FEATURE SELECTION

❖ The goals:

- ➢ Select the "optimum" number l of features
- ➢ Select the "best" l features

❖ Large l has a three-fold disadvantage:

- ➢ High computational demands
- ➢ Low generalization performance
- ➢ Poor error estimates

1

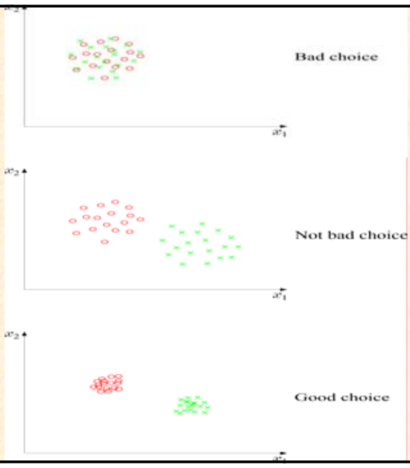
➢ Given N

- l must be large enough to learn
  - what makes classes different
  - what makes patterns in the same class similar
- l must be small enough not to learn what makes patterns of the same class different
- In practice, $l < N/3$ has been reported to be a sensible choice for a number of cases

➢ Once l has been decided, choose the l most informative features

- Best: **Large** between class distance, **Small** within class variance

2



3

❖ The basic philosophy

- ➢ Discard individual features with poor information content
- ➢ The remaining information rich features are examined jointly as vectors

4

❖ Feature Selection based on statistical Hypothesis Testing

➢ The Goal: For each individual feature, find whether the values, which the feature takes for the different classes, differ significantly.

That is, answer

- $H_1: \ \theta_1 \neq \theta_0$: The values differ significantly
- $H_0: \ \theta_1 = \theta_0$: The values do not differ significantly

If they do not differ significantly reject feature from subsequent stages.

❖ Hypothesis Testing Basics

5

➢ The steps:

- N measurements $x_i, i = 1,2,...,N$ are known
- Define a function of them

  $q = f(x_1, x_2, ..., x_N)$: test statistic

  so that $\boxed{p_q(q;\theta)}$ is easily parameterized in terms of θ.
- Let D be an interval, where q has a high probability to lie under H_0, i.e., $p_q(q|\theta_0)$
- Let $\bar{D}$ be the complement of D

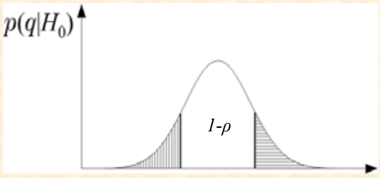
  $D \longrightarrow$ Acceptance Interval

  $\bar{D} \longrightarrow$ Critical Interval
- If q, resulting from $x_1, x_2, ..., x_N$, lies in D we accept H_0, otherwise we reject it.

6

- Probability of an error

$$p_q(q \in \bar{D}|H_0) = \rho$$



- ρ is preselected and it is known as the significance level.

❖ Application: The known variance case:

- Let x be a random variable and the experimental samples, $x_i = 1,2,\dots,N$, are assumed mutually independent. Also let

$$E[x] = \mu$$
$$E[(x-\mu)^2] = \sigma^2$$

- Compute the sample mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This is also a random variable with mean value

$$E[\bar{x}] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

That is, it is an Unbiased Estimator

- The variance $\sigma_{\bar{x}}^2$

$$E[(\bar{x}-\mu)^2] = E[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)^2]$$
$$= \frac{1}{N^2}\sum_{i=1}^{N} E[(x_i-\mu)^2] + \frac{1}{N^2}\sum_i\sum_j E[(x_i-\mu)(x_j-\mu)]$$

Due to independence

$$\sigma_{\bar{x}}^2 = \frac{1}{N}\sigma_x^2$$

That is, it is Asymptotically Efficient

- Hypothesis test

$$H_1: E[x] \neq \hat{\mu}$$
$$H_0: E[x] = \hat{\mu}$$

- Test Statistic: Define the variable

$$\boxed{q = \frac{\bar{x}-\hat{\mu}}{\sigma/\sqrt{N}}}$$

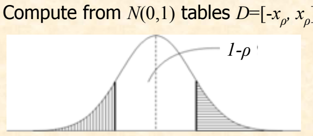
- Central limit theorem under H_0

$$p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{N(\bar{x}-\hat{\mu})^2}{2\sigma^2}\right)$$

- Thus, under H_0

$$p_q(q) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{q^2}{2}\right) \quad q \approx N(0,1)$$

- The decision steps
  - Compute q from x_i, $i=1,2,\dots,N$
  - Choose significance level ρ
  - Compute from $N(0,1)$ tables $D=[-x_\rho, x_\rho]$



  - if $q \in D$ accept H_0

    if $q \in \bar{D}$ reject H_0

- An example: A random variable x has variance $\sigma^2=(0.23)^2$. $N=16$ measurements are obtained giving $\bar{x}=1.35$. The significance level is $\rho=0.05$.

Test the hypothesis

$$H_0: \mu = \hat{\mu} = 1.4$$
$$H_1: \mu \neq \hat{\mu}$$

- Since σ² is known, $q = \frac{\bar{x}-\hat{\mu}}{\sigma/4}$ is $N(0,1)$.

From tables, we obtain the values with acceptance intervals $[-x_\rho, x_\rho]$ for normal $N(0,1)$

$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.28	1.44	1.64	1.96	2.32	2.57	3.09	3.29

- Thus

$$\text{Prob}\left\{-1.967 < \frac{\bar{x}-\hat{\mu}}{0.23/4} < 1.967\right\} = 0.95$$

or

$$\text{Prob}\{-0.113 < \bar{x}-\hat{\mu} < 0.113\} = 0.95$$

or

$$\boxed{\text{Prob}\{1.237 < \hat{\mu} < 1.463\} = 0.95}$$

- Since $\hat{\mu} = 1.4$ lies within the above acceptance interval, we accept H_0, i.e.,

$$\mu = \hat{\mu} = 1.4$$

The interval [1.237, 1.463] is also known as confidence interval at the $1\text{-}\rho = 0.95$ level.

We say that: There is no evidence at the 5% level that the mean value is not equal to $\hat{\mu}$

13

- The Unknown Variance Case

  - Estimate the variance. The estimate

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

is unbiased, i.e.,

$$E[\hat{\sigma}^2] = \sigma^2$$

  - Define the test statistic

$$q = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{N}}$$

14

- This is no longer Gaussian. If x is Gaussian, then q follows a t-distribution, with N-1 degrees of freedom

- An example:

x is Gaussian, $N = 16$, obtained from measurements, $\bar{x} = 1.35$ and $\hat{\sigma}^2 = (0.23)^2$. Test the hypothesis

$$H_0:\ \mu = \hat{\mu} = 1.4$$

at the significance level $\rho = 0.025$.

15

- Table of acceptance intervals for t-distribution

Degrees of Freedom	1-ρ	0.9	0.95	0.975	0.99
12		1.78	2.18	2.56	3.05
13		1.77	2.16	2.53	3.01
14		1.76	2.15	2.51	2.98
15		1.75	2.13	2.49	2.95
16		1.75	2.12	2.47	2.92
17		1.74	2.11	2.46	2.90
18		1.73	2.10	2.44	2.88

- $$\text{Prob}\left\{-2.49 < \frac{\bar{x} - \hat{\mu}}{\hat{\sigma}/4} < 2.49\right\}$$

$$1.207 < \hat{\mu} < 1.493$$

Thus, $\hat{\mu} = 1.4$ is accepted

16

- Application in Feature Selection

  - The goal here is to test against zero the difference μ_1-μ_2 of the respective means in ω_1, ω_2 of a single feature.

  - Let x_i $i=1,\ldots,N$, the values of a feature in ω_1

  - Let y_i $i=1,\ldots,N$, the values of the same feature in ω_2

  - Assume in both classes $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown or not)

  - The test becomes

$$H_0:\ \Delta\mu = \mu_1 - \mu_2 = 0$$

$$H_1:\ \Delta\mu \neq 0$$

17

- Define

z=x-y

- Obviously

$E[z]$=μ_1-μ_2

- Define the average

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i) = \bar{x} - \bar{y}$$

- Known Variance Case: Define

$$q = \frac{(\bar{x} - \bar{y}) - (\hat{\mu}_1 - \hat{\mu}_2)}{\sigma\sqrt{\frac{2}{N}}}$$

- This is $N(0,1)$ and one follows the procedure as before.

18

- Unknown Variance Case:
  Define the test statistic

$$q = \frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{S_z\sqrt{\frac{2}{N}}}$$

$$S_z^2 = \frac{1}{2N-2}\left(\sum_{i=1}^{N}(x_i-\bar{x})^2 + \sum_{i=1}^{N}(y_i-\bar{y})^2\right)$$

  - q is t-distribution with $2N$-2 degrees of freedom,
  - Then apply appropriate tables as before.

- Example: The values of a feature in two classes are:

  ω_1: 3.5, 3.7, 3.9, 4.1, 3.4, 3.5, 4.1, 3.8, 3.6, 3.7

  ω_2: 3.2, 3.6, 3.1, 3.4, 3.0, 3.4, 2.8, 3.1, 3.3, 3.6

  Test if the mean values in the two classes differ significantly, at the significance level ρ=0.05

- We have

$$\omega_1: \ \bar{x} = 3.73, \ \hat{\sigma}_1^2 = 0.0601$$

$$\omega_2: \ \bar{y} = 3.25, \ \hat{\sigma}_2^2 = 0.0672$$

  For N=10

$$S_z^2 = \frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$$

$$q = \frac{(\bar{x}-\bar{y})-0}{S_z\sqrt{\frac{2}{10}}}$$

$$\boxed{q = 4.25}$$

- From the table of the t-distribution with $2N$-2=18 degrees of freedom and ρ=0.05, we obtain D=[-2.10,2.10] and since q=4.25 is outside D, H_1 is accepted and the feature is selected.

- Class Separability Measures

  The emphasis so far was on individually considered features. However, such an approach cannot take into account existing correlations among the features. That is, two features may be rich in information, but if they are highly correlated we need not consider both of them. To this end, in order to search for possible correlations, we consider features jointly as elements of vectors. To this end:

  - Discard poor in information features, by means of a statistical test.
  - Choose the maximum number, ℓ, of features to be used. This is dictated by the specific problem (e.g., the number, N, of available training patterns and the type of the classifier to be adopted).

- Combine remaining features to search for the "best" combination. To this end:

  - Use different feature combinations to form the feature vector. Train the classifier, and choose the combination resulting in the best classifier performance.
    A major disadvantage of this approach is the high complexity. Also, local minima, may give misleading results.
  - Adopt a class separability measure and choose the best feature combination against this cost.

- Class separability measures: Let $\underline{x}$ be the current feature combination vector.
  - Divergence. To see the rationale behind this cost, consider the two – class case. Obviously, if on the average the value of $\ln\frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)}$ is close to zero, then $\underline{x}$ should be a poor feature combination. Define:

    – $D_{12} = \int_{-\infty}^{+\infty} p(\underline{x}|\omega_1)\ln\frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)}d\underline{x}$

    – $D_{21} = \int_{-\infty}^{+\infty} p(\underline{x}|\omega_2)\ln\frac{p(\underline{x}|\omega_2)}{p(\underline{x}|\omega_1)}d\underline{x}$

    – $d_{12} = D_{12} + D_{21}$

    d_{12} is known as the divergence and can be used as a class separability measure.

– For the multi-class case, define d_{ij} for every pair of classes ω_i, ω_j and the average divergence is defined as

$$d = \sum_{i=1}^{M}\sum_{j=1}^{M} P(\omega_i)P(\omega_j)d_{ij}$$

– Some properties:

$$d_{ij} \geq 0$$

$$d_{ij} = 0, \text{if } i = j$$

$$d_{ij} = d_{ji}$$

– Large values of d are indicative of good feature combination.

➢ Scatter Matrices. These are used as a measure of the way data are scattered in the respective feature space.

- Within-class scatter matrix

$$S_w = \sum_{i=1}^{M} P_i S_i$$

where

$$S_i = E\left[\left(\underline{x} - \underline{\mu}_i\right)\left(\underline{x} - \underline{\mu}_i\right)^T\right]$$

and

$$P_i \equiv P(\omega_i) \approx \frac{n_i}{N}$$

n_i the number of training samples in ω_i.

Trace $\{S_w\}$ is a measure of the average variance of the features.

25

- Between-class scatter matrix

$$S_b = \sum_{i=1}^{M} P_i \left(\underline{\mu}_i - \underline{\mu}_0\right)\left(\underline{\mu}_i - \underline{\mu}_0\right)^T$$

$$\underline{\mu}_0 = \sum_{i=1}^{M} P_i \underline{\mu}_i$$

Trace $\{S_b\}$ is a measure of the average distance of the mean of each class from the respective global one.

- Mixture scatter matrix

$$S_m = E\left[\left(\underline{x} - \underline{\mu}_0\right)\left(\underline{x} - \underline{\mu}_0\right)^T\right]$$

It turns out that:

$$S_m = S_w + S_b$$

26

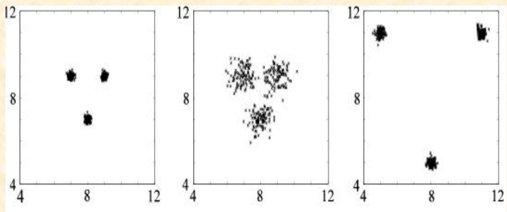
➢ Measures based on Scatter Matrices.

- $J_1 = \dfrac{\text{Trace}\{S_m\}}{\text{Trace}\{S_w\}}$
- $J_2 = \dfrac{|S_m|}{|S_w|} = \left|S_w^{-1} S_m\right|$
- $J_3 = \text{Trace}\left\{S_w^{-1} S_m\right\}$
- Other criteria are also possible, by using various combinations of S_m, S_b, S_w.

The above J_1, J_2, J_3 criteria take high values for the cases where:

- Data are clustered together within each class.
- The means of the various classes are far.

27



(a) (b) (c)

28

- Fisher's discriminant ratio. In one dimension and for two equiprobable classes the determinants become:

$$|S_w| \propto \sigma_1^2 + \sigma_2^2$$

$$|S_b| \propto (\mu_1 - \mu_2)^2$$

and

$$\frac{|S_b|}{|S_w|} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

known as Fischer's ratio.

29

❖ Ways to combine features:

Trying to form all possible combinations of ℓ features from an original set of m selected features is a computationally hard task. Thus, a number of suboptimal searching techniques have been derived.

➢ Sequential forward selection. Let x_1, x_2, x_3, x_4 the available features (m=4). The procedure consists of the following steps:

- Adopt a class separability criterion (could also be the error rate of the respective classifier). Compute its value for ALL features considered jointly $[x_1, x_2, x_3, x_4]^T$.
- Eliminate one feature and for each of the possible resulting combinations, that is $[x_1, x_2, x_3]^T$, $[x_1, x_2, x_4]^T$, $[x_1, x_3, x_4]^T$, $[x_2, x_3, x_4]^T$, compute the class reparability criterion value C. Select the best combination, say $[x_1, x_2, x_3]^T$.

30

- From the above selected feature vector eliminate one feature and for each of the resulting combinations, $[x_1, x_2]^T$, $[x_2, x_3]^T$, $[x_1, x_3]^T$ compute C and select the best combination.

The above selection procedure shows how one can start from m features and end up with the "best" ℓ ones. Obviously, the choice is suboptimal. The number of required calculations is:

$$1+\frac{1}{2}\left((m+1)m-\ell(\ell+1)\right)$$

In contrast, a full search requires:

$$\binom{m}{\ell}=\frac{m!}{\ell!(m-\ell)!}$$

operations.

31

➢ Sequential backward selection. Here the reverse procedure is followed.

- Compute C for each feature. Select the "best" one, say x_1
- For all possible 2D combinations of x_1, i.e., $[x_1, x_2]$, $[x_1, x_3]$, $[x_1, x_4]$ compute C and choose the best, say $[x_1, x_3]$.
- For all possible 3D combinations of $[x_1, x_3]$, e.g., $[x_1, x_3, x_2]$, etc., compute C and choose the best one.

The above procedure is repeated till the "best" vector with ℓ features has been formed. This is also a suboptimal technique, requiring:

$$\ell m-\frac{\ell(\ell-1)}{2}$$

operations.

32

➢ Floating Search Methods

The above two procedures suffer from the nesting effect. Once a bad choice has been done, there is no way to reconsider it in the following steps.

In the floating search methods one is given the opportunity in reconsidering a previously discarded feature or to discard a feature that was previously chosen.

The method is still suboptimal, however it leads to improved performance, at the expense of complexity.

33

➢ Remarks:

- Besides suboptimal techniques, some optimal searching techniques can also be used, provided that the optimizing cost has certain properties, e.g., monotonic.
- Instead of using a class separability measure (filter techniques) or using directly the classifier (wrapper techniques), one can modify the cost function of the classifier appropriately, so that to perform feature selection and classifier design in a single step (embedded) method.
- For the choice of the separability measure a multiplicity of costs have been proposed, including information theoretic costs.

34

Optimal Feature Generation

❖ In general, feature generation is a problem-dependent task. However, there are a few general directions common in a number of applications. We focus on three such alternatives.

➢ Optimized features based on Scatter matrices (Fisher's linear discrimination).

- The goal: Given an original set of m measurements $\underline{x}\in\Re^m$, compute $\underline{y}\in\Re^\ell$, by the linear transformation

$$\underline{y}=A^T\underline{x}$$

so that the J_3 scattering matrix criterion involving S_w, S_b is maximized. A^T is an $\ell x m$ matrix.

35

- The basic steps in the proof:
  - $J_3 = \text{trace}\{S_w^{-1} S_m\}$
  - $S_{yw} = A^T S_{xw} A$, $S_{yb} = A^T S_{xb} A$,
  - $J_3(A) = \text{trace}\{(A^T S_{xw} A)^{-1} (A^T S_{xb} A)\}$
  - Compute A so that $J_3(A)$ is maximum.
- The solution:
  - Let B be the matrix that diagonalizes simultaneously matrices S_{yw}, S_{yb}, i.e:

$$B^T S_{yw} B = I\,,\; B^T S_{yb} B = D$$

where B is a $\ell x \ell$ matrix and D a $\ell x \ell$ diagonal matrix.

36

– Let $C=AB$ an $m\times\ell$ matrix. If A maximizes $J_3(A)$ then

$$\left(S_{xw}^{-1}S_{xb}\right)C = CD$$

The above is an eigenvalue-eigenvector problem. For an M-class problem, $S_{xw}^{-1}S_{xb}$ is of rank M-1.

- If ℓ=M-1, choose C to consist of the M-1 eigenvectors, corresponding to the non-zero eigenvalues.

$$\underline{y} = C^T \underline{x}$$

The above guarantees maximum J_3 value. In this case: $J_{3,x} = J_{3,y}$.

- For a two-class problem, this results to the well known Fisher's linear discriminant

$$\underline{y} = \left(\underline{\mu}_1 - \underline{\mu}_2\right) S_{xw}^{-1}\underline{x}$$

For Gaussian classes, this is the optimal Bayesian classifier, with a difference of a threshold value .

37

- If ℓ<M-1, choose the ℓ eigenvectors corresponding to the ℓ largest eigenvectors.
- In this case, $J_{3,y}$<$J_{3,x}$, that is there is loss of information.

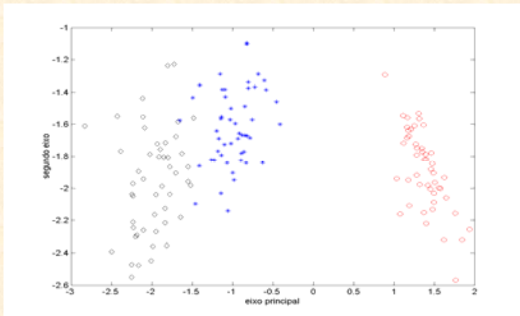
– Geometric interpretation. The vector $\underline{y}$ is the projection of $\underline{x}$ onto the subspace spanned by the eigenvectors of $S_{xw}^{-1}S_{xb}$.

38

```
loadiris
data=iris(:,1:4);
m1=mean(data(1:50,:));
m2=mean(data(51:100,:));
m3=mean(data(101:150,:));

m=(m1+m2+m3)/3;

sb=(m1-m)'*(m1-m)+(m2-m)'*(m2-m)+(m3-m)'*(m3-m);

s1=zeros(4,4);
s2=s1;
s3=s1;
for i=1:50
    s1=s1+(data(i,:)-m1)'*(data(i,:)-m1);
end
for i=51:100
    s2=s2+(data(i,:)-m2)'*(data(i,:)-m2);
end
for i=101:150
    s3=s3+(data(i,:)-m3)'*(data(i,:)-m3);
end

sw=s1+s2+s3;

[v,d]=eig(inv(sw)*sb)
w=[v(:,1), v(:,2)]
k=w'*data';
plot(k(1,1:50),k(2,1:50),'ro',k(1,51:100),k(2,51:100),'b*',k(1,101:150),k(2,101:150),'kd');
xlabel('eixo principal');
ylabel('segundo eixo');
```

39

Resultado LDA - Iris



40

Resultado LDA - Iris

```
v =

   0.2049   -0.0090    0.3398   -0.6672
   0.3871   -0.5890    0.1988    0.4427
  -0.5465    0.2543    0.2728    0.4688
  -0.7138   -0.7670   -0.8779   -0.3729

d =

   0.6454         0         0         0
        0    0.0056         0         0
        0         0    0.0000         0
        0         0         0   -0.0000

w =

   0.2049   -0.0090  sepal length
   0.3871   -0.5890  sepal width
  -0.5465    0.2543  petal length
  -0.7138   -0.7670  petal width
  1º eixo    2º eixo
```

41

❖ Principal Components Analysis

(The Karhunen – Loève transform):

➢ The goal: Given an original set of m measurements $\underline{x} \in \Re^m$ compute $\underline{y} \in \Re^\ell$

$$\underline{y} = A^T \underline{x}$$

for an orthogonal A, so that the elements of $\underline{y}$ are optimally mutually uncorrelated.

That is

$$E\left[y(i)y(j)\right] = 0,\ i \neq j$$

➢ Sketch of the proof:

$$R_y = E\left[\underline{y}\,\underline{y}^T\right] = E\left[A^T \underline{x}\,\underline{x}^T A\right] = A^T R_x A$$

42

- If A is chosen so that its columns $\underline{a}_i$ are the orthogonal eigenvectors of R_x, then

$$R_y = A^T R_x A = \Lambda$$

where Λ is diagonal with elements the respective eigenvalues λ_i.
- Observe that this is a sufficient condition but not necessary. It **imposes** a specific orthogonal structure on A.

➢ Properties of the solution

- Mean Square Error approximation.

Due to the orthogonality of A:

$$\underline{x} = \sum_{i=0}^{m} y(i)\underline{a}_i, \quad y(i) = \underline{a}_i^{\,T}\underline{x}$$

43

– Define

$$\hat{\underline{x}} = \sum_{i=0}^{\ell-1} y(i)\underline{a}_i$$

– The Karhunen – Loève transform minimizes the square error:

$$E\left[\left\|\underline{x}-\hat{\underline{x}}\right\|^2\right] = E\left[\left\|\sum_{i=\ell}^{m} y(i)\underline{a}_i\right\|^2\right]$$

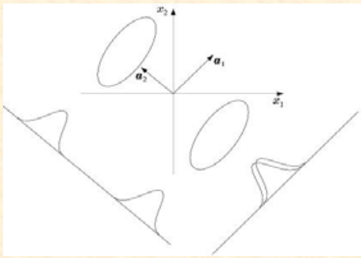
– The error is:

$$E\left[\left\|\underline{x}-\hat{\underline{x}}\right\|^2\right] = \sum_{i=\ell}^{m} \lambda_i$$

It can be also shown that this is the minimum mean square error compared to **any** other representation of x by an ℓ-dimensional vector.

44

– In other words, $\hat{\underline{x}}$ is the projection of $\underline{x}$ into the subspace spanned by the principal ℓ eigenvectors. However, for Pattern Recognition this is not the always the best solution.



45

- Total variance: It is easily seen that

$$\sigma^2_{y(i)} = E\left[y^2(i)\right] = \lambda_i$$

Thus Karhunen – Loève transform makes the total variance maximum.

- Assuming $\underline{y}$ to be a zero mean multivariate Gaussian, then the K-L transform maximizes the entropy:

$$H_y = -E\left[\ln P_{\underline{y}}(\underline{y})\right]$$

of the resulting $\underline{y}$ process.

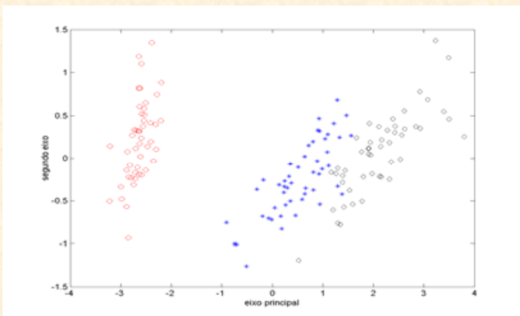
46

PCA

❖ loadiris
❖ data=iris(:,1:4)-repmat(mean(iris(:,1:4)),size(iris,1),1)
❖ [v,d]=eig(data'*data)
❖ w=[v(:,4), v(:,3)]
❖ k=w'*data';
❖ plot(k(1,1:50),k(2,1:50),'ro',k(1,51:100),k(2,51:100),'b*',k(1,101:150), k(2,101:150),'kd');
❖ xlabel('eixo principal');
❖ ylabel('segundo eixo');

47

Resultado PCA - Iris



48

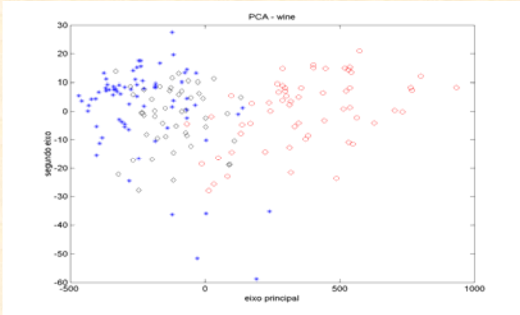
Resultados PCA - Iris

- v =
- -0.3173 0.5810 0.6565 0.3616
- 0.3241 -0.5964 0.7297 -0.0823
- 0.4797 -0.0725 -0.1758 0.8566
- -0.7511 -0.5491 -0.0747 0.3588

- d =
- 3.5288 0 0 0
- 0 11.7001 0 0
- 0 0 36.0943 0
- 0 0 0 629.5013

- w =
- 0.3616 0.6565 sepal length
- -0.0823 0.7297 sepal width
- 0.8566 -0.1758 petal length
- 0.3588 -0.0747 petal width
- 1º eixo 2º eixo

Resultados PCA - Wine

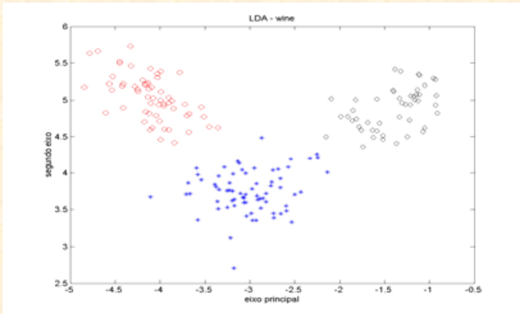


Resultados PCA - Wine

- w =
- 0.0017 -0.0012
- -0.0007 -0.0022
- 0.0002 -0.0046
- -0.0047 -0.0265
- 0.0179 -0.9993
- 0.0010 -0.0009
- 0.0016 0.0001
- -0.0001 0.0014
- 0.0006 -0.0050
- 0.0023 -0.0151
- 0.0002 0.0008
- 0.0007 0.0035
- 0.9998 0.0178

- wine_fields =
- Origin
- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

Resultados LDA - Wine



Resultados LDA - Wine

- w =
- -0.1241 0.2644
- 0.0631 0.0878
- -0.0848 0.7003
- 0.0511 -0.0458
- -0.0008 -0.0001
- 0.2144 -0.0193
- -0.5869 -0.1194
- -0.5506 -0.4592
- 0.0409 -0.0930
- 0.1282 0.0694
- -0.3127 -0.4357
- -0.4017 0.0334
- -0.0009 0.0009

- wine_fields =
- Origin
- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

- Subspace Classification. Following the idea of projecting in a subspace, the subspace classification classifies an unknown $\underline{x}$ to the class whose subspace is closer to $\underline{x}$. The following steps are in order:

  - For each class, estimate the autocorrelation matrix R_i, and compute the m largest eigenvalues. Form A_i, by using respective eigenvectors as columns.

  - Classify $\underline{x}$ to the class ω_i, for which the norm of the subspace projection is maximum

$$\left\|A_i^T \underline{x}\right\| > \left\|A_j^T \underline{x}\right\| \quad \forall i \neq j$$

  According to Pythagoras theorem, this corresponds to the subspace to which $\underline{x}$ is closer.

❖ Independent Component Analysis (ICA)

In contrast to PCA, where the goal was to produce uncorrelated features, the goal in ICA is to produce statistically independent features. This is a much stronger requirement, involving higher to second order statistics. In this way, one may overcome the problems of PCA, as exposed before.

- The goal: Given $\underline{x}$, compute $\underline{y} \in \Re^{\ell}$

$$\underline{y} = W\underline{x}$$

so that the components of $\underline{y}$ are statistically independent. In order the problem to have a solution, the following assumptions must be valid:

  - Assume that $\underline{x}$ is indeed generated by a linear combination of independent components

$$\underline{x} = \Phi\underline{y}$$

Φ is known as the mixing matrix and W as the demixing matrix.

- Φ must be invertible or of full column rank.
- Identifiability condition: All independent components, $y(i)$, must be non-Gaussian. Thus, in contrast to PCA that can always be performed, ICA is meaningful for non-Gaussian variables.
- Under the above assumptions, $y(i)$'s can be uniquely estimated, within a scalar factor.

- Common's method: Given $\underline{x}$, and under the previously stated assumptions, the following steps are adopted:
  - Step 1: Perform PCA on $\underline{x}$:

$$\underline{y} = A^T\underline{x}$$

  - Step 2: Compute a unitary matrix, $\hat{A}$, so that the fourth order cross-cummulants of the transform vector

$$\underline{y} = \hat{A}^T\hat{\underline{y}}$$

are zero. This is equivalent to searching for an $\hat{A}$ that makes the squares of the auto-cummulants maximum,

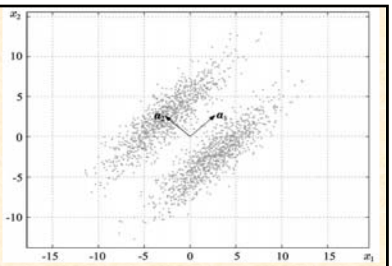
$$\max_{\hat{A}\hat{A}^T} \Psi(\hat{A}) = \sum k_4\left(y(i)\right)^2$$

where, $k_4(\cdot)$ is the 4th order auto-cumulant.

- Step 3: $W = \left(A\hat{A}\right)^T$

- A hierarchy of components: which ℓ to use? In PCA one chooses the principal ones. In ICA one can choose the ones with the least resemblance to the Gaussian pdf.

- Example:



The principal component is $\underline{a}_1$, thus according to PCA one chooses as y the projection of $\underline{x}$ into $\underline{a}_1$. According to ICA, one chooses as y the projection on $\underline{a}_2$. This is the least Gaussian. Indeed:

$$K_4(y_1) = -1.7$$
$$K_4(y_2) = 0.1$$

Observe that across $\underline{a}_2$, the statistics is bimodal. That is, no resemblance to Gaussian.

Measures of nongaussianity

❖ To use nongaussianity in ICA estimation, we must have a quantitative measure of nongaussianity of a random variable, say y. To simplify things, let us assume that y is centered (zero-mean) and has variance equal to one.

❖ **Kurtosis**

- The classical measure of nongaussianity is kurtosis or the fourth-order cumulant. The kurtosis of y is classically defined by

$\mathrm{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2$

❖ **Negentropy**

- A second very important measure of nongaussianity is given by negentropy. Negentropy is based on the information-theoretic quantity of (differential) entropy.
- Negentropy J is defined as follows

$J(\mathbf{y}) = H(\mathbf{y}gauss) - H(\mathbf{y})$