VECTOR)



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

CONSIDERING $w = \sum_i \alpha_i z_i x_i$ AND $z_i (w^T x_i + b) = 1$ FOR SUPPORT VECTORS (so $b = 1 - z_i w^T x_i$)

WE GET

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

NOW WE CONSIDER 2

NON SEPARABLE CLASSES

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



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

$$z_i(w^T x_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \quad \text{ie. we allow a margin}$$

WE NOW CONSIDER THE FUNCTIONAL

$$\Phi = \sum_{i=1}^m \xi_i \quad G \geq 0$$

FOR SUFFICIENTLY SMALL VALUES OF G , Φ INDICATES THE NUMBER OF POINTS INSIDE THE MARGIN

IF WE REMOVED THESE POINTS WE COULD DRAW A MAXIMUM MARGIN HYPERPLANE OVER THE

REMAINING POINTS. FORMALLY THIS CORRESPONDS TO THE MINIMIZATION OF

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

CONSTANT
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

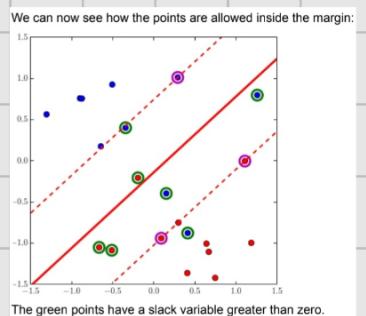
OF C ALLOW A LARGER NUMBER OF POINTS INSIDE THE MARGIN.

WE SIMPLIFY THE PROBLEM CONSIDERING $G = 1$ WHICH GUARANTEES A UNIQUE SOLUTION FOR THE SEPARATION SURFACE AND $F(w) = M$. NOW WE GET

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

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
- $\sum_i \xi_i$ 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 CONDITION AND LAGRAGIAN FORMULATION
 FOR OUR CONCERN THE SOLUTION REMAINS THE SAME BUT WITH AN ADDITIONAL
 CONSTRAINT:

$$\alpha_i \leq C$$

(BEFORE IT COULD GROW INDEFINITELY)

AND WE CAN COMPUTE THE DUAL PROBLEM

$$\max_{\alpha} L_D(\alpha) = \max_{\alpha} \left[\sum_i^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j z_i z_j x_i^T x_j \right] \quad \text{s.t.} \quad \begin{cases} 0 \leq \alpha_i \leq C & b_i \\ \sum_{i=1}^m \alpha_i z_i = 0 \end{cases}$$

$$L_D(\alpha) = \alpha^T z - \frac{1}{2} \alpha^T H \alpha$$

THE SCORES ARE STILL

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

THE FORMULA OF THE SUPPORT VECTOR HAS CHANGED:

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

IF A POINT IS MISSCLASSIFIED AND LIES IN THE MARGIN ξ_i WILL BE 2 (WHICH BRINGS ALL TO ZERO)

SEVERAL METHODS TO SOLVE THE DUAL PROBLEM. IF WE ARE INTERESTED IN LINEAR SEPARATION

WE CAN DIRECTLY SOLVE THE PRIMAL SVA PROBLEM

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

SO DATA ON THE CORRECT SIDE OF THE MARGIN WILL HAVE $\xi_i = 0$ WHILE THE OTHERS

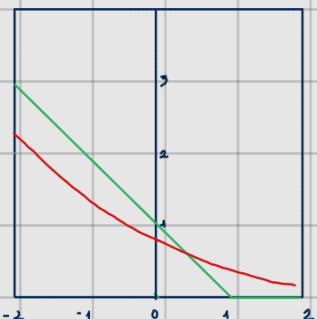
WILL HAVE $\xi_i = 1 - z_i (w^T x_i + b)$. SO NOW WE HAVE

ERROR OF PREDICTION

LOGISTIC LOSS
HINGE LOSS

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

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



WE CAN REWRITE THE OBJECTIVE FUNCTION AS:

DEPENDS ON C

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

TO GET THE SEPARATION HYPER PLANE WE CAN SOLVE:

PRIMAL

MINIMIZE wrt w, b

$$L_p(w, b) = \frac{1}{2} \|w\|^2 + C \sum_i [1 - z_i(w^T x_i + b)]$$

PARAMETERS $w \in \mathbb{R}^D, b \in \mathbb{R}$

SCORING COMPLEXITY IS $O(D)$

$$s(x_r) = w^T x_r + b$$

DUAL

MAXIMIZE wrt α

$$L_D = \alpha^T \mathbf{1} - \frac{1}{2} \alpha^T H \alpha$$

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

$$\text{S.T. } \begin{cases} 0 \leq \alpha_i \leq C \\ \sum_i \alpha_i z_i = 0 \end{cases}$$

PARAMETERS $\alpha \in \mathbb{R}^N$

SCORING COMPLEXITY $O(SV)$

$$\begin{aligned} s(x_r) &= \sum_i \alpha_i z_i x_i^T x_r + b \\ &= \sum_{i: \alpha_i > 0} \alpha_i z_i x_i^T x_r + b \end{aligned}$$

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:

$$k(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$$

IS CALLED KERNEL FUNCTION

THEN BOTH TRAINING AND SCORING CAN BE DONE USING JUST k

$$\begin{aligned} H_{ij} &= z_i z_j \Phi(x_i)^T \Phi(x_j) \\ &= z_i z_j k(x_i, x_j) \end{aligned}$$

$$\begin{aligned} s(x_r) &= \sum_{i: \alpha_i > 0} \alpha_i z_i \Phi(x_i)^T \Phi(x_r) + b \\ &= \sum_{i: \alpha_i > 0} \alpha_i z_i k(x_i, x_r) + b \end{aligned}$$

A KERNEL FUNCTION ALLOWS TO TRAIN A SVM IN A LARGE DIMENSIONAL HILBERG SPACE WITHOUT HAVING TO EXPLICITLY 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

$$k(x_i, x_j) = (x_i^T x_j + 1)^d$$

GAUSSIAN RADIAL BASIS FUNCTION KERNEL

$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

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 SLOPES 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 WEIGHTING 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)

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

$$\underset{\alpha}{\operatorname{argmax}} \quad \alpha^T 1 - \frac{1}{2} \alpha^T H \alpha \quad \text{S.T.} \quad \left\{ \begin{array}{l} \sum_{i=1}^m \alpha_i z_i = 0 \\ 0 \leq \alpha_i \leq C_i \quad \forall i \end{array} \right.$$

CORRESPONDING DUAL PROBLEM

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

$$C_i = C_F \quad \text{FOR SAMPLES BELONGING TO } H_F$$

$$C_F = C \cdot \frac{\bar{n}_T^{\text{EMP}}}{\bar{n}_F^{\text{EMP}}} \quad \text{EMPIRICAL PRIORS (SAMPLE PROPORTIONS)}$$

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

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

THE DISTRIBUTION PARAMETERS ARE:

$$\boldsymbol{\mu} = [\mu_1 \dots \mu_k] \quad \text{COMPONENT MEANS}$$

$$\boldsymbol{\Sigma} = [\Sigma_1 \dots \Sigma_k] \quad \text{COMPONENT COVARIANCES}$$

$$\boldsymbol{w} = [w_1 \dots w_k] \quad \text{WEIGHTS}$$

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

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

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

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

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

$$X_i \sim X \sim \text{GMM}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w})$$

AS WE DID FOR GAUSSIAN DENSITY WE CAN RESORT TO ML TO ESTIMATE THE MODEL PARAMETERS THAT BEST DESCRIBES THIS 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

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

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

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

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

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

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

WE CAN COMPUTE COMPONENT POSTERIOR PROBABILITY

$$y_{c,i} = P(C=c_i | X_i=x_i) = \frac{f_{x_i|C_i}(x_i|c_i)}{\sum_c f_{x_i|C_i}(x_i|c_i)} = \frac{w_c N(x_i; \mu_c, \Sigma_c)}{\sum_c w_c N(x_i; \mu_c, \Sigma_c)}$$
RESPONSIBILITY

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

$$c^* = \arg \max_c P(C=c | X_i=x_i)$$

WE CAN NOW ESTIMATE BY ML THE NEW GMN PARAMETERS $\theta^{\text{NEW}} = (\mu^{\text{NEW}}, s^{\text{NEW}}, w^{\text{NEW}})$

WE TREAT CLUSTER ASSIGNMENTS AS IF THEY WERE KNOWN CLASS LABELS

$$\begin{aligned} l(\theta) &= \sum_i^m [\log f_{x_i|C_i}(x_i|c_i^*) + \log P(c_i^*)] \\ &= \sum_i^m [\log N(x_i; \mu_{c_i^*}, \Sigma_{c_i^*})] + \sum_i^m [\log w_{c_i^*}] \\ &= l_N(\mu, s) + l_c(w) \end{aligned}$$

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

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

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 $\theta^{\text{NEW}} = (\mu^*, s^*, w^*)$

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 GMN)

IF $P(C=c_i | X_i=x_i) \approx 1$ THEN IT IS CORRECT TO ASS THAT THE POINT BELONGS TO THAT COMPONENT, BUT IF $P(C=c_1 | X_i=x_i) \approx P(C=c_2 | X_i=x_i)$ 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 :

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:

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

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 (EXPLICITLY THE MODEL PARAMETERS)

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

COMPUTING THE GRADIENT WRT μ_c WE GET

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

IF WE KNEW THE RESPONSABILITIES WE COULD GET μ_c

IF WE INTERPRET IT AS A WEIGHTED EMPIRICAL MEAN, THE WEIGHT OF EACH SAMPLE IS THE CORRESPONDING RESPONSIBILITY

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

0 ORDER STATISTIC

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

1st ORDER STATISTIC

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

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

2nd ORDER STATISTIC

WEIGHTS CAN BE RE-ESTIMATED AS:

$$w_c = \frac{N_c}{N}$$

SINCE WE ARE NOT GIVEN $y_{c,i}$ WE FOLLOW THE SAME PROCEDURES OF THE HARD-ASSIGNMENT SOLUTION: (EXPECTATION-MAXIMIZATION EM)

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

$$y_{c,i} = P(C_i=c | X_i, \theta)$$
- GIVEN THE RESPONSABILITIES WE RE-ESTIMATE THE GMN PARAMETERS θ

EXPECTATION MAXIMIZATION

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

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

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

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

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 $f_{x,n}(x, h|\theta)$

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

X IS A RV WITH REALIZATION

H HIDDEN RANDOM VARIABLE

RV. 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 $\log f_{X,H}(x,h|\theta)$

E-STEP : COMPUTE $\gamma_{c,i} = P(C_i=c | X_i=x_i, \theta^t)$

COMPUTE POSTERIOR DISTRIBUTION AND
AUXILIARY FUNCTION $Q(\theta, \theta_t)$

M-STEP: MAXIMIZE $Q(\theta, \theta_t)$ wrt $\theta = (M, S, w)$ WITH $\sum_{k=1}^K w_k = 1$

MAXIMIZE $Q(\theta, \theta_t)$ wrt θ
TO OBTAIN θ_{t+1}

$$M_c^* = \frac{\sum_i Y_{c,i} x_i}{\sum_i Y_{c,i}}$$

$$\Sigma_c^* = \frac{\sum_i Y_{c,i} (x_i - M_c)(x_i - M_c)^T}{\sum_i Y_{c,i}}$$

$$w_c^* = \frac{\sum_i Y_{c,i}}{\sum_i \sum_c Y_{c,i}}$$

NEW ESTIMATES

GMM FOR CLASSIFICATION

WE ASSUME THAT SAMPLES OF CLASS c ARE GENERATED BY A GMM WITH PARAMETERS (M_c, S_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 DISTRIBUTIONS OR LOG-LIKELIHOOD RATIOS

- CHOOSE THE NUMBER OF COMPONENTS FOR EACH CLASS k_c

FIT A GMM WITH k_c COMPONENTS TO SAMPLES OF EACH CLASS

$$X_i | C_i = c \sim \text{GMM } (M_c, S_c, w_c)$$

$$f_{X|C=c}(x_i | c) = \sum_{k=1}^{k_c} w_{c,k} N(x_i | M_{c,k}, \Sigma_{c,k})$$

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

WE CAN ALSO EXPLOIT GMM CLUSTERING CAPABILITIES FOR OPEN-SET MULTICLASS CLASSIFICATION.

DIFFICULT TO DEFINE THE "NOT-OTHERS" CLASS WHICH IS USUALLY VERY HETEROGENEOUS

WE COULD MODEL UNLABELED SAMPLES USUALLY AND A LARGE SET OF

UNLABELED DATA FOR THE "NOTO" CLASS USING GMM. GMM WILL FIND HOMOGENEOUS

CLUSTERS OF THE "NOTO" POPULATION.

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)

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