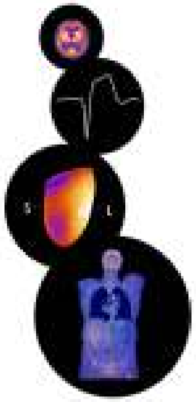


METU Informatics Institute

Min720

Pattern Classification
Bio-Medical Applications

Lecture Notes

by

Neşe Yalabık

Spring 2011

Part 3: Estimation of Parameters

Estimation of Parameters

- Most of the time, we have random samples but not the densities given.
- If the parametric form of the densities are given or assumed, then, using the labeled samples, the parameters can be estimated. (supervised learning)

Maximum Likelihood Estimation of Parameters

- Assume we have a sample set:

$$D = \{X_1, X_2, \ldots, X_n\}$$

- as belonging to a given class. Drawn from $P(X|\omega_j)$

  iid (independently drawn from identically distributed r.v.)

  samples

$\theta_j = [t_1, t_2, ..., t_p]^T$ (unknown parameter vector)

$\theta_j = (\mu_j, \Sigma_j)^T = [\mu_{j1}, \mu_{j2}, ..., \Sigma_{j11}, ...]$ for gaussian

The density function $P(X|\omega_j)$ - assumed to be of known form

So our problem: estimate θ_j using sample set:

$$D_j = \{X_{j1}, X_{j2}, ..., X_{jn}\} \quad iid$$

Now drop j and assume a single density function.

$\hat{\theta}$: estimate of θ

Anything can be an estimate. What is a good estimate?

- Should converge to actual values
- Unbiased etc

Consider the mixture density $L(\theta) = P(D|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$

(due to statistical independence)

$L(\theta)$ is called “likelihood function”

$\hat{\theta} - \theta$ that maximizes $L(\theta)$

(Best agrees with drawn samples.)

if θ is a singular,

Then find θ such that $\frac{dL}{d\theta} = 0$ and for solving for θ.

When θ is a vector, then $L = L(t_1, t_2, ..., t_p)$

$$\nabla_\theta L = 0$$

∇ : gradient of L wrt θ

Where: $$\nabla_\theta = \begin{bmatrix} \frac{\partial L}{\partial t_1} \\ \frac{\partial L}{\partial t_2} \\ .. \\ \frac{\partial L}{\partial t_p} \end{bmatrix} = 0$$

Therefore $\hat{\theta} = \arg\max L(\theta)$

or $\hat{\theta} = \arg\max \ln L(\theta) = \arg\max l(\theta)$ (log-likelihood)

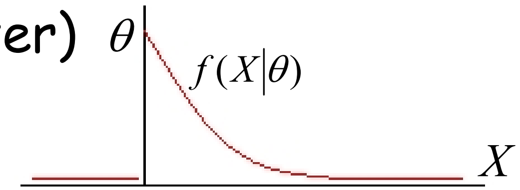
(Be careful not to find the minimum with derivatives)

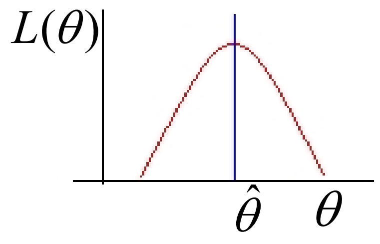
Example 1:

Consider an exponential distribution

$$f(X;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(single feature, single parameter)





With a random sample

$\{X_1, X_2, ..., X_n\}$

:

$$L(\theta) = f(X_1, X_2, ..., X_n|\theta) = \prod_{i=1}^{n} \theta . e^{-\theta . x_i} \qquad x \geq 0$$

valid for

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln\theta - \theta \sum_{i=1}^{n} x_i = n \ln\theta - \theta \sum_{i=1}^{n} x_i$$

$$\frac{dl}{d\theta} = \frac{d \ln L(\theta)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \frac{n}{\hat{\theta}} = \sum_{i=1}^{n} x_i \Rightarrow \boxed{\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}}$$

(inverse of average)

Example 2:

- Multivariate Gaussian with unknown mean vector M. Assume $\sum$ is known.
- k samples from the same distribution:

$$X_1, X_2, \ldots\ldots\ldots, X_k \qquad \text{(iid)}$$

$$L(X \mid M) = \prod_{i=1}^{k} \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(X_i - M)^T \Sigma^{-1} (X_i - M)}$$

$$\nabla l = \nabla_M \log L = \sum_{i=1}^{k} \nabla_M \log \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(X_i - M)^T \Sigma^{-1} (X_i - M)}$$

$$= \sum_{i=1}^{k} \nabla_M \left(\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X_i - M)^T \Sigma^{-1} (X_i - M)\right)$$

$$= \sum_{i=1}^{k} (\Sigma^{-1}(X_i - \hat{M})) \qquad \text{(linear algebra)}$$

$$\Rightarrow 0 = \Sigma^{-1}\left(\sum_{i=1}^{k} X_i - k\hat{M}\right)$$

$$\hat{M} = \frac{1}{k}\sum_{i=1}^{k} X_i$$ (sample average or sample mean)

Estimation of $\sum$ when it is unknown.

(Do it yourself: not so simple)

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \hat{M})(X_k - \hat{M})^T$$ $\hat{\Sigma}$:sample covariance

where $\hat{M}$ is the same as above.

Biased estimate : $E(\hat{\sigma}^2) \neq \sigma^2$

$$= \frac{n-1}{n}\sigma^2$$

use $\frac{1}{n-1}\sum \cdots\cdots$ for an unbiased estimate.

Example 3:

Binary variables with unknown parameters $p_i, 1 \le i \le n$

(n parameters)

$$\log P(X) = \sum_{i=1}^{n} x_i \log p_i + \sum_{i=1}^{n} (1 - x_i) \log(1 - p_i)$$

So,

$$l = \log L = \sum_{j=1}^{k} \log P(X_j) \text{ k samples}$$

$$= \sum_{j=1}^{k} (\sum_{i=1}^{n} x_{ij} \log p_i + \sum_{i=1}^{n} (1 - x_{ij}) \log(1 - p_i))$$

here x_{ij} is the i^{th} element of j^{th} sample X_j .

So,

$$\nabla_{p_i} \log L = \begin{bmatrix} \frac{\partial}{\partial p_1} \log L \\ \frac{\partial}{\partial p_2} \log L \\ \vdots \\ \frac{\partial}{\partial p_n} \log L \end{bmatrix}$$

$$\frac{\partial}{\partial p_i} \log L = \sum_{j=1}^{k} (\frac{x_{ij}}{p_i} - ((1 - x_{ij})(1 - p_i))$$

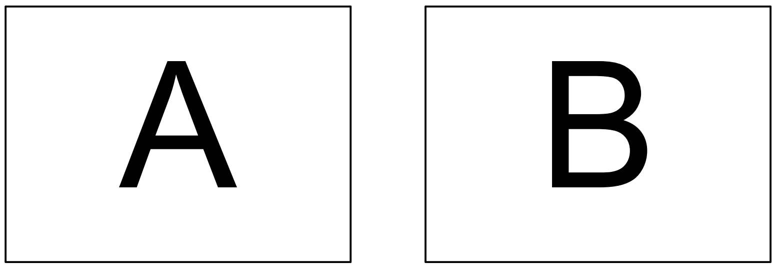
$$\Rightarrow 0 = \frac{1}{\hat{p}_i} \sum_{j=1}^{k} x_{ij} - \frac{1}{1 - \hat{p}_i} \sum_{j=1}^{k} (1 - x_{ij})$$

$$\Rightarrow \hat{p}_i = \frac{1}{k} \sum_{j=1}^{k} x_{ij}$$

❖ $\hat{p}_i$ is the sample average of the feature.

- Since X_i is binary, $\sum_{j=1}^{k} x_{ij}$ will be the same as counting the occurances of '1'.

- Consider character recognition problem with binary matrices.



- For each pixel, count the number of 1's and this is the estimate of p_i .

METU Informatics Institute

Min720

Pattern Classification Bio-Medical Applications

Lecture Notes

by

Neşe Yalabık

Spring 2011

Part 4: Features and Feature Extraction

Problems of Dimensionality and Feature Selection

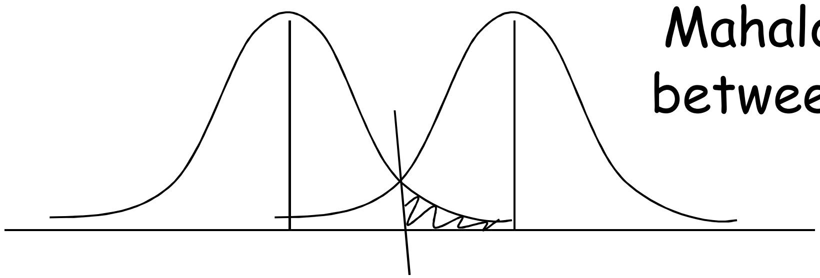
- Are all features independent? Especially in binary features, we might have >100.
- The classification accuracy vs. size of feature set.
- Consider the Gaussian case with same Σ for both categories.

$$P(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du$$

(assuming a priori probabilities are the same) (e:error)

- where r^2 is the square of mahalonobis distance between class means.

$$r^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$



Mahalonobis distance between μ_1 and μ_2 (the means)

- P(e) decreases as r increases (the distance between the means).

If
$$\Sigma = \begin{bmatrix} \sigma_1^{\;2} & & & 0 \\ & \sigma_2^{\;2} & & \\ & & . & \\ 0 & & & \sigma_d^{\;2} \end{bmatrix}$$
(all features statistically independent.)

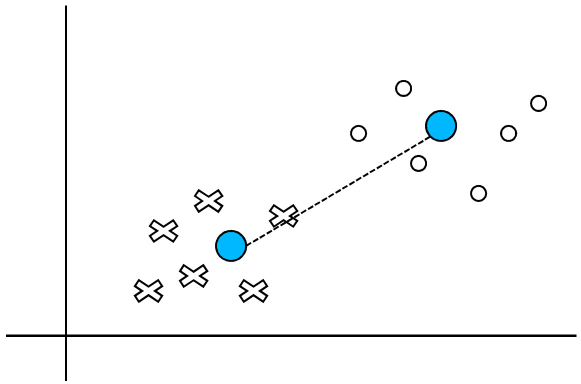
then

$$\boxed{r^2 = \sum_{i=1}^{d} \left(\frac{m_{i1} - m_{i2}}{\sigma_i}\right)^2 = \sum_{i=1}^{d} \frac{(m_{i1} - m_{i2})^2}{\sigma_i^{\;2}}}$$

We conclude from here that

1-Most useful features are the ones with large distance and small variance.

2-Each feature contributes to reduce the probability of error.



- When r increases, probability of error decreases.
- Best features are the ones with distant means and small variances.

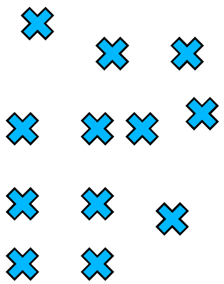
➢ <u>So add new features if the ones we already have are not adequate (more features, decreasing prob. of error.)</u>

➢ But it was shown that adding new features after some point leads to worse performance.

- ✓ Find statistically independent features
- ✓ Find discriminating features
- ✓ Computationally feasible features

Principal Component Analysis (PCA) (Karhunen-Loeve Transform)

- Finds (reduces the set) to statistically independent features.



$X_1, X_2, \ldots\ldots, X_n$ vectors

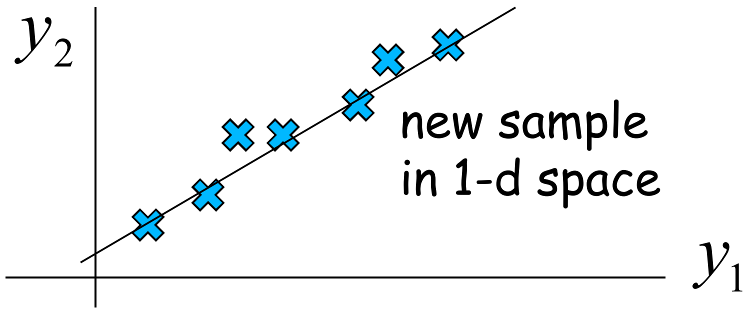
Find a representative X_0

Squared error criterion

Eliminating Redundant Features

$X = [x_1, \ldots\ldots\ldots, x_d]^T$ is to be found using a larger set

$$Y = [y_1, \ldots\ldots\ldots, y_m]^T$$



y_1, y_2 Features that are linearly dependent

So we either

- Throw one away
- Generate a new feature using y_1 and y_2 (ex:projections of the points to a line)
- Form a linear combination of features.

$$\left.\begin{array}{l} x_1 = f_1(y_1, \ldots\ldots, y_m) \\ x_2 = f_2(y_1, \ldots\ldots, y_m) \\ \\ x_d = f_d(y_1, \ldots\ldots, y_m) \end{array}\right\}$$ Linear functions

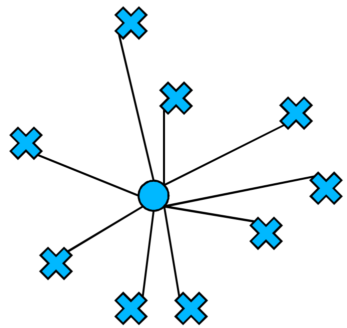
$X = WY$ A linear transformation

W? Can be found by: K-L expansion, Principal Component Analysis

W :above are satisfied (class discrimination and independent x1, x2,..).

$X_1, \ldots\ldots\ldots, X_n$ represented with a single vector X_0 .

-Find a vector X_0 so that sum of the squared distances to X_0 is minimum(Zero degree representation).



$$J_0(X_0) = \sum_{k=1}^{n} \left\| X_0 - X_k \right\|^2$$ squared-error criterion

Find X_0 that maximizes J_0 .

Solution is given by the sample mean.

$$M = \frac{1}{n} \sum X_k$$

$$J_0(X_0) = \sum \left\| (X_0 - M) - (X_k - M) \right\|^2$$

$$= \sum \left\| X_0 - M \right\|^2 - \sum 2(X_0 - M)^T (X_K - M) + \sum \left\| X_k - M \right\|^2$$

$$= \sum \left\| X_0 - M \right\|^2 - 2(X_0 - M)^T \overbrace{\sum (X_K - M)}^{0} + \sum \left\| X_k - M \right\|^2$$

$$= \sum \|X_0 - M\|^2 + \underbrace{\sum \|X_k - M\|^2}_{\text{Independent of } X_0}$$

Where $X_0 = M$,

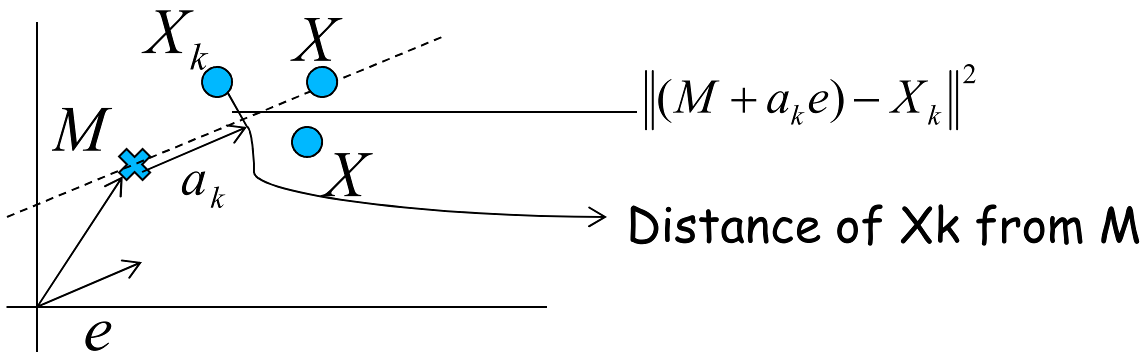
this expression is minimized.

Consider now 1-d representation from 2-d.

-The line should pass through the sample mean.

$$X = M + ae$$

unit vector in the direction of line



- Now how to find best e that minimizes $J_1 = \sum \|(M + a_k e) - X_k\|^2$
- It turns out that given the scatter matrix

$$S = \sum_{k=1}^{n} (X_k - M)(X_k - M)^T$$

- e must be the eigenvector of the scatter matrix with the largest eigenvalue lambda λ .

$$Se = \lambda e$$

- That is, we project the data onto a line through the sample mean in the direction of the eigenvector of the scatter matrix with largest eigenvalue.
- Now consider d dimensional projection

$$\boxed{X = M + \sum_{i=1}^{d} a_i e_i}$$

- Here $e_1, \ldots\ldots, e_d$ are d eigenvectors of the scatter matrix having largest eigenvalues.

Coefficients a_i are called principal components.

- So each m dimensional feature vector is transferred to d dimensional space since the components a_i are given as

$$\boxed{a_{ki} = e_i^T (X_k - M)}$$

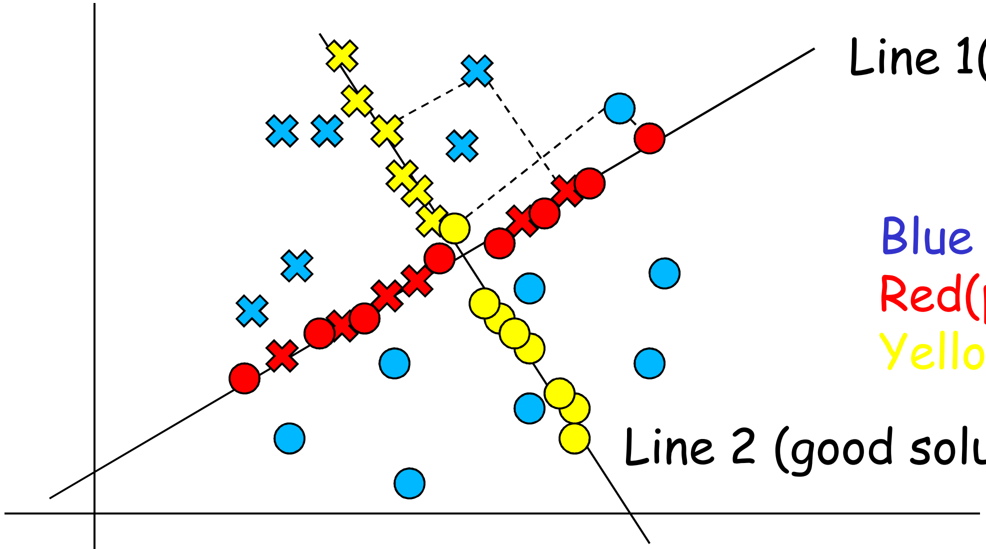
- Now represent our new feature vector's elements

So

$$a_{1i} = e_1^T (X_k - M)$$
$$a_{2i} = e_2^T (X_k - M)$$
$$\vdots$$
$$a_{d'i} = e_d^T (X_k - M)$$

FISHER'S LINEAR DISCRIMINANT

- Curse of dimensionality. More features, more samples needed.
- We would like to choose features with more discriminating ability.
- Reduces the dimension of the problem to one in simplest form.
- Seperates samples from different categories.
- Consider samples from 2 different categories now.



-Find a line so that the projection separates the samples best.

Same as:

Apply a transformation to samples X to result with a scalar such that $y = W^T X$

Fisher's criterion function

$$J(W) = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}$$ is maximized, where

$$\mu_i = \frac{1}{n_i} \sum_{y \in C_i} y$$

$$\sigma_i^{\ 2} = \frac{1}{n_i} \sum_{y \in C_i} (y - \mu_i)^2$$

- This reduces the problem to <u>1d</u>, by keeping the classes most distant from each other.
- But if we write μ_i and $\sigma_i^{\ 2}$ in terms of M_i and Σ_i

$$M_i = \frac{1}{n_i} \sum_{X \in C_i} X$$

$$\mu_i = \frac{1}{n_i} \sum \underbrace{W^T X}_{y} = W^T M$$

$$\sigma_i^{\ 2} = \frac{1}{n_i} \sum_{X \in C_i} (W^T X - W^T M_i)^2 = \frac{1}{n_i} \sum_{X \in C_i} (W^T (X - M_i))^2$$

$$= \frac{1}{n_i} \sum W^T (X - M_i)(X - M_i)^T W = W^T (\frac{1}{n_i} (\sum (X - M_i)(X - M_i)^T) W)$$

$$= W^T S_i W$$

Then, $$(\mu_1 - \mu_2)^2 = (W^T M_1 - W^T M_2)^2 = \left[W^T (M_1 - M_2)\right]^2$$

$$= W^T \underbrace{(M_1 - M_2)(M_1 - M_2)^T}_{S_B} W = W^T S_B W$$

$$\sigma_1{}^2 + \sigma_2{}^2 = W^T \underbrace{(S_1 + S_2)}_{S_W} W = W^T S_W W$$

S_B- within class scatter matrix

S_W- between class scatter matrix

- Then , maximize

$$J(W) = \frac{W^T S_B W}{W^T S_W W}$$

Rayleigh quotient

- It can be shown that W that maximizes J can be found by solving the eigenvalue problem again:

$$S_W^{-1} S_B W = \lambda W$$

and the solution is given by

$$\boxed{W = S_W^{-1}(M_1 - M_2)}$$

- Optimal if the densities are gaussians with equal covariance matrices. That means reducing the dimension does not cause any loss.

Multiple Discriminant Analysis: c category problem.

A generalization of 2-category problem.

Non-Parametric Techniques

- Density Estimation
- Use samples directly for classification
  - Nearest Neighbor Rule
  - 1-NN
  - k-NN
- Linear Discriminant Functions: *gi*(X) is linear.