- **ESTIMATION THEORY**

  - □ Noisy deterministic signals
    - ex transmission signal noise (known source)
  - □ Noisy parametric signals
    - target speed velocity (partially known source)
  - □ Noisy random signal
    - Buried object detection (UNKNOWN SOURCE)

PDF = PROBABILITY DENSITY FUNCTION

ESTIMATION

=> capire P(x) dato un sottoinsieme X

X (iid indipendente ed identicamente distribuito) **SAMPLES**

x => features vector

- ESTIMATION
  - PARAMETRIC
    - ML Estim
    - BAYESIAN
  - NON PARAMETRIC
    - KNN
    - PARZEN
    - ETC

PARAMETRIC ESTIMATION

Model to estimate known parameter vector => $\theta = (\theta_1 .. \theta_r)$ che caratterizzano la $p(x)$ del modello

- Likelehood Function

perché i samples sono iid

$$p(X|\theta) = \prod_{k=1}^{N} p(x_k|\theta)$$

dipendenza del modello a θ

defines the likelehood of θ respect to X

- buona misura di compatibilità fra θ e X

ESTIMATION

2 alternative

- Maximum likelehood estimation (ML estim.)
  - maximise the probability of obtaining the samples observed
    optimization process

- Bayesian estimation
  - parametri visti come variabili random che hanno una prior distribution (unknown)
  - observation → posterior density

Estimation error

ipotizzato

$$\varepsilon = \hat{\theta} - \theta = \varepsilon(X, \theta) = [\hat{\theta}_i - \theta_i : i = 1..r]$$

ideal estimate

- no bias

Bias is shift in value constant
Variance

UNCERTAINTY IN THE ESTIMATION

BIAS

MEAN OF ε

$$E\{\varepsilon\} = 0 \Rightarrow E\{\hat{\theta}\} = \theta$$

VALORE ATTESO o valore medio

GENERALMENTE $\mathbb{E}[x] = \sum_{i=0}^{N} x_i P(x_i)$

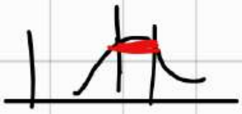
- NO VARIANCE (NO UNCERTAINTY)

HOW MUCH THE ERROR FLUCTUATES
DEVIATES RISPETTO MEAN

$$VAR\{\varepsilon_i\} = E\{(\hat{\theta}_i - \bar{\hat{\theta}}_i)^2\}$$

$(i = 1, 2 \ldots r)$

NO CONFIDENCE



PER STABILIRE LA GOODNESS USO

Cramér-RAO BOUND

COMPUTE THE PEAKNESS

(NO BIAS)

- THE VARIANCE OF SUCH ESTIMATOR IS AT LEAST AS HIGH AS THE INVERSE OF THE FISHER INFORMATION

Fisherman

- rappresenta un limite inferiore sulla varianza di un UNBIASED STATISTICAL ESTIMATOR

Si basa su fisher information

Fisher

è la varianza dello score (derivata logaritmica) associata ad una data la likelihood function

$$VAR\{\varepsilon_i\} \geq [I^{-1}(\theta)]_{ii} \qquad i=1\dots r$$

EFFICIENT ESTIMATOR REACHES THE CRAMER RAO BOUND

- SOSTANZIALMENTE È UN MODO PER MISURARE IL NUMERO DI INFORMAZIONI CHE $\underline{X}$ PORTA DI UNA SCONOSCIUTO PARAMETRO θ

FORMALMENTE LA VARIANZA DELLO SCORE

SCORE

=> IL GRADIENTE (vettore derivate parziali) del logaritmo della likelihood

ASIMPTOTIC PROPRETIES

- ESTIMATOR EVALUATION

- ASYMPTOTICALLY UNBAIASED

if

$$\lim_{N\to+\infty} E\{\varepsilon\}=0 \Longrightarrow \lim_{N\to+\infty} E\{\hat{\theta}\}=\theta$$

- ASYMPTOTICALLY EFFICIENT

if

$$\lim_{N\to+\infty} \frac{var\{\varepsilon\}}{[I^{-1}(\theta)]_{ii}}=1 \qquad i=1,2\dots r$$

fisher

- CONSISTENT

$$\lim_{N \to +\infty} P\{\|\varepsilon\| < \delta\} = 1 \quad \forall \delta > 0$$

CONVERGES TO THE TRUE VALUE WHEN $N \to +\infty$

NECESSARY CONDITION

=> ASYN UNBIASED AND $\text{VAR} \xrightarrow[N \to +\infty]{} 0$

GAUSSIAN

- RAPPRESENTS THE PDF OF A NORMALLY DISTRIBUTED RANDOM VARIABLE

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

EXPECTED VALUE μ

VARIANCE σ^2

- MAXIMUM LIKELEHOOD ESTIMATION

(ML) => $\hat{\theta} = \arg\max_{\theta} p(X|\theta)$ likelehood

=> $\hat{\theta} = \arg\max_{\theta} \ln p(X|\theta)$

OBS

- ML estimates θ that agrees best with training data

- easier to work with log of likeleHood estim

- log increase monotonically => the θ MAX for log is the same for

Propreties

- ASYMPTOTICALLY UNBAIASED
- " EFFICIENT
- CONSISTENT

likelihood

STATISTICAL MODEL SELECTION

ASPECT FOR STATISTICAL MODEL SELECTION

- INTRINSIC STATISTICAL VALUE
- NOISE
- PRE PROCESSING AND FEATURE EXTRACTION

Examples

SUPER USED

GAUSSIAN

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

CENTRAL LIMIT THEOREM

if the sum of many iid variables has a finite variance, then it will be approximately normally distributed

iid INDIPENDENT AND IDENTICALLY DISTRIBUTED

GENERALIZED GAUSSIAN

$$f(x; m, b, \beta) = \frac{d\beta}{2\Gamma(1/\beta} e^{-[b|x-m|]^{\beta}}$$

MULTIVARIATE GAUSSIAN PDF

$$P(x|\theta) = P(x|m, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-m)^t \Sigma^{-1}(x-m)\right]$$

COVARIANCE MATRIX

mean vector

$N(m, \Sigma)$

MEAN VECTOR

$m = E\{x\}$ [MEAN of a large number of indipendently selected OUTCOMES of random variable] (media pesata)

Expected value

$$E[X] = \sum_{i=1}^{\infty} x_i \, P(x_i)$$

COVARIANCE MATRIX

$$\Sigma = Cov\{x\} = E\left\{(x-m)(x-m)^t\right\} = E\left\{x x^t\right\} - m m^t$$

Properties

- SYMMETRIC $\Sigma = \Sigma^t$
- POSITIVE SEMI DEFINITE
- FOR INDIPENDENT FEATURES

SEI UN BEL PATATONE ♡

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix} \Rightarrow P(X) = P(x_1) P(x_2) \dots P(x_n)$$

SHAPE OF 2D GAUSSIAN PDF

- MAIN AXIS OF ECLIPSE = EIGEN VECTOR → MAX EIGEN VALUE

ML ESTIMATOR EXAMPLE

POISSON DISTRIBUTION

$$P(x_i \mid \lambda) = \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}$$

TRAINING SAMPLES

$X = \{x_1 .. x_N\}$

$\hat{X} = ?$

$$P(X \mid \lambda) = \prod_{i=1}^{N} P(x_i \mid \lambda)$$

log likelihood $\Rightarrow \sum_{i=1} \log P(x_i | \lambda)$

$$\Rightarrow \sum_{i=1}^{N} [x_i \log \lambda - \lambda - \ln x_i!]$$

MAXIMIZE

$$\frac{d \log P(X|\lambda)}{\lambda} = \sum_{i=1}^{N} \left[\frac{x_i}{\lambda} - 1\right] = \overset{\text{MAX}}{0}$$

$$\Rightarrow \frac{1}{\lambda} \sum_{i=1}^{N} x_i = N$$

$$\boxed{\hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} x_i}$$

BAYESIAN ESTIMATION

APPLICATION

PARAMETRIC ESTIMATION

θ random variable

Prior $\longrightarrow$ POSTERIOR

EXPERT

PRIOR KNOWLEDGE

The goal is to compute $\overset{\text{POSTERIOR}}{p(x|X)}$ with is close as possible to $p(x)$

- $p(x|X) = \int p(x, \theta | X)\, d\theta$

BAYESIAN THEOREM

- $p(x|X) = \int p(x|\theta, X)\, p(\theta|X)\, d\theta$

$\downarrow$

$$p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta$$

▲ Since θ is unknown Bayesian estimation [illegible]

average $p(x|\theta)$ over all possible θ

Δ $\mathcal{X}$ influence $p(x|\mathcal{X})$ can be posterior $p(\theta|\mathcal{X})$

$$\square \quad p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)\,p(\theta)}{\int p(\mathcal{X}|\theta)\,p(\theta)\,d\theta}$$

BAYESIAN

$$\square \quad p(\mathcal{X}|\theta) = \prod_{k=1}^{N} p(x_k|\theta)$$

Example

- $p(x|\theta)$ GAUSSIAN
- $\theta = [\mu, \sigma^2]$ ONLY UNKNOWN PARAMETER IS μ

$$\Rightarrow \quad p(x|\mu) \sim N(\mu, \sigma^2)$$

$\hookrightarrow \quad p(\mu) \sim N(\mu_0, \sigma_0^2)$ PARAM a priori known

$$\Rightarrow \quad p(\mu|\mathcal{X}) = \alpha \prod_{k}^{N} p(x_k|\mu)\,p(\mu)$$

α NORM FACTOR

ML VS BAYESIAN ESTIM

① - ML estimates a point θ

- Bayesian estimates a distribution $p(x|\mathcal{X})$ then inferes a θ

② TRADE OFFS

ML BETTER

- Lower computational complexity
- EASYER INTERPRETABILITY

NON PARAMETRIC ESTIMATION

- No prior knowledge
- Parametric not good approximation of PDF

→ APPLIED DIRECTLY ON DATA

Basic concepts

let x^* sample, R a region of the feature space such that $x^* \in R$

K n of training samples in R

→ A CONSISTENT ESTIMATE OF P_R

RELATIVE FREQUENCY

$$\hat{P}_R = \frac{K}{N} \qquad \lim_{N \to +\infty} P\{|\hat{P}_R - P_R| < \delta\} = 1 \quad \forall \delta > 0$$

law of big Numbers

ASSUMING true pdf $p(x)$ continuous

R sufficiently small that $p(x)$ does not vary

$$P_R = P\{x \in R\} = \int_R p(x)\,dx = p(x^*)V$$

volume of R

PDF ESTIMATION

$$\hat{p}(x^*) = \frac{\hat{P}_R}{V} = \frac{K}{NV}$$

R DIM:

- ENOUGH NUMBERS TO APPLY LAW OF LARGE NUMBER
- ENOUGH SMALL TO LIMIT THE VARIABILITY OF $p(x)$ WITHIN IT

THE PARAMETERS V AND K DETERMINE 2 DIFFERENT

PARZEN METHOD

$R(\to V)$ is FIXED AND K is calculated

K-NEAREST NEIGHBOR

K is FIXED AND V of R is computed

BOTH CONVERGE ALTHO is DIFFICULT

TO MAKE statements on finite-sample behavior

K-NN

- K is set a priori
- A shape is chosen a priori centered in x^*

    - Expand the cell up to including k training samples

        → Resulting volume $V_K(x^*)$

        → pdf $\hat{p}(x^*) = \dfrac{k}{N V_K(x^*)}$

K can be chosen as a function of N

ex

$k = \sqrt{N}$

PARZEN WINDOW

- TEMPORARY ASSUMPTION R is a n-dimensional hipercube

    ↳ $V = h^n$

    volume — edge — dimensions

- CALCULATE k based on how many training samples lyes in the hyper cube

non ti distrarre anche se sono BELLISSIMA!! ♡

WINDOW FUNCTION $\quad x_K \in R$ if $\gamma[(x_K - x^*)/h] = 1$

FUNCTION THAT GIVES 1 if ∈ TO HYPERCUBE 0 OTHERWISE

CENTER

$x_K \notin R$ if $\gamma[(x_K - x^*)/h] = 0$

- number of samples in hyper cube

$$k = \sum_{k}^{N} \gamma\left(\frac{x_k - x^*}{h}\right)$$

ESTIMATE

↳ $$\hat{p}(x^*) = \frac{\hat{P}_R}{V} = \frac{k}{NV}$$

$$\hat{p}(x^*) = \frac{1}{N} \sum_{k}^{N} \underbrace{\frac{1}{h^n}}_{\text{VOLUME}} \gamma\Bigg(\overbrace{\frac{x_k - x^*}{h}}^{\text{distanza}}\Bigg)$$

$$\frac{x_k - x^*}{h} = \left(\frac{x^* - x_k}{h}\right)$$

IN ESSENCE

- Window function γ é usata per interpolazione ogni sample peso per $\hat{p}(x^*)$ in base alla sua distanza da x

- The function $\gamma(\cdot)$ it's called parzen window or kernel h is the width of the window

DOES THIS SATISFIES TO BE A DENSITY FUNCTION?

- NON NEGATIVE
- INTEGRATE TO 1

this can be assured for the window function to be a density function

$\Rightarrow \quad \gamma(x) \geq 0 \quad \forall x \in \mathbb{R}^n, \quad \int_{\mathbb{R}^n} \gamma(x)\,dx = 1$

if we mantain the condition $V = h^n$

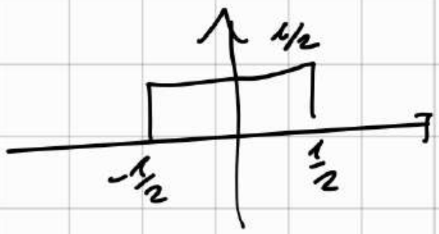
$\Rightarrow$ ✓

GOOD ESTIMATE

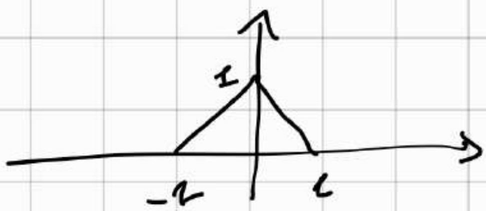
- $\gamma(\cdot)$ MAX VALUE AT THE ORIGIN
- $\gamma(\cdot)$ CONTINUOS
- $\gamma(x) \to 0$ AS $x \to +\infty$

KERNEL EXAMPLES

① Rectangular $\quad \gamma(x) = \Pi(x)$



② TRIANGULAR $\quad \gamma(x) = \Lambda(x)$



③ GAUSSIAN $\quad \gamma(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$

④ EXPONENTIAL

$$\gamma(x) = \frac{1}{2} \exp(-|x|)$$

⑤ CAUCHY

$$\gamma(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$$

⑥ "$\sin c^2(\cdot)$"

$$\gamma(x) = \frac{1}{2\pi} \left(\frac{\sin(x/2)}{x/2} \right)^2$$

Propreties

$$\underline{X} \rightarrow \boxed{\text{PARZEN WINDOW}} \rightarrow \hat{p}(x) \begin{cases} \nearrow \text{BIAS} \\ \searrow \text{VARIANCE} \end{cases}$$

BIAS

- Since the samples are iid

$$\underbrace{E\{\hat{p}(x)\}}_{\text{A BLURRED VERSION BY the kernel}} = p(x) \underbrace{*}_{\text{convolution}} \frac{1}{h^n} \gamma\left(\frac{x}{h}\right)$$

- PARZEN ESTIMATOR is <u>BIASED</u>

BUT ASIMPTOTICALLY UNBIASED

IF KERNEL WIDTH IS CHOSEN PROPERLY

i.e. $h_N \rightarrow 0$ AS $N \rightarrow +\infty$

$$\lim_{N \to +\infty} \left\{ \frac{1}{h_N^n} \gamma\left(\frac{x}{h_N}\right) \right\} = \delta(x)$$

$$\Rightarrow \lim_{N \to +\infty} E\{\hat{p}(x)\} = p(x)$$

Bias

VARIANCE

- ESTIMATED VARIANCE is bounded by

real

$$E\{(\hat{p}(x) - \bar{\hat{p}}(x))^2\} \leq \frac{\sup(\gamma(\cdot))\,\bar{\hat{p}}(x)}{N h_N^n}$$

it can be demostrated to be consistent if

$$\lim_{N \to +\infty} h_N = 0 \qquad \lim_{N \to +\infty} N h_N^n = +\infty$$

example

$$h_N = \frac{1}{\sqrt[2n]{N}}$$

GAUSSIAN KERNEL with spherical symmetry (SPECHT)

$$\hat{p}(x) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|x - x_k\|^2}{2\sigma^2}\right)$$

STANDARD DEVIATION

RAPPRESENTS a smoothing parameter

CHOOSE wisely to prevent OVER AND UNDER FITTING

σ it's the same along all features

NORMALIZATION is NEEDED

- YOU MAY compute σ adaptively to each sample

- σ_K could be defined MEAN EUCLIDEAN DISTANCE X_K and L nearest training samples $y_1, y_2 \dots y_L$

$$\sigma_K = \frac{1}{L} \sum_{l=1}^{L} \| x_K - y_l \|$$

THIS method can be implemented by means of Probabilistic NN

ESTIMATION WITH INCOMPLETE DATA

- N iid observations $x_i = (i = 1, 2, 3 \dots N)$ in n-dimensional space $\mathbb{X}$

- ASSUME we know the distribution and its MIX GAUSSIAN

$$p(x) = \sum_{i=1}^{M} \underbrace{P_i}_{\text{Prior Probability (WEIGHT)}} \, p(x \mid \underbrace{m_i}_{\text{MEAN}}, \underbrace{\Sigma_i}_{\text{SHAPE}})$$

are the mean and covariance MATRIX

let $\theta = [P_1, P_2 \ldots P_M; m_1, m_2 \ldots m_M; \Sigma_1, \Sigma_2, \ldots \Sigma_M]$

vector of param to be estimated

EM ALGORITHM

X DATASET

$Z = (X, Y)$ complete dataset

Y MISSING DATA

- JOINT density function

$$p(Z|\theta) = p(X, Y|\theta) = p(Y|X, \theta) \cdot p(X|\theta)$$

COMPLETE DATA likelihood

$$l(\theta|Z) = p(Z|\theta) = p(X, Y|\theta)$$

incomplete DATA LIKELIHOOD

$$l(\theta|X) = p(X|\theta)$$

EM

find the expected value of $\log p(X, Y|\theta)$

w.r.t Y

given X and a current estimate of θ

CONDITIONAL EXPECTATION

$$Q(\theta, \theta^{(k)}) = E\{\log(p(X, Y|\theta)) | X, \theta^{(k)}\}$$

$$Q(\theta, \theta^{(k)}) = E\{\log(p(X, Y \mid \theta)) \mid X, \theta^{(k)}\}$$

$\Downarrow$ Expectation step

$$\int \log(p(X, y \mid \theta))\, p(y \mid X, \theta^{(k)})\, dy$$

$\Rightarrow$ EM alterna fra expectation E della likelihood con latent variables come se fossero osservate.
MASSIMIZZA M ML ESTIMATO dello step precedente
: parametri trovati si usano in un altro STEP(E)

Expectation step $\quad Q(\theta, \theta^{(k)}) = E\{\log(p(X, Y \mid \theta)) \mid X, \theta^{(k)}\}$

Maximization step $\quad \theta^{(k+1)}$

Feature reduction

- COURSE OF DIMENSIONALITY OR HUGES EFFECT
  - ADDING new feature leads to worse rather than better performances

    CAREFULL
    - How is the classification accuracy affected by dimensionality
    - Save for computational complexity

OBJectives of feature reduction

- Minimize implementation cost
- reduce computational load
- overcome HUGES EFFECT

HUGES EFFECT

- UNBALANCE Between number of training samples and features
- BALANCE depends on classifier complexity

  ↳ UNRELIABLE

- is caused by exponential increase in volume

⚠ the feature space can be represented by an hypersphere of radius r in n dimensions

for convergence you usually set a $\lim_{n \to +\infty}$

but higher the dimensionality the more samples you need to converge

GEOMETRICAL ANALYSIS

→ High dimensional space are MOSTLY EMPTY

IMPLIES multivariate data in $\mathbb{R}^n$ is in a lower dimensional STRUCTURE

→ NORMALLY DISTRIBUTED DATA concentrate in the tails

uniformly distributed in the corners

△ Filter method

- Filter features indipendently from the classifier
- Wrapper methods where best feature are guided by acc. of classifier
- embedded feature selection is part of the classifier

SEPARABILITY MEASURE

↳ SHOULD BE COMPATIBLE with the discrimination criterion of the classifier : ex

① ACCURACY of the classifier

② Measure based on error probability of bayes classifier

TYPICAL SEP. MEASURES

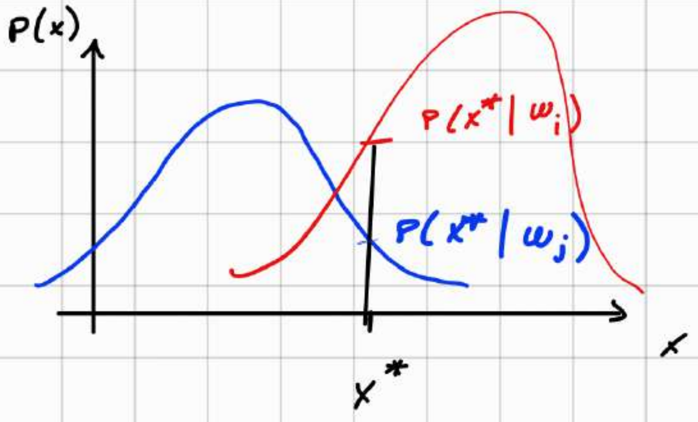
- DIVERGENCE measure
- BHATTACHARYYA
- JEFFRIES - MATUSITA

★ $E\{f(x)\} = \int f(x)\, p(x)\, dx$

DIVERGENCE MEASURE

CONSISTE nell' overlap degree

- Likelihood ratio



$$L_{ij}(x) = \frac{P(x | w_i)}{P(x | w_j)}$$

$\simeq 1$ STRONG OVERLAP

$\gg$ OR $\ll$ small overlap

NOT SYMMETRIC

Divergence measure

log ratio

$x \in F'$

$$D_{ij}(F') = \underbrace{E}_{\text{average}} \{ L'_{ij}(x) \} + E \{ L'_{ji}(x) \}$$

where

$$L'_{ij}(x) = \ln [L_{ij}(x)] = \ln [P(x | w_i)] - \ln [P(x | w_j)]$$

$$\Rightarrow \quad D_{ij}(F') = \int_x \left\{ [P(x | w_i) - P(x | w_j)] \ln \left[\frac{P(x | w_i)}{P(x | w_j)} \right] \right\} dx$$

SPECIAL CASE GAUSSIAN

$$D_{ij}(F') = \frac{1}{2} Tr \{ (\Sigma_i - \Sigma_j) \cdot (\Sigma_i^{-1} - \Sigma_j^{-1}) \} + \frac{1}{2} Tr \{ (\Sigma_i^{-1} - \Sigma_j^{-1}) \cdot (m_i - m_j) \cdot (m_i - m_j)^t \}$$

COVARIANCE

MEAN VECTOR

MATRIX TRACE OPERATOR

SOMMA VALORI DIAGONALE

Δ Proprietà

- $\omega_1 = \omega_j \Rightarrow D_{ij} = 0$

- $\omega_1 \neq \omega_j \Rightarrow D_{ij} > 0$

- $D_{ij} = D_{ji}$

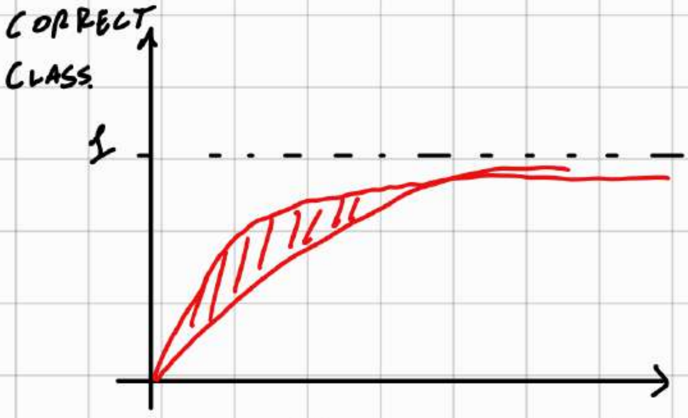
- $D_{ij}(f_1 \dots f_K) \leq D_{ij}(f_1 \dots f_K, f_{K+1})$

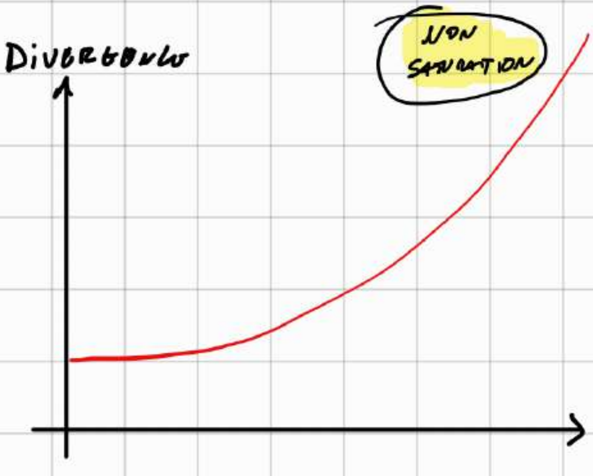
- if f INDIPENDENT FEATURES

$$D_{ij}(f_1 \dots f_K) = \sum_{q=1}^{K} D_{ij}(f_q)$$

LARGER THE DIVERGENCE BETTER SEPARABILITY BETWEEN CLASSES

DRAWBACK





IN MULTI CLASS

AVERAGE corresponding to all couples of classes

$$D_{ij}(F') = \sum_{i=1}^{c} \sum_{j>i}^{c} P(\omega_i) \cdot P(\omega_j) \cdot D_{ij}(F')$$

MATRIX WITH Distance Between classes

WORST CASE REASONING

Cov $\frac{c(c-1)}{2}$ entry

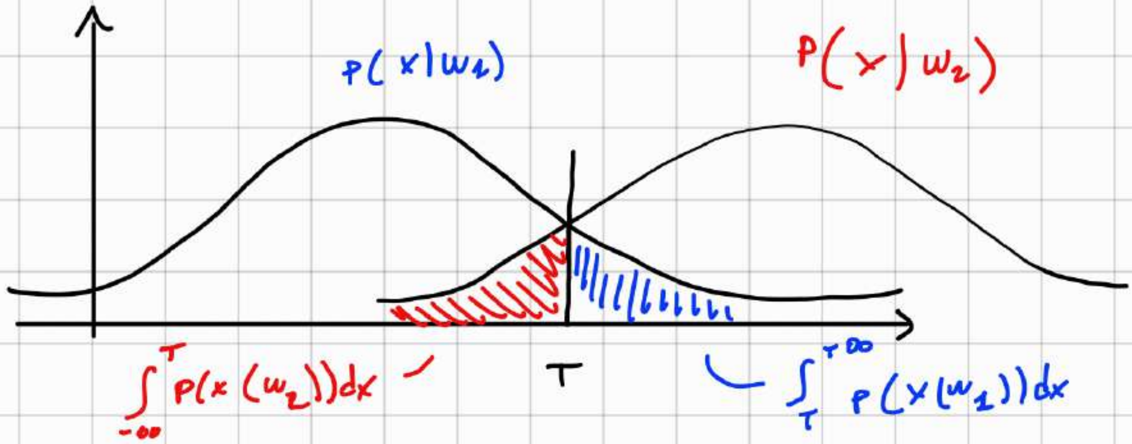
$$D_{min}(F') = \min_{\substack{1 \le i < c \\ i < j}} \{ D_{ij}(F') \}$$

BHATTACHARYA DISTANCE

- is important to set the distance based on the classifier so the bayesian classifier needs are

- upper bound on error probability

- error in bayesian classifier



Errore bayesian classifier

$$P_{err} = P(x \to w_2 | w_1)\, P(w_1) + P(x \to w_1 | w_2)\, P(w_2)$$

$\downarrow$ integrale di PDF e densità

$$\int_{-\infty}^{T} P(x|w_2)\,dx + \int_{T}^{+\infty} P(x|w_1)\,dx$$

devo minimizzare

$$\int_{-\infty}^{+\infty} \min \left\{ p(x|\omega_1) P(\omega_1) , P(x|\omega_2) P(\omega_2) \right\} dx$$

↳ HARD TO COMPUTE $\Rightarrow$ APPROXIMATION

USING THE RELATIONSHIP $\min[a,b] \leq a^s b^{1-s} \quad 0 \leq s \leq 1$

TO EXPLICIT THE MINIMUM

CHERNOFF BOUND

$$\varepsilon_u = P(\omega_i)^s P(\omega_j)^{1-s} \int_x p(x|\omega_i)^s \cdot p(x|\omega_j)^{1-s} dx$$

$$= P(\omega_i)^s P(\omega_j)^{1-s} \exp\left[-\underbrace{\mu_{ij}(s)}_{\text{CHERNOFF DISTANCE}}\right]$$

SMALLER => BETTER SEPARABILITY

→ for a gaussian distribution

$$\mu_{ij}(s) = \frac{s(1-s)}{2} (m_i - m_j)^t \left\{ s\Sigma_i + (1-s)\Sigma_j \right\}^{-1} (m_i - m_j) + \frac{1}{2} \ln \left\{ \frac{|s\Sigma_i + (1-s)\Sigma_j|}{|\Sigma_i|^s |\Sigma_j|^{1-s}} \right\}$$

BUT s ?

① s calculatex maximising $\mu_{ij}(s)$

② ARBITRARY value

③ Particular case $\boxed{s = \frac{1}{2}}$

$\Downarrow$

Bhattasharyya bound

$$\varepsilon_U = \sqrt{P(\omega_i)\, P(\omega_j)} \int_X \sqrt{p(x|\omega_i) \cdot p(x|\omega_j)}\, dx$$

$$= \sqrt{P(\omega_i)\, P(\omega_j)}\, \exp\left[-\mu\left(\frac{1}{2}\right)\right]$$

BHATTASHARYYA DISTANCE GAUSSIAN

$$B_{ij} = \mu_{ij}\left(\frac{1}{2}\right) = \frac{1}{8}(m_i - m_j)^t \left\{\frac{\Sigma_i + \Sigma_j}{2}\right\}^{-1}(m_i - m_j) + \frac{1}{2}\ln\left\{\frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{|\Sigma_i|^{1/2}|\Sigma_j|^{1/2}}\right\}$$

Propreties

- $\omega_i = \omega_j \Rightarrow B_{ij} = 0$
- $\omega_i \neq \omega_j \Rightarrow B_{ij} > 0$
- $B_{ij} = B_{ji}$
- $B_{ij}(f_1 \dots f_k) \leq B_{ij}(f_1 \dots f_k, f_{k+1})$

- indipendent features

$$B_{ij}(f_1 \dots f_k) = \sum_{q}^{k} B_{ij}(f_q)$$

- Non SATURATING behaviour

Jeffries - Matusita Distance

- OBJECTIVE

SATURATING BEHAVIOR

IS AN AVERAGE DISTANCE BETWEEN 2 DENSITY FUNCTION

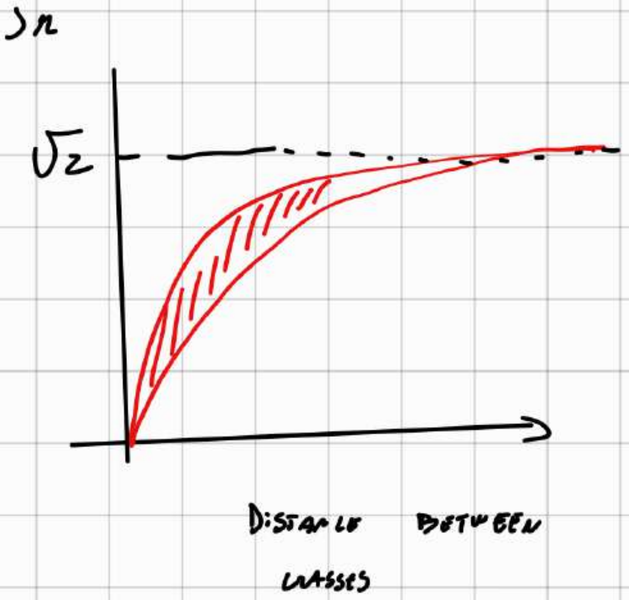
$$JM_{ij} = \left\{ \int_x \left[\sqrt{p(x|\omega_i)} - \sqrt{p(x|\omega_j)} \right]^2 dx \right\}^{1/2}$$

si può esprimere come una funzione della BHATTACHARYYA

$$JM_{ij} = \sqrt{2(1 - \exp(-B_{ij}))}$$

if $B = 0 \Rightarrow JM = 0$

$B \to +\infty \Rightarrow JM \to \sqrt{2}$



SEARCH STRATEGIES

DATA → FEATURES

DISTANCE

- EXAUSTIVE SEARCH

complexity $\binom{n}{m}$

SUB OPTIMAL SEARCH STRATEGIES

SEQUENTIAL FORWARD SELECTION

- Find feature that optimises singularly the class sep. measure

- INCLUDE FEATURE IN DESIRED SUBSET

- ITERATIVELY CHOOSE OTHER FEATURES TO INCLUDE ALONG THE OTHERS

- iterated m times (DESIRED number of features)

SFS SUBOPTIMAL

- Nesting effect a features previously selected cannot be removed

ex $m = 3$ → DESired features

$n = 4$ → initial features

BEST

1 2 3 4

1 2   2 3 BEST   2 4

1 2 3   2 3 4   Final Solution

SEQUENTIAL BACKWARD SELECTION

OPPOSITE FROM SFS AT each iteration remove the lowest decreased of adopted class until cardinality m

TRADE OFFS

- VERY FAST
- same n lower m
  Better SFS

HIGH m Better SBF

FEATURE EXTRACTION

COMBINING FEATURES potentially loose less information than feature selection

PCA	**LDA**
Principal component analysis	Linear discriminant analysis
SEEK PROJECTION that best rappresents the DATA in least	SEEKS BEST DATA SEPARATION

SQUARE SENSE

PRINCIPAL COMPONENT ANALYSIS

Given $x_1, x_2 \dots x_N \in R^n$ the goal is to find the m-dimensional sub space where the reconstruction error is minimized

ORTHONORMAL TO EACH OTHER

$$\begin{cases} \langle \phi_i \cdot \phi_k \rangle = 0 \\ \| \phi_k \| = 1 \end{cases}$$

ORTHONORMAL

$$\hat{x}_i = \sum_{k}^{m} y_{k_i} \phi_k$$

$$J_m = \sum_{i=1}^{N} \left\| \sum_{k=1}^{m} y_{ik} \phi_k - x_i \right\|^2$$

distance between points

real feature

projected point

BASE of subspace

This is effectively minimized when $\phi_1 \dots \phi_m$ are the m eigenvectors of the scatter matrix HAVING THE LARGEST eigenvalues

$$S = N \cdot \Sigma$$

$$S = \sum_{i}^{N} (x_i - m)(x_i - m)^t$$

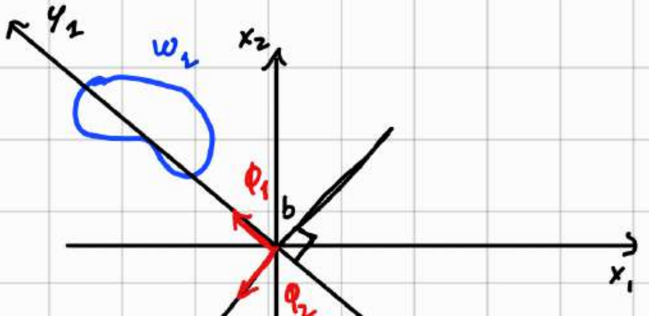
THE coefficient $y = (y_1 \dots y_m)^t$ are called principal components

S is n-1 times the Σ

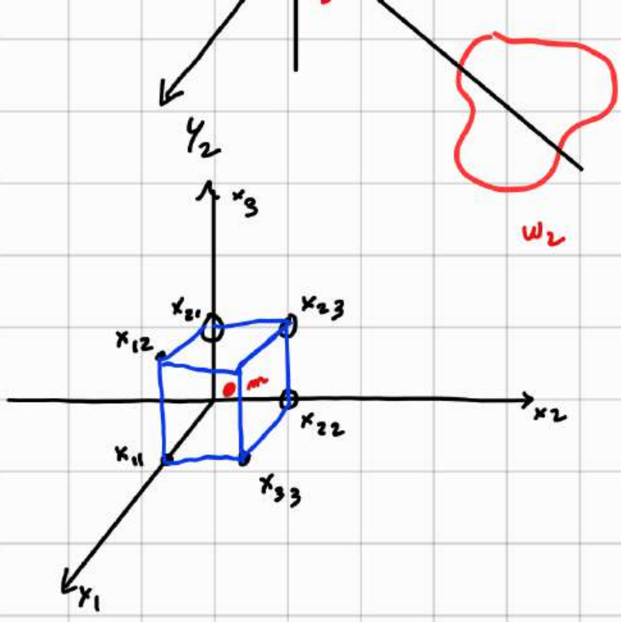
$\Rightarrow \phi_k$ is eigenvector of $S(\Sigma)$

Δ WHEN THE EIGENVECTORS ARE SORTED IN DESCENDING ORDER OF CORRESPONDING EIGENVALUES THE GREATEST VARIANCE OF THE DATA LIES ON THE FIRST PRINCIPAL COMPONENT

ex



w_1	w_2
$x_{11} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$	$x_{21} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$
$x_{12} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$	$x_{22} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$



$$x_{13} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad x_{23} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

$n = 3$ $\quad m = 2$ ①

$m = 1$ ②

①

1. Barycenter

$$m = \begin{bmatrix} 0,5 \\ 0,5 \\ 0,5 \end{bmatrix}$$

2. DATA SHIFTING $\quad x' = x - m$

$$x'_{11} = \begin{bmatrix} 0,5 \\ -0,5 \\ -0,5 \end{bmatrix} \qquad x'_{21} = \begin{bmatrix} -0,5 \\ -0,5 \\ 0,5 \end{bmatrix}$$

$$x'_{12} = \begin{bmatrix} 0,5 \\ -0,5 \\ 0,5 \end{bmatrix} \qquad x'_{22} = \begin{bmatrix} -0,5 \\ 0,5 \\ -0,5 \end{bmatrix}$$

$$x'_{13} = \begin{bmatrix} 0,5 \\ 0,5 \\ -0,5 \end{bmatrix} \qquad x'_{23} = \begin{bmatrix} -0,5 \\ 0,5 \\ 0,5 \end{bmatrix}$$

3. COVARIANCE MATRIX

$$\Sigma_{x'} = \frac{1}{N} \sum_{i=1}^{k} (x_i - \bar{x})(x_i - \bar{x})^t$$

$$\begin{bmatrix} 0,5 \\ -0,5 \\ -0,5 \end{bmatrix} \cdot \begin{bmatrix} 0,5, & -0,5, & -0,5 \end{bmatrix}$$

$$\begin{bmatrix} 0,25 & -0,25 & -0,25 \\ -0,25 & 0,25 & 0,25 \\ -0,25 & 0,25 & 0,25 \end{bmatrix} + \begin{bmatrix} 0,25 & -0,25 & 0,25 \\ -0,25 & 0,25 & -0,25 \\ 0,25 & -0,25 & 0,25 \end{bmatrix} \ldots$$

$$\frac{1}{} \frac{1}{} \begin{bmatrix} 3 & -1 & -1 \end{bmatrix}$$

$$\Sigma_{x'} = \frac{1}{6} \cdot \frac{1}{2} \begin{bmatrix} -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$$

$$\det \begin{bmatrix} 3-\lambda & -1 & -1 \\ -1 & 3-\lambda & -1 \\ -1 & -1 & 3-\lambda \end{bmatrix} = 0$$

$$(3-\lambda) \det \begin{bmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{bmatrix} + 1 \det \begin{bmatrix} -1 & -1 \\ -1 & 3-\lambda \end{bmatrix} - 1 \det \begin{bmatrix} -1 & 3-\lambda \\ -1 & -1 \end{bmatrix}$$

$$(3-\lambda)\left[(3-\lambda)^2 - 1\right] + (-3+\lambda-1) - (1+3-\lambda) = 0$$

$$(3-\lambda)(9-6\lambda+\lambda^2-1) - 3 + \lambda - 1 - 1 - 3 + \lambda = 0$$

$$27 \underline{- 18\lambda} + 3\lambda^2 - 3 \underline{- 9\lambda} + 6\lambda^2 - \lambda^3 + \underline{\lambda} - 3 + \lambda - 5 + \underline{\lambda} = 0$$

$$-\lambda^3 + 9\lambda^2 - 24\lambda + 16 = 0$$

$$-(\lambda-1)\cdot(\lambda^2-8\lambda+16) = (-\lambda+1)\cdot(\lambda-4)^2 = 0$$

$\lambda_1 = \frac{1}{3}$

$\lambda_2 = \frac{1}{3}$ eigenvector

$\lambda_3 = \frac{1}{12}$

$$Ax = \lambda x$$

$$\begin{bmatrix} 3-\lambda & -1 & -1 \\ -1 & 3-\lambda & -1 \end{bmatrix} \quad \begin{bmatrix} \frac{8}{3} & -1 & -1 \\ -1 & 8 & -1 \end{bmatrix} \bar{X} = 0$$

$$\begin{bmatrix} \cdots \\ -1 & -1 & 3-\lambda \end{bmatrix} \qquad \begin{bmatrix} \cdots \\ -1 & -1 & \frac{8}{3} \end{bmatrix}$$

B

$\lambda = \frac{1}{3}$

$$B \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \left[\begin{array}{ccc|c} \frac{8}{3} & -1 & -1 & 0 \\ -1 & \frac{8}{3} & -1 & 0 \\ -1 & -1 & \frac{8}{3} & 0 \end{array}\right]$$

$$\frac{3}{8} R_1 + R_2$$

$$\left[\begin{array}{ccc|c} \frac{8}{3} & -1 & -1 & 0 \\ 0 & \frac{8}{3} - \frac{3}{8} & -1 - \frac{3}{8} & 0 \\ 0 & -1 - \frac{3}{8} & \frac{8}{3} - \frac{3}{8} & 0 \end{array}\right] \quad \ldots\ldots$$

Linear discriminant ANALYSIS

Seeks directions efficient for discrimination

$y = w^t x$ — line to project

where points from ω_1 and ω_2 are well separated

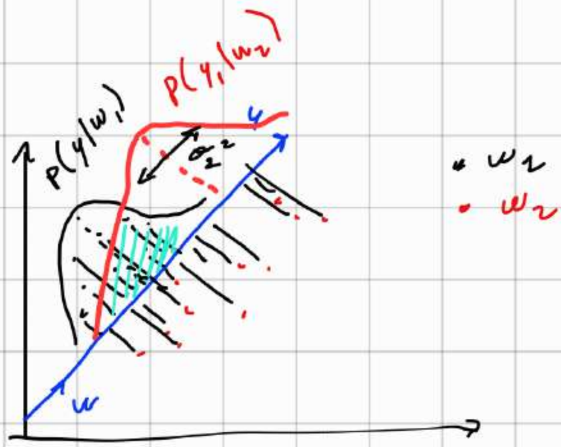
- CRITERION FUNCTION

SAMPLE MEAN

where $\tilde{m}_i = \frac{1}{\#(X_i)} \sum_{y \in \omega_i} y$

SCATTER OF PROJECTED LABLES

and $\tilde{S}_i^2 = \sum_{y \in \omega_i} (y - \tilde{m}_i)^2$



① Between class distance

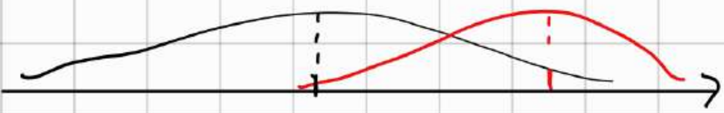
$$(\tilde{m}_1 - m_2)^2$$

MAX

② within class measure

$$\bar{\sigma}_1^2 + \sigma_2^2$$

MIN



$\tilde{S}_1 = N_1 \bar{\sigma}_1^2$

$\bar{S}_2^2 = N_2 \bar{\sigma}_2^2$

FEATURE MEASURE

MAXimize $J(w) = \frac{\| \tilde{m}_1 - \tilde{m}_2 \|^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$

?

- within class scatter matrix

$$S_W = S_1 + S_2$$

$$S = N \cdot \Sigma$$

- Between class S

$$S_B = (m_1 - m_2)(m_1 - m_2)^t \qquad J(w) = \frac{w^t S_B w}{w^t S_w w}$$

$$w = S_w^{-1}(m_1 - m_2)$$

C-CLASSES

- $$S_W = \sum_{i=1}^{c} S_i$$

$$S_B = \sum_{i=1}^{c} [\#(X_i)](m_i - m)(m_i - m)^t$$

GLOBAL MEAN VECTOR

$$m = \frac{1}{N} \sum_{x \in X} x$$

MATRIX (c-1) (volume of means)

$$\Rightarrow J(W) = \frac{|W^t S_B W|}{|W^t S_W W|}$$

MAXIMIZATION

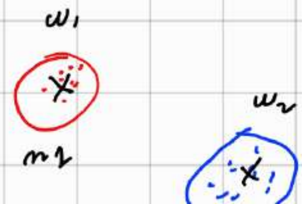
DETERMINANT

W = eigenvector of $S_W^{-1} S_B$

$$\boxed{m \text{ eigenvectors} \leq c-1}$$

SUPERVISED CLASSIFICATION

MDM Minimal distance to mean classifier

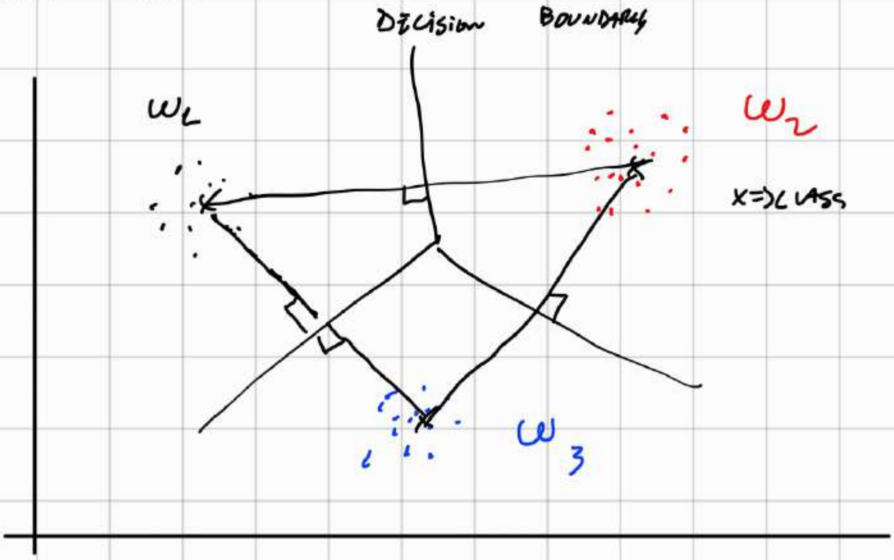


model on classes union

- SMALL SAMPLE DISPERSION

- SAME STATISTICAL BEHAVIOUR

- CLASSIFICATION with euclidean distance to the mean vector



BOX CLASSIFIER

- MDM Does not take into account 2° ORDER STATISTICS (VARIANCE

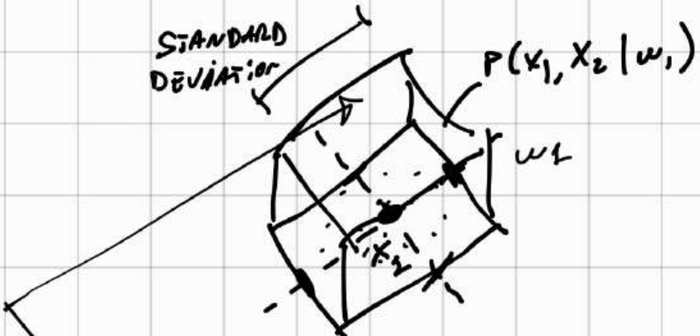
• MODELS THE CLASS in UNIFORM PDF
• CENTER ON THE BARICENTER
• with a SIZE PROPORTIONAL TO STDV in each dimension of feature space

-> CREATE BOXES AS DECISION REGIONS

CLASSIFIES INSIDE THE BOXES

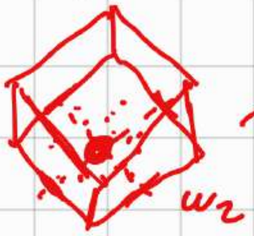
side $= \alpha \cdot \sigma_{11}$

CONSTANT



$P(x_1, x_2 | \omega_2)$ uniform

$P(x_1, x_2 | \omega_2)$ "



$\rightsquigarrow P(x_1 x_2 \mid w_2)$

HOW ABOUT AN x THAT is in NO BOX?

① CLOSEST TO BOX

K-NN CLASSIFIER

Derives from the density estimation KNN

- K is hyper parameter
- calculate the distance between new point to other samples and the majority of the samples IS THE DOMINANT CLASS

High K very non linear OVERFITTING

BAYESIAN CLASSIFICATION

CRITERION

FULL PROBLEM — MINIMUM RISK

INCOMPLETE PROBLEM — MINMAX, NEYMAN PEARSON

MAP CRITERION   ML CRITERION

MAP CRITERION

- Hipotesis — Posterior prob $P(w_i|x)$ are assumed known
- GOAL: Minimize the average prob of error
- DECISION rule

  A pattern x is assigned to the class that maximise the posterior

$$x \in w_j \iff P(w_j|x) \geq P(w_i|x) \ \forall i \in C$$

-> since the posterior are not known use the Bayes theorem

$$\text{post} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

MAX POSTERIOR

$$\Rightarrow \quad x \in w_j \iff P(w_j)\,p(x|w_j) \geq P(w_i)\,p(x|w_i) \quad \forall i \in C$$

BAYES $$P(w_i|x) = \frac{P(x|w_i)\,P(w_i)}{P(x)}$$

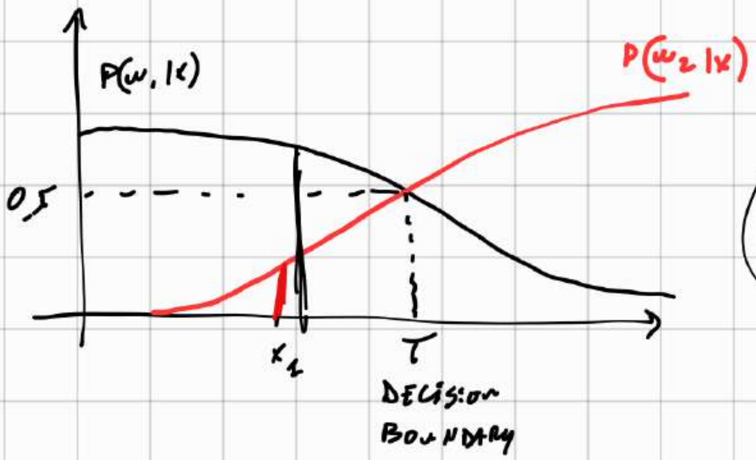
$$\hat{w} = \arg\max_{w_i} \{P(w_i|x)\}$$

- ERROR

$$P(\text{error}) = P(\text{err}|w_1)\,P(w_1) + P(\text{err}|w_2)\,P(w_2)$$

$$= \int_{-\infty}^{T} P(\omega_2 | x)\, P(x)\, dx + \int_{T}^{+\infty} P(\omega_1 | x) \cdot P(x)$$

$$\text{MIN} \quad P(err) = \int_{-\infty}^{+\infty} \min \{ P(\omega_1 | x), P(\omega_2 | x) \}\, P(x)\, dx$$



MIN

if $x_1 \rightarrow \omega_1$

$P(err | x_1) = P(\omega_2 | x_1)$

if $x_1 \rightarrow \omega_2$

$P(err | x_1) = P(\omega_1 | x_1)$

AVERAGE

$$P_e = \sum_{1}^{C} P(err | \omega_i)\, P(\omega_i)$$

ML CRITERION

- HYP. ASSUMED CLASS CONDITIONAL PDF'S

- GOAL MIN AVERAGE PROB of error

- DECISION RULE

$$x \in \omega_j \iff P(x | \omega_j) \geq P(x | \omega_i) \quad \forall_i \in C$$

- CLASS PROB is $\frac{1}{C}$

if Prior is equal $\Rightarrow$ ML = MAP

$$w^2 = \underset{w_i}{\arg\max} \{ P(x|w_i) \}$$

RISK THEORY

EX

$$\Lambda = \begin{bmatrix} \lambda(\alpha_1|w_1) & \lambda(\alpha_1|w_2) \\ \lambda(\alpha_2|w_1) & \lambda(\alpha_2|w_2) \end{bmatrix}$$

DECISION

$$R(\alpha_1|x) \gtrless R(\alpha_2|x)$$

$$\Leftrightarrow \quad \lambda_{11} P(w_1|x) + \lambda_{12} P(w_2|x) \gtrless \lambda_{21} P(w_1|x) + \lambda_{22} P(w_2|x)$$

BAYES

$$\lambda_{11} \frac{P(x|w_1) P(w_1)}{\cancel{P(x)}} + \lambda_{12} \frac{P(x|w_2) P(w_2)}{\cancel{P(x)}} \gtrless \lambda_{21} \frac{P(x|w_1) P(w_1)}{\cancel{P(x)}} + \lambda_{22} \frac{P(x|w_2) P(w_2)}{\cancel{P(x)}}$$

$$P(x|w_1) \cdot P(w_1) [\lambda_{11} - \lambda_{21}] \gtrless P(x|w_2) \cdot P(w_2) [\lambda_{22} - \lambda_{12}]$$

$$\frac{P(x|w_1)}{P(x|w_2)} \gtrless \frac{P(w_2)}{P(w_1)} \frac{[\lambda_{22} - \lambda_{12}]}{[\lambda_{11} - \lambda_{21}]}$$

$\parallel$

L ratio

$\downarrow$

η

DISCRIMINANT FUNCTIONS

$g(x)$ $(i = 1, 2 \dots c)$ where the classifier assigns a feature vector x to class w_j if

$$g_i(x) > g_j(x) \quad \forall j \neq i$$

these function separates the space into c decision regions with decision boundaries

For risk min

$$g_i(x) = -R(\alpha_i | x)$$

MIN ERROR

$$g_i(x) = P(\omega_i | x)$$

CASE 1 $\Sigma_i = \sigma^2 I$

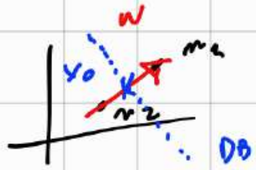
- LINEAR DISCRIMINANT FUNCTION

$$g_i(x) = w^t_i x + w_{i0}$$

Threshold/bias

where

$$\begin{cases} w_i = \frac{1}{\sigma^2} m_i \\ w_{i0} = \frac{1}{2\sigma^2} m^t_i m_i + \ln p(\omega_i) \end{cases}$$



DECISION BOUNDARIES

$$g_i(x) = g_j(x)$$

$$\boxed{w^t(x - x_0) = 0}$$

PERPENDICULAR

where

$$\begin{cases} w = m_i - m_j \\ x_0 = \frac{1}{2}(m_i + m_j) - \underbrace{\frac{\sigma^2}{\|m_i - m_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (m_i - m_j)}_{\text{correction term}} \end{cases}$$

mid point

BASED ON PRIOR

CASE 2 $\Sigma_i = \Sigma$ SAME SHAPE

$$g_i(x) = w^t_i x + w_{i0}$$

rotated on the shape of the covariance

$$\begin{cases} w_i = \Sigma^{-1}(m_i - m_j) \\ w_{i0} = -\frac{1}{2} m_i^t \Sigma^{-1} m_i + \ln P(\omega_i) \end{cases}$$

CASE ③ Σ_i = arbitrary

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

QUADRATIC

$$\begin{cases} W_i = -\frac{1}{2} \Sigma_i^{-1} \\ w_i = \Sigma_i^{-1} m_i \\ w_{i0} = -\frac{1}{2} m_i^t \Sigma^{-1} \ldots \end{cases}$$

DECISION TREE

- INTERPRETABILITY
- FAST CLASSIFIER
- EASY INCORPORATION OF PRIOR KNOWLEDGE

CART

CLASSIFICATION REGRESSION TREES

- SPLIT

DECISION DIVISION

* SIMPLICITY

WE PREFERR DECISION THAT LEADS TO SMALLER TREE

IMPURITY

Gini $\quad i(N) = \sum_{i \neq j} P(\omega_i) P(\omega_j)$

Drop in impurity

$$\Delta i(N) = i(N) - P_L\, i(N_L) - (1 - P_L)\, i(N_R)$$

left $\quad$ right

Max $\Delta i(N)$

LOCAL MINIMA

TOWING CRITERION

SUPER CLASS ASSUMED TO USE DROP IMPURITY with all combination of classes

SEARCH in 3 DIMENSION
- BEST CLASS DIVISION
- BEST PARAMETER
- BEST VALUE

STOPPED SPLITTING

EARLY STOP TO AVOID OVERFITTING

STOPPING CRITERION

- CROSS VALIDATION USE DATASET TO TEST
- DROP OF IMPURITY THRESHOLD ON $\Delta i(N)$

TRADE OFF TOTAL IMPURITY AND

$$\sum i(N) + \alpha \text{ Complexity}$$

Hypothesis testing

Pruning

grow as much as the tree an then pruning

RISK UNBALANCED TREE

DT

VERY UNSTABLE

MULTI VARIATE DECISION TREE

DT HAVE ONLY PARALLEL LINES

TO MAKE NOT PARALLEL ⇒

USE NON LINEAR DISCRIMINANT BOUNDARY

⇒ USE MORE FEATURE IN A SINGLE SPLIT

RANDOM FOREST

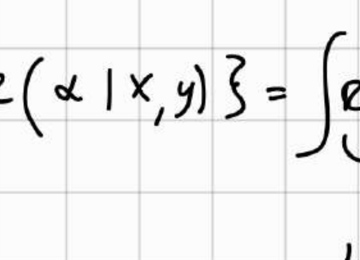
+ TREE with split of dataset

DECISION MAJORITY RULE

GENERALIZATION ERROR

CLASSIFIER WORK

GEN ERROR = APPROXIMATION ERROR + ESTIMATION ERROR



NOT HAVE true joint PDF

$$R(\alpha) = E\{R(\alpha \mid x, y)\} = \int \underbrace{R(\alpha(x,y))}_{L(y,\hat{y}) = L(y, f(x,\alpha))} P(x,y)\, dx\, dy$$

NEED APPROXIMATION

EMPIRICAL RISK

$$R(\alpha) \simeq \frac{1}{N} \sum_{i}^{l} L(y_i, f(x_i, \alpha))$$

$N \to \infty \qquad R_{emp}(\alpha) \to R(\alpha)$

ACCURACY EVALUATION

USED TO

- TUNE HYPER PARAMETERS
- EFFECTIVENESS
- MAKE STATISTICAL COMPARISONS BETWEEN MACHINES

- CROSS VALIDATION ROUND VALIDATION

  - EXAUSTIVE (LEAVE ONE OUT)

ex 100 samples

i consider 99 LABLED AND TRAIN 1 TEST

- ITERATION OVER ALL SAMPLES

  TRAIN OVER 99 TEST 1

  - TRAIN 100 MACHINES

MISCLASSIFIED → ACC

- LEAVE P OUT

  SAME BUT WITH P instead of 1

- Hold OUT

- SET LABELED SAMPLES ; DIVIDE 50%

TRAIN AND TEST

1 TRAINING

MOST USED

- K FOLD — Semplificazione of leave 1 out

DIVIDE in K part (5)

TRAIN OVER A FOLD OUT

ex LEAVE 20 OUT

TRAIN 5 TIMES INSTEAD OF 100

- MONTECARLO — MANY HOLD OUT

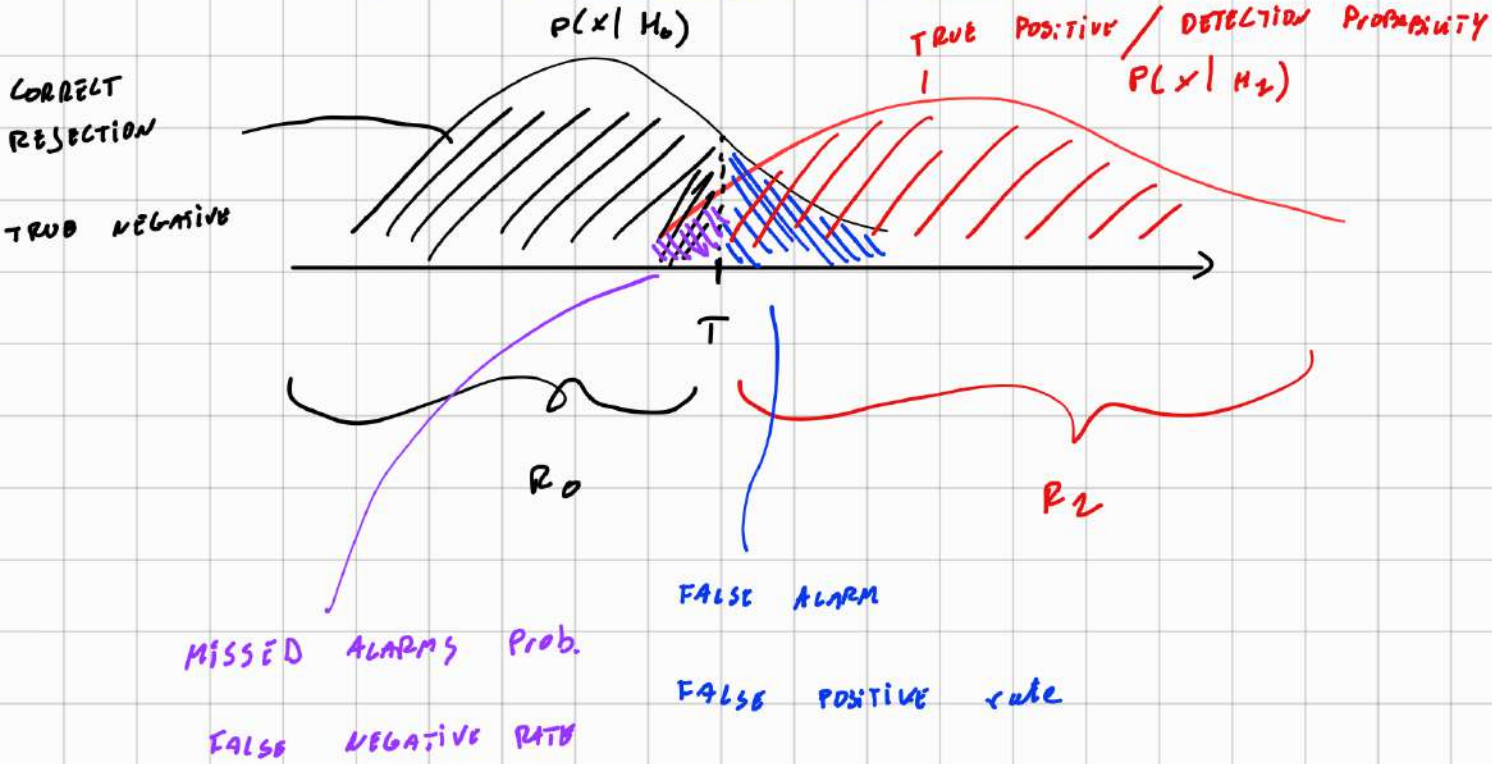
- RANDOM SPLIT
- ITERATE ON DIFFERENT SPLIT

- NESTED CROSS VALIDATION

└ WHEN YOU NEED ACC AND HYPER PARAM

- SPLIT K FOLDS
  - SPLIT L FOLDS
  - BEST HYPER PARAMETER

ACC. IN BINARY CLASSIFICATION



CONFUSION MATRIX

TRUE / DECISION

	D_0	D_1
H_0	TN	FP
H_1	FN	TP

BEST SCENARIO

$\omega_1 \to 100$

$\omega_2 \to 100$

	$\hat{\omega}_1$	$\hat{\omega}_2$
ω_1	100	0
ω_2	0	100

$$ACC = \frac{TN + TP}{TN + TP + FN + FP}$$

$$\text{PRECISION} = \frac{TP}{TP + FP}$$

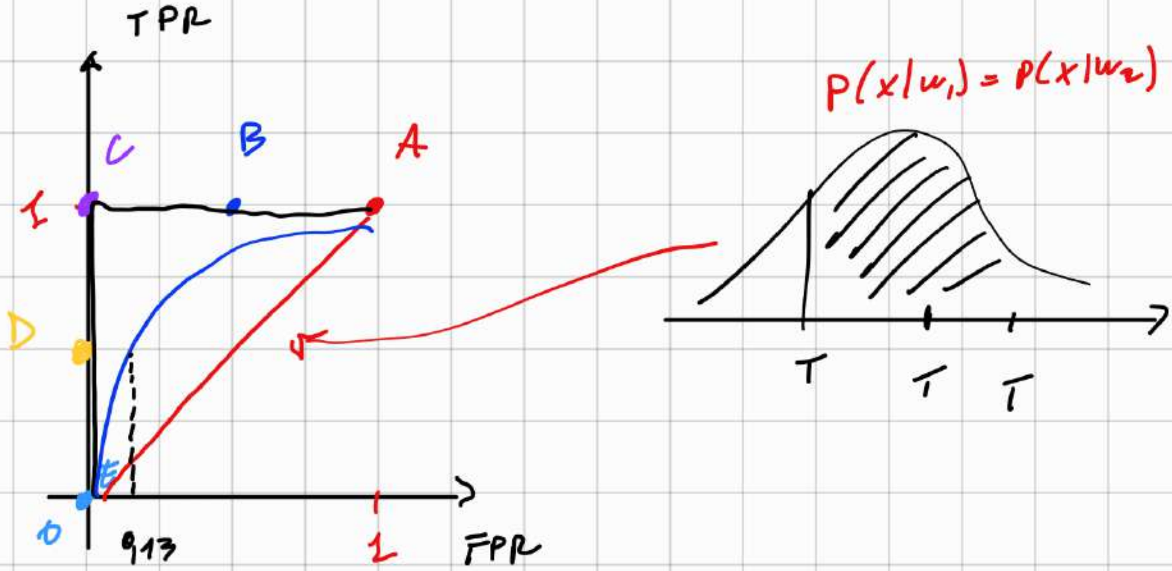
TPR

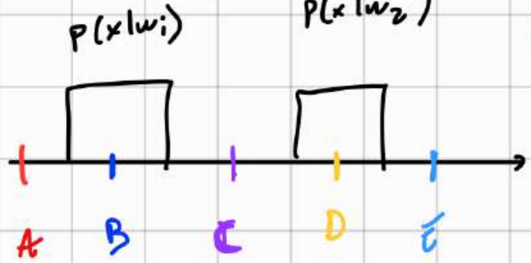
$$\text{SENSITIVITY} = \frac{TP}{TP + FN}$$

$$\text{SPECIFICITY} = \frac{TN}{TN + FP}$$

$$F_1 = \frac{2P \cdot S}{P + S}$$

ROC CURVE





$$\frac{p(x|w_1)}{p(x|w_2)} \gtrless \eta$$

L RATIO TEST

- 1 => ML
- $\frac{P_1}{P_2}$ => MAP
- $\frac{P_1}{P_2} \cdot \frac{(C\,-\,1)}{(\,-\,)}$ MIN RISK
- 3 ARBITRARY VALUE

COMPARING CLASSIFIER

T TEST