

# Machine learning: basi e sue applicazioni

Christian Salvatore  
Scuola Universitaria Superiore IUSS Pavia

[christian.salvatore@iusspavia.it](mailto:christian.salvatore@iusspavia.it)

# FEATURE SELECTION

# Feature Selection

“is the process of reducing to subset of selected features (variables, predictors) for use in model construction

Informative / Relevant

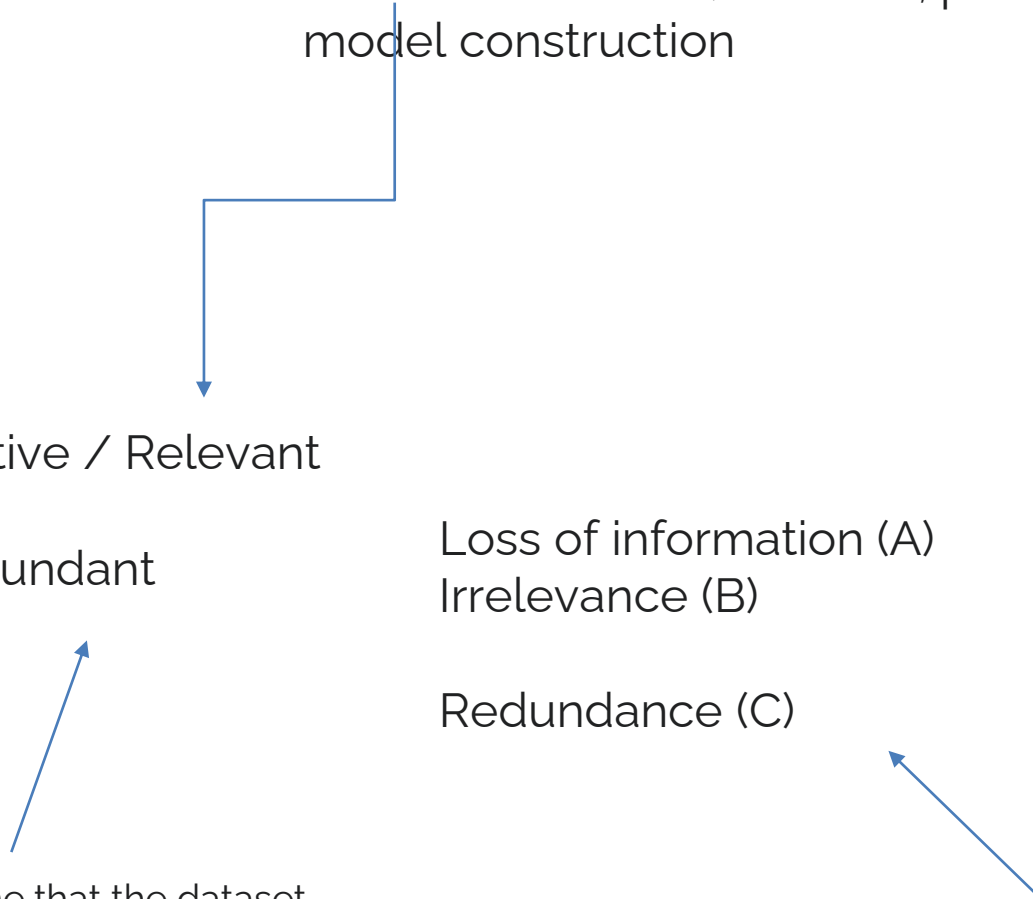
Non redundant

Loss of information (A)  
Irrelevance (B)

Redundance (C)

We assume that the dataset  
is characterized by these issues

Reducing (B) (C)  
without affecting (A)



# Feature Selection

“is the process of reducing to subset of selected features (variables, predictors) for use in model construction



Informative / Relevant

Non redundant

Loss of information  
Irrelevance

Redundance

Reasons for performing a feature-selection step

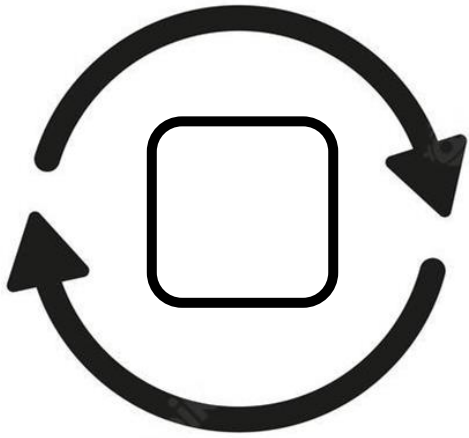
Improving explainability

Reducing training time  
by  
Reducing computational costs

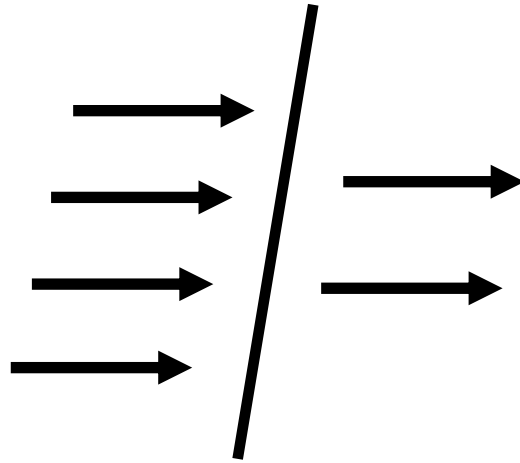
Improving generalization ability  
by  
Reducing model complexity

(Trying to) avoid the curse of dimensionality

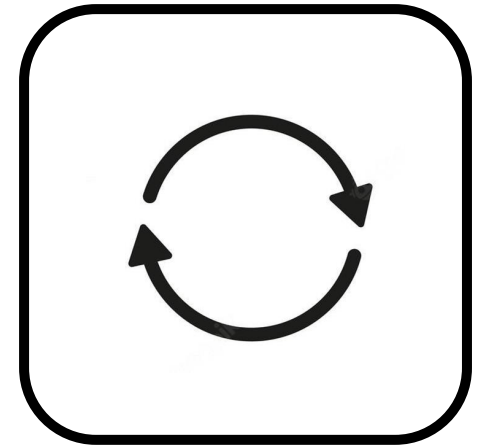
# Feature Selection



WRAPPER



FILTER



EMBEDDED

# Feature selection

	Feature #1	Feature #2	Feature #3	Feature #4
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2
...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7

# Non-redundant features

		Feature #1	Feature #2	Feature #3	Feature #4
CLASS 1	Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7
	Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2
	Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4
CLASS 2	Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2
	...	...	...	...	...
	Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7

Is there a “best approach” to calculate the dependence between two variables?

# Mutual Information

Mutual information is a way to calculate statistical dependence between two variables

(it is related to information gain, which will be introduced in classification trees)

$$I(X,Y) = H(X) - H(X | Y)$$



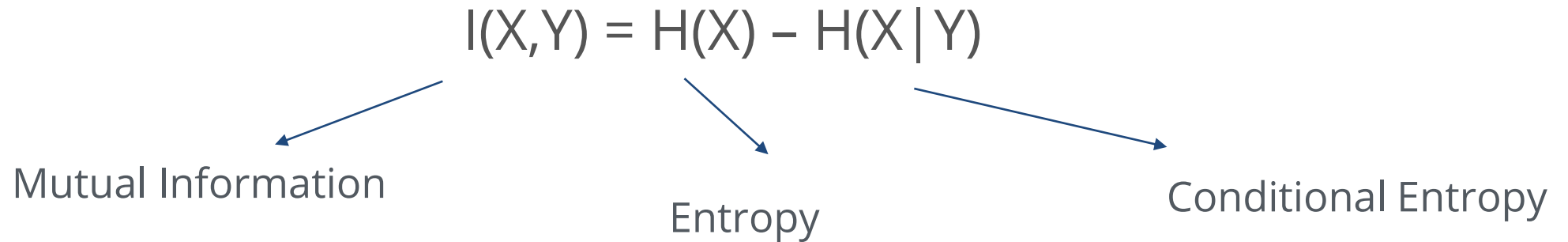
*A quantity called mutual information measures the amount of information one can obtain from one random variable given another.*



# Mutual Information

Mutual information is a way to calculate statistical dependence between two variables

(it is related to information gain, which will be introduced in classification trees)



*A quantity called mutual information measures the amount of information one can obtain from one random variable given another.*

# Mutual Information

$$I(X,Y) = H(X) - H(X | Y)$$

Mutual information is symmetric and non-negative (and it is measures in bits)

Which range does it span?

In our example...

# Mutual Information

In our example...

	Feature #1	Feature #2	Feature #3	Feature #4
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2
...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7

Entropy(feature #1) =  
 $-\sum(p \cdot \log_2(p))$

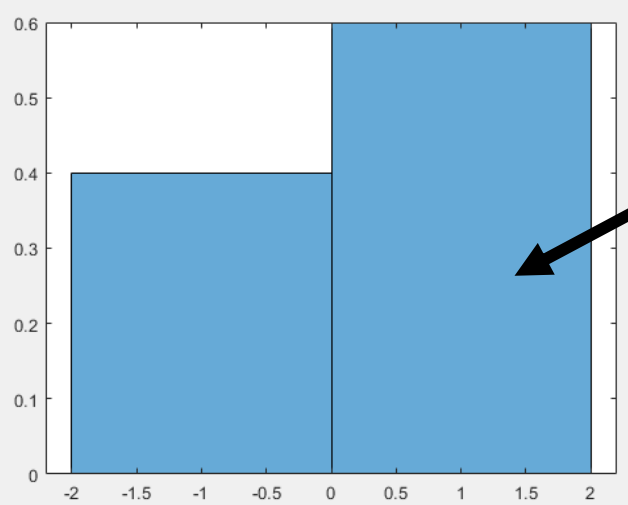
where

$P([0,25 \ 0,03 \ -0,91 \ 1,20 \ -0,87]) =$   
 $[0,40 \ 0,60]$

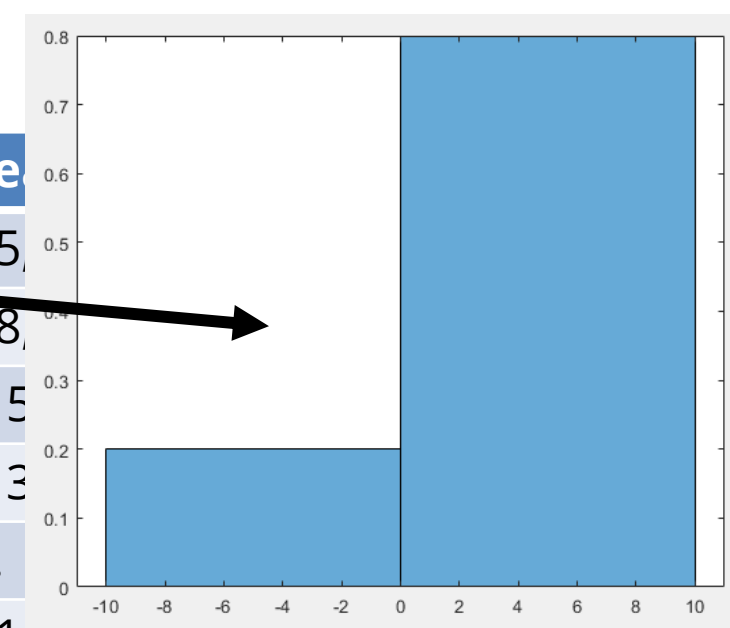
Entropy = 0,971

Conditional entropy(feature #1 | feature #2) =  
 $\sum [ p(x,y) \cdot \log_2( p(x) / p(x,y) ) ] =$   
 $\sum [ p(y) \cdot H(X|Y=y) ]$

# Mutual Information



	Feature #1	Feature #2	Feature #3	Feature #4
...	...	...	...	...
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,1
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,1
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,1
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,1
...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,1



Entropy(feature #1) =  
 $-\sum(p_i \cdot \log_2(p_i))$

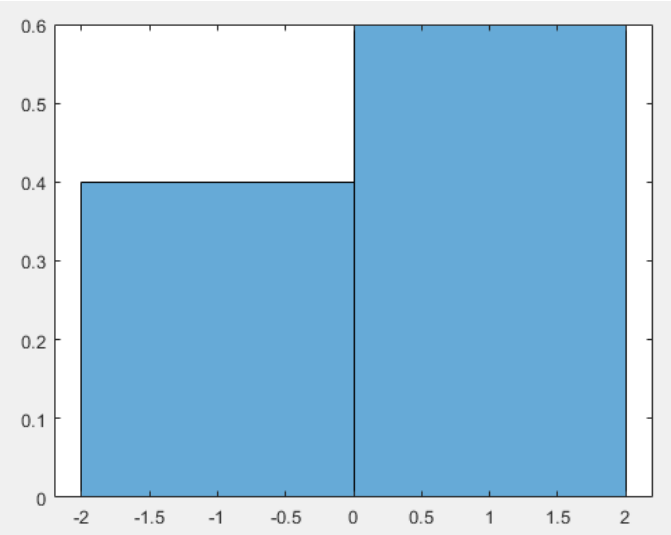
where

$P([0,25 \ 0,03 \ -0,91 \ 1,20 \ -0,87]) =$   
 $[0,40 \ 0,60]$

Entropy = 0,971

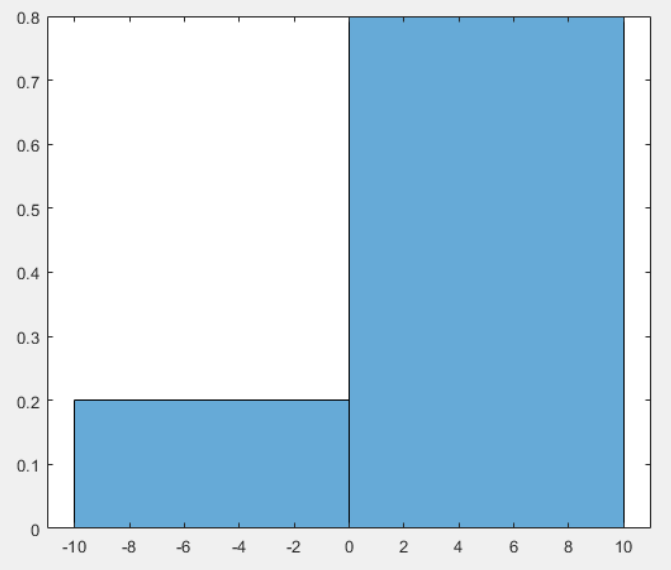
Conditional entropy(feature #1 | feature #2) =  
 $\sum [ p(x,y) \cdot \log_2( p(x) / p(x,y) ) ] =$   
 $\sum [ p(y) \cdot H(X|Y=y) ]$

# Mutual Information



Feature #1	Feature #2
+ 0,25	+ 1,47
+ 0,03	+ 1,81
- 0,91	+ 9,70
+ 1,20	- 1,71
...	...
- 0,87	+ 0,88

1 / 2	0	1
0	0	2/5
1	1/5	2/5



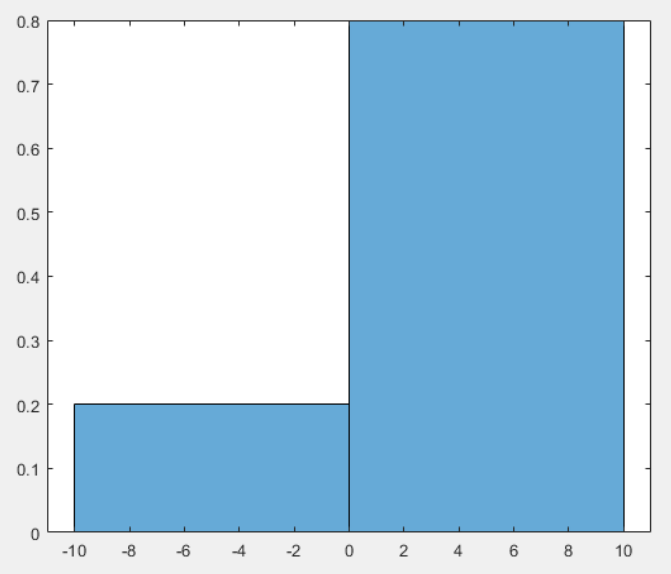
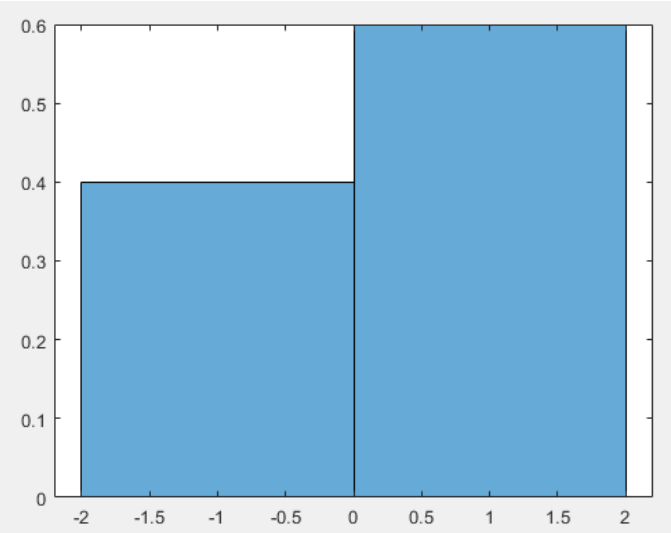
#1) =

)

1 1,20 -0,87]) =

Conditional entropy(feature #1 | feature #2) =  
sum[ p(x,y) .\* log2( p(x) / p(x,y) ) ] =  
sum[ p(x) .\* H(Y | X=x) ]

# Mutual Information



Feature #1	Feature #2
+ 0,25	+ 1,47
+ 0,03	+ 1,81
- 0,91	+ 9,70
+ 1,20	- 1,71
...	...
- 0,87	+ 0,88

1 / 2	0	1
0	0	2/5
1	1/5	2/5

#1) =

)

1 1,20 -0,87]) =

Conditional entropy(feature #1 | feature #2) =  
 $\sum [ p(x,y) \cdot \log_2( p(x) / p(x,y) ) ] =$   
 $\sum [ p(x) \cdot H(Y | X=x) ] =$

$0,4 \cdot H(Y | X=0) + 0,6 \cdot H(Y | X=1) =$   
 $0,4 \cdot H(0, 1) + 0,6 \cdot H(1/3, 2/3) =$   
 $0,6 \cdot 0,918 =$   
0,55

# Mutual Information

	Feature #1	Feature #2
Sample #1	+ 0,25	+ 1,47
Sample #2	+ 0,03	+ 1,81
Sample #3	- 0,91	+ 9,70
Sample #4	+ 1,20	- 1,71
...	...	...
Sample #n	- 0,87	+ 0,88

$$I(X,Y) = H(X) - H(X|Y)$$

Entropy(feature #1) =  
0,971

Conditional entropy(feature #1 | feature #2) =  
0,55

MI = I(feature #1, feature #2) =  
 $0,971 - 0,55 = 0,421$

# Informative / relevant features

	Feature #1	Feature #2	Feature #3	Feature #4
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2
...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7



# Informative / relevant features

	Feature #1	Feature #2	Feature #3	Feature #4	Feature #5
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7	+ 1
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2	+ 1
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4	+ 1
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2	+ 1
...	...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7	+ 1

How to capture this  
information?

# Informative / relevant features

	Feature #1	Feature #2	Feature #3	Feature #4	Feature #5
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7	+ 1
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2	+ 1
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4	+ 1
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2	+ 1
...	...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7	+ 1

How to capture this  
information?



$$\text{Var}(X) = \text{E}[(X - \mu)^2]$$

# Near-Zero Variance

To remove constant and almost-constant predictors across samples



A diagram consisting of two blue arrows. The first arrow starts from the text 'To remove constant and almost-constant predictors across samples' and points down and to the left towards the text 'Zero variance'. The second arrow starts from the same top text and points down and to the right towards the text 'Near-zero variance'.

Zero variance

Near-zero variance

# Near-Zero Variance

Usually performed following these two criteria: a feature is removed if

- (1) the fraction of unique values over the whole set of samples is low (typically  $<10\%$ )
- (1) the ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (typically, around 20).

In our example...

# Near-Zero Variance

(1) the fraction of unique values over the whole set of samples is low (typically <10%)

(2) the ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (typically, around 20).

	Feature #1	Feature #2	Feature #3	Feature #4	Feature #5
Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7	+ 1
Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2	+ 1
Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4	+ 1
Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2	+ 1
...	...	...	...	...	...
Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7	+ 1

(1)  $\frac{2}{5} = 40\%$

$\frac{1}{5} = 20\%$

(2)  $\frac{3}{5} / \frac{2}{5} = 1,5$

$5/5 / 0 > 20$

# Fisher's Discriminant Ratio / Linear Discriminant Analysis

Linear and supervised reduction of dimensionality

Aim: maximizing the separation between (among) classes

In order to formulate the optimization criterion of maximum separation between (among) classes, we define the following scattering matrices:

- within-class
- between-class

The first says how much the vectors are scattered with respect to the centre of the corresponding classes (each set of data with respect to the corresponding - belonging- class)

The second says how the centres of the classes are scattered with respect to the centre of the entire distribution (it measures how much the classes are scattered)

# Fisher's Discriminant Ratio / Linear Discriminant Analysis

Given a dataset containing  $n$  samples  $(x_1, y_1) \dots (x_n, y_n)$   
where  $x_i$  are the multidimensional patterns and  $y_i$  are the labels of the  $s$  classes,

let us assume that  $n_i$  is the number of patterns  
and  $\bar{x}_i$  the mean vector of the  $i$ -th class

then the scattering matrices are defined as follows:

*within-class*

$$\mathbf{S}_w = \sum_{i=1 \dots s} \mathbf{S}_i, \quad \mathbf{S}_i = \sum_{\mathbf{x}_j | y_j=i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)^t$$

covariance matrix  
(without normalizing for the number of  
samples)

← sample of the  $i$ -th class

*between-class*

$$\mathbf{S}_b = \sum_{i=1 \dots s} n_i \cdot (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0)(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0)^t, \quad \bar{\mathbf{x}}_0 = \frac{1}{n} \sum_{i=1 \dots s} n_i \cdot \bar{\mathbf{x}}_i$$

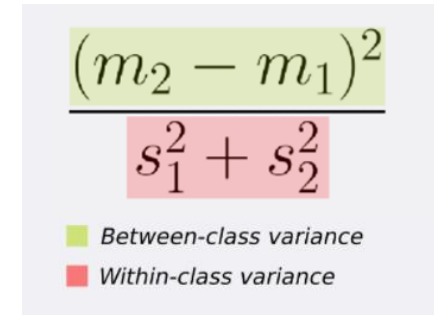
← global mean

# Fisher's Discriminant Ratio / Linear Discriminant Analysis

The criterion for the optimal solution is intuitive, as it tries to maximize the scattering between classes  $S_b$ , minimizing at the same time the within-class scattering  $S_w$  (within each class).

Thus, this means maximizing the following quantity:

$$J_1 = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b) = \sum_{i=1 \dots d} \lambda_i$$



The diagram shows the Fisher's Discriminant Ratio formula: 
$$\frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$
 The numerator  $(m_2 - m_1)^2$  is highlighted in light green. The denominator  $s_1^2 + s_2^2$  is highlighted in light red. Below the formula, a legend indicates: 

- Between-class variance (green square)
- Within-class variance (red square)

where  $tr$  is the trace of the matrix (sum of the eigenvalues).

It can be shown that in order to maximize  $J_1$ , the LDA space is defined by the eigenvectors corresponding to the first  $k$  eigenvalues of the matrix  $\mathbf{S}_w^{-1} \mathbf{S}_b$  ( $k < n$ ,  $k < s$ ,  $k < d$ ) (analogy with PCA).



Maximum value of  $k = s-1$



# Fisher's Discriminant Ratio / Linear Discriminant Analysis

In our example...

		Feature #1	Feature #2	Feature #3	Feature #4
CLASS 1	Sample #1	+ 0,25	+ 1,47	+ 0,02	- 5,7
	Sample #2	+ 0,03	+ 1,81	+ 0,02	- 8,2
	Sample #3	- 0,91	+ 9,70	+ 0,01	+ 5,4
CLASS 2	Sample #4	+ 1,20	- 1,71	+ 0,01	+ 3,2
	...	...	...	...	...
	Sample #n	- 0,87	+ 0,88	+ 0,02	- 1,7

$$\frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

■ Between-class variance  
■ Within-class variance

$$\begin{aligned} m_1 &= -0,21 \\ m_2 &= 0,17 \\ s_1^2 &= 0,38 \\ s_2^2 &= 2,14 \end{aligned}$$

$$0,057$$

$$0,898$$

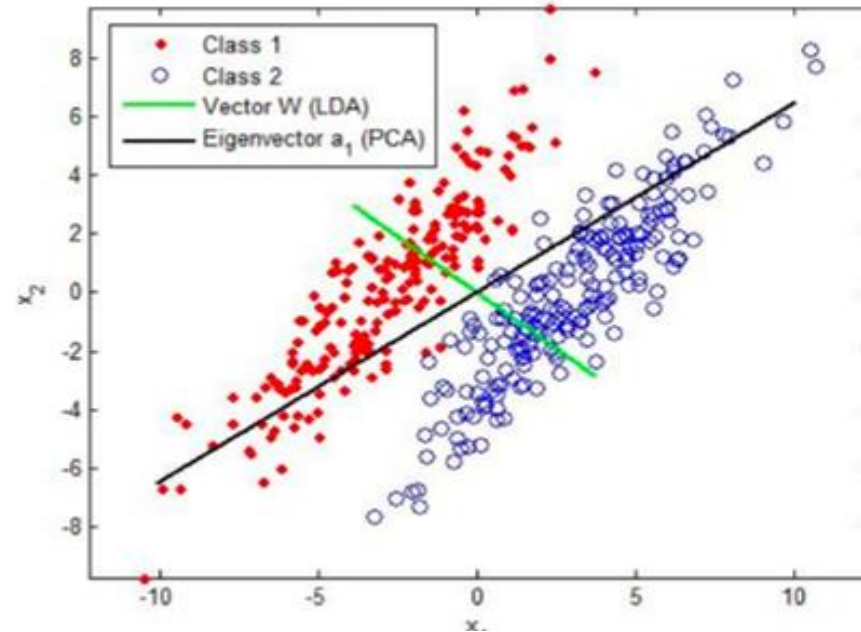
$$0,033$$

$$0,199$$

# Linear Discriminant Analysis vs. Principal Components Analysis

Dimensionality reduction from  $d = 2$  to  $k = 1$

Linear mapping are performed in both cases, but the solution is different:



$$\frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

■ Between-class variance  
■ Within-class variance

The black line identifies the PCA solution, that is, the hyperplane on which the projected data samples (independently from their belonging class) preserve the maximum information.

The green line identifies the LDA solution, that is, the hyperplane on which the projected data samples allow to best discriminating the two classes.

While PCA privileges the dimensions that best represent the distribution of the data samples, LDA privileges the dimensions that best discriminate them in the two considered classes.