Authentication

Prof. Dr. Helene Dörksen

helene.doerksen@th-owl.de

Learned before

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

Supervised Learning:

1. Bayes' decision rule (based on probability distributions)

2. Classification rules (for one feature) based on:

- **likelihood ratio test** (errors are of equal importance)
- **loss matrix** and **generalised likelihood ratio test** (errors are of non equal importance)

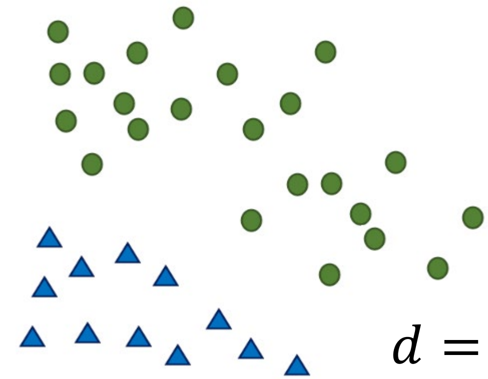
3. Classification rules for more general considerations (e.g. more than one features):

- **Decision boundaries** obtained from **discriminant functions**

Lecture 6:

Feature Extraction and Feature Selection

Reducing Dimensionality



$d = 2$
$N = 35$

Complexity depends on:

- number of features, $\boldsymbol{d}$ (which result in a $\boldsymbol{d}$-dimensional feature space)
- number of (training) data samples, $\boldsymbol{N}$

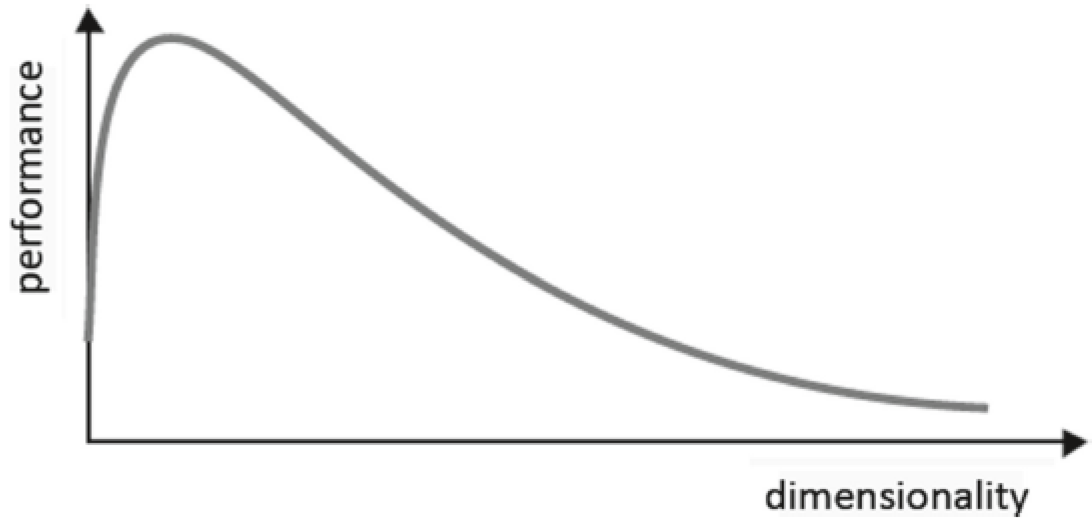
feature selection = choosing the most informative subset of features, and removing as many irrelevant and redundant features as possible

feature extraction = combining the existing feature set into a smaller set of new, more informative features

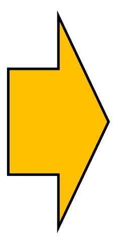
vs.

"**curse of dimensionality**"

Reducing Dimensionality



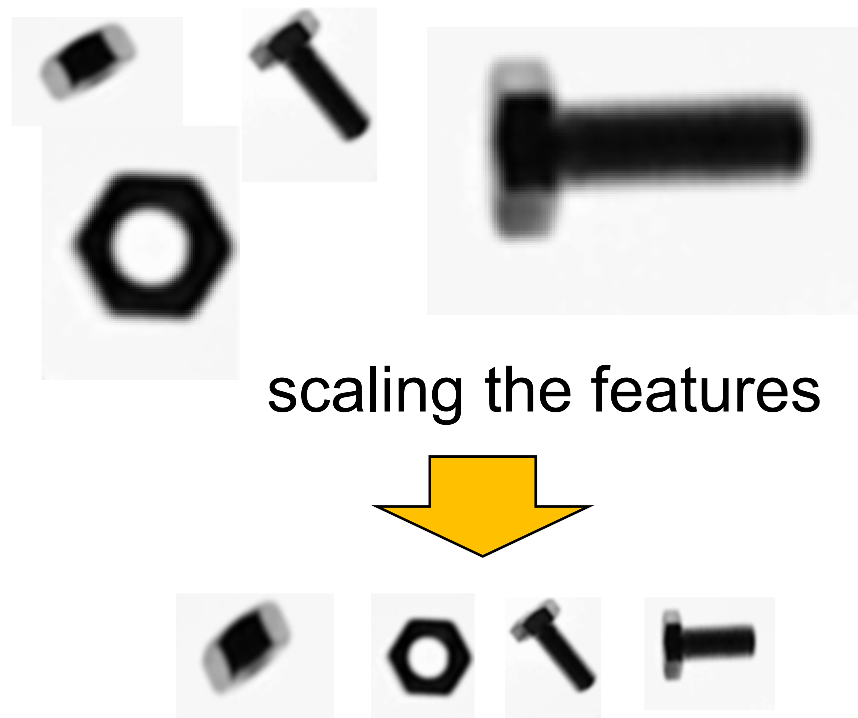
schematic representation of performance of a classifier in terms of the dimensionality of the feature space



Principal Components Analysis (**PCA**)
Linear Discriminant Analysis (**LDA**)

Preprocessing

- scale the features to comparable dynamic ranges (i.e., normalization)
- remove outliers
- treat incomplete data

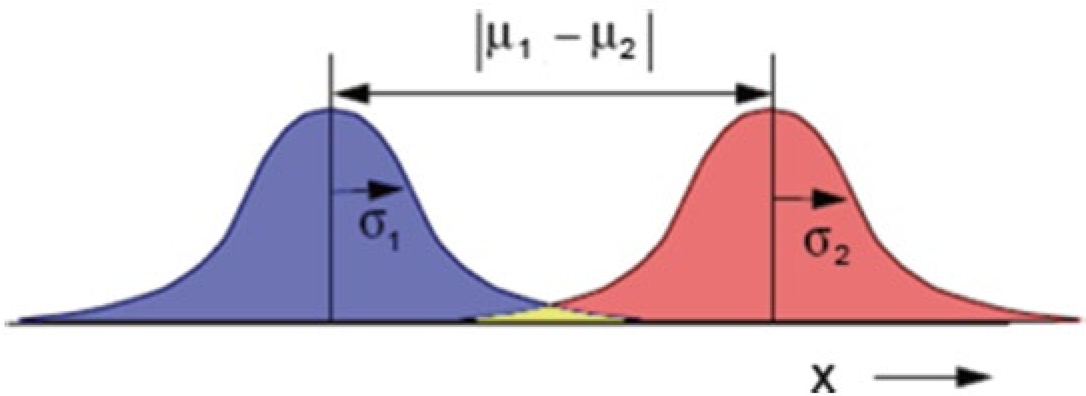


normalization: $$x' = \frac{(x - \bar{x})}{\sigma}$$

softmax scaling: $$x' = \frac{1}{1 + \exp(-x')}$$

Feature Selection

Intraclass distance & **Interclass** distance



$$\text{FDR} = \frac{\mu_1{}^2 - \mu_2{}^2}{\sigma_1{}^2 + \sigma_2{}^2}$$

Feature Selection

Subset Selection

There are $\mathbf{2^{\boldsymbol{d}}}$ possible subsets of $\boldsymbol{d}$ features (4,8,16,…,1014,…, 1048576)

Two main approaches:
- forward selection
- backward selection

not guarantee finding optimal subset

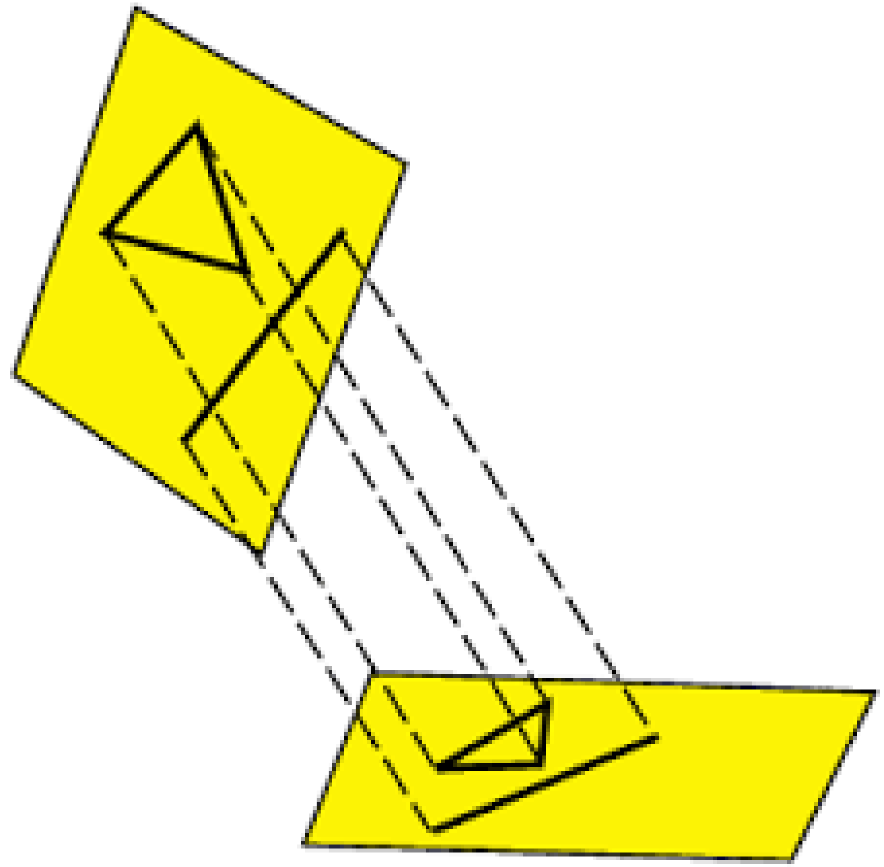
E.g. metrics for selection:
- FDR
- misclassification error

Feature Extraction

Projections Methods

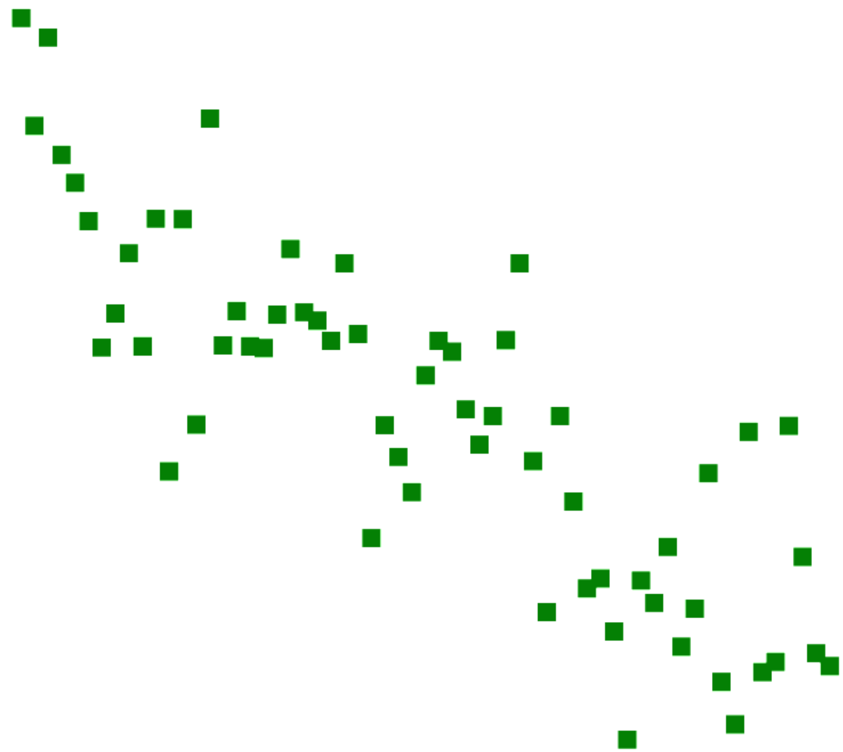
e.g. shape of data:

- scales, rotates
- simplifies
- disappears



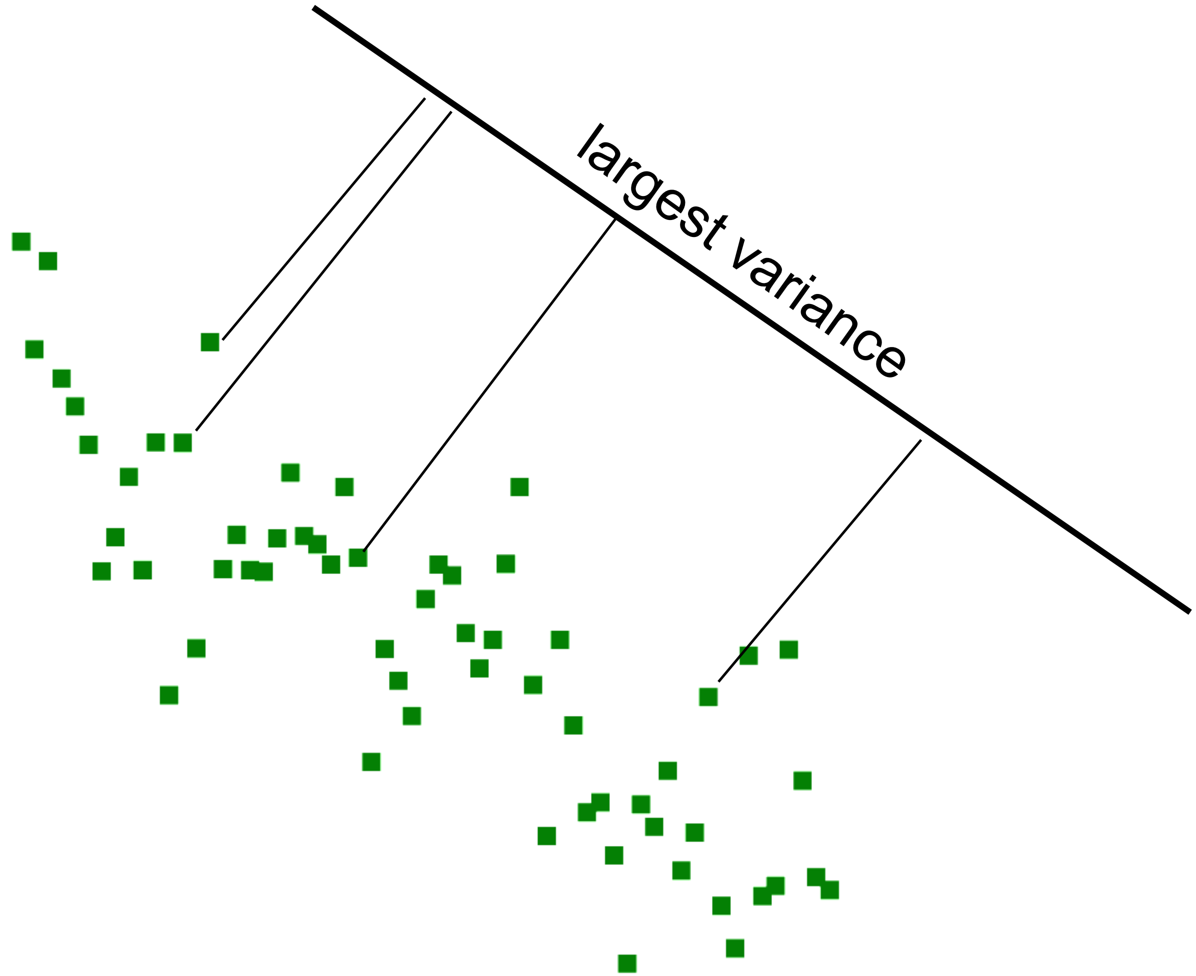
Feature Extraction

Projections Methods



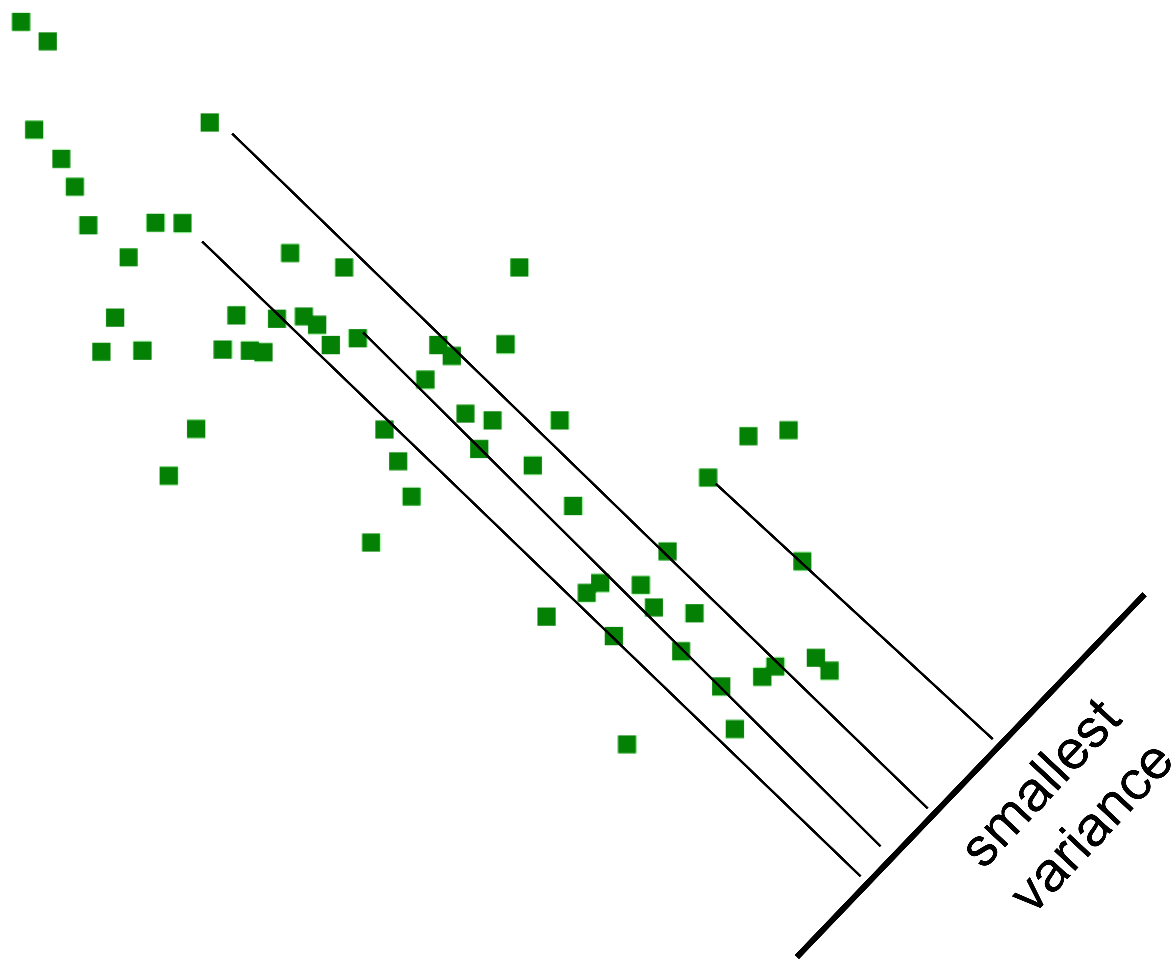
Feature Extraction

Projections Methods



Feature Extraction

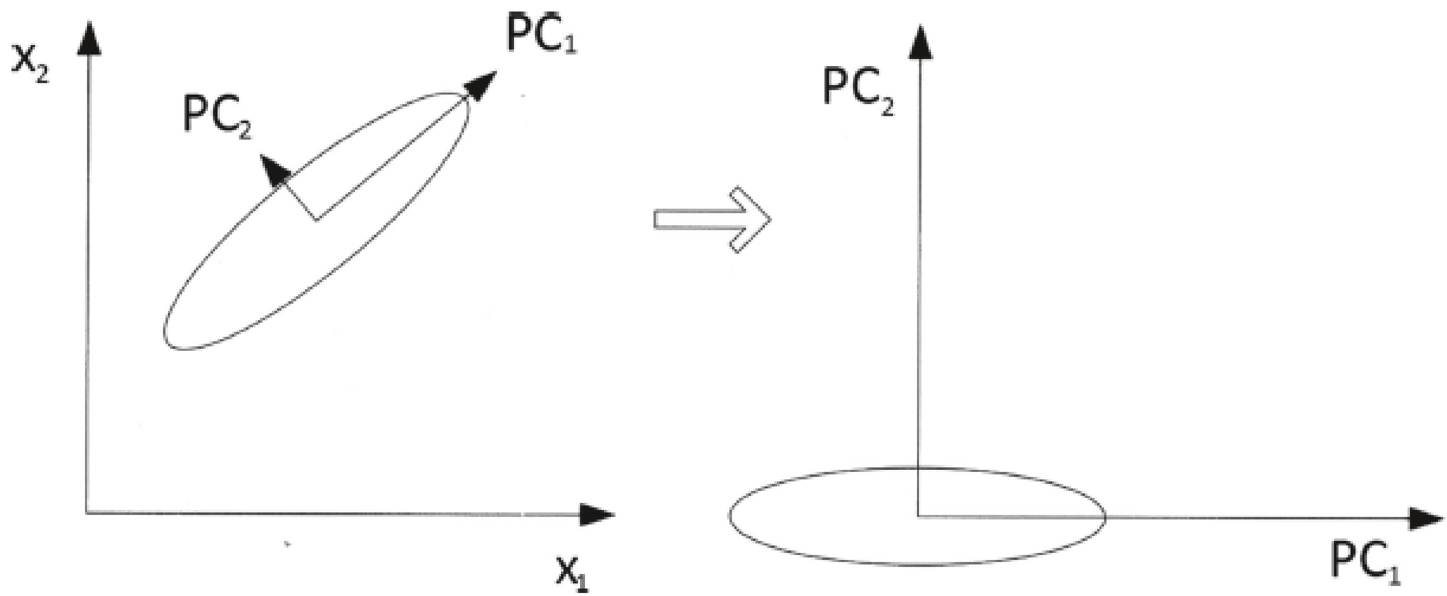
Projections Methods



Feature Extraction

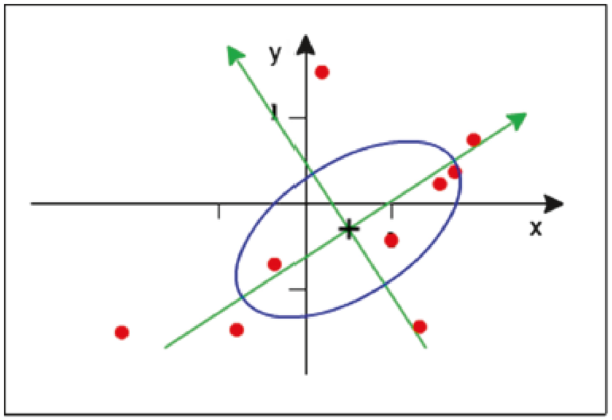
Principal Component Analysis

also known as the Karhunen–Loe`ve (KL) transform



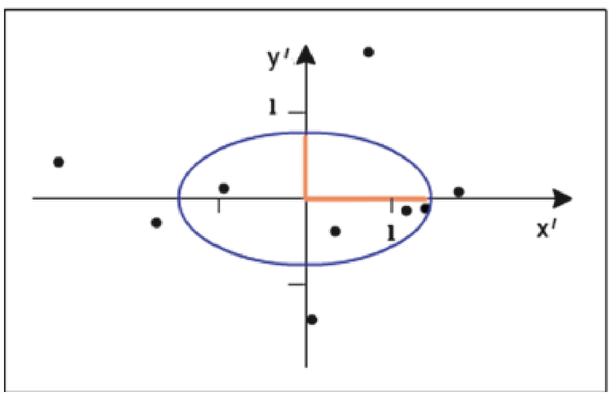
Feature Extraction

Principal Component Analysis



Covariance matrix

$$\left[\begin{array}{c|c} 1.67 & 0.70 \\ \hline 0.70 & 1.05 \end{array}\right]$$



Diagonalized covariance matrix

$$\left[\begin{array}{c|c} 2.13 & 0 \\ \hline 0 & 0.59 \end{array}\right]$$

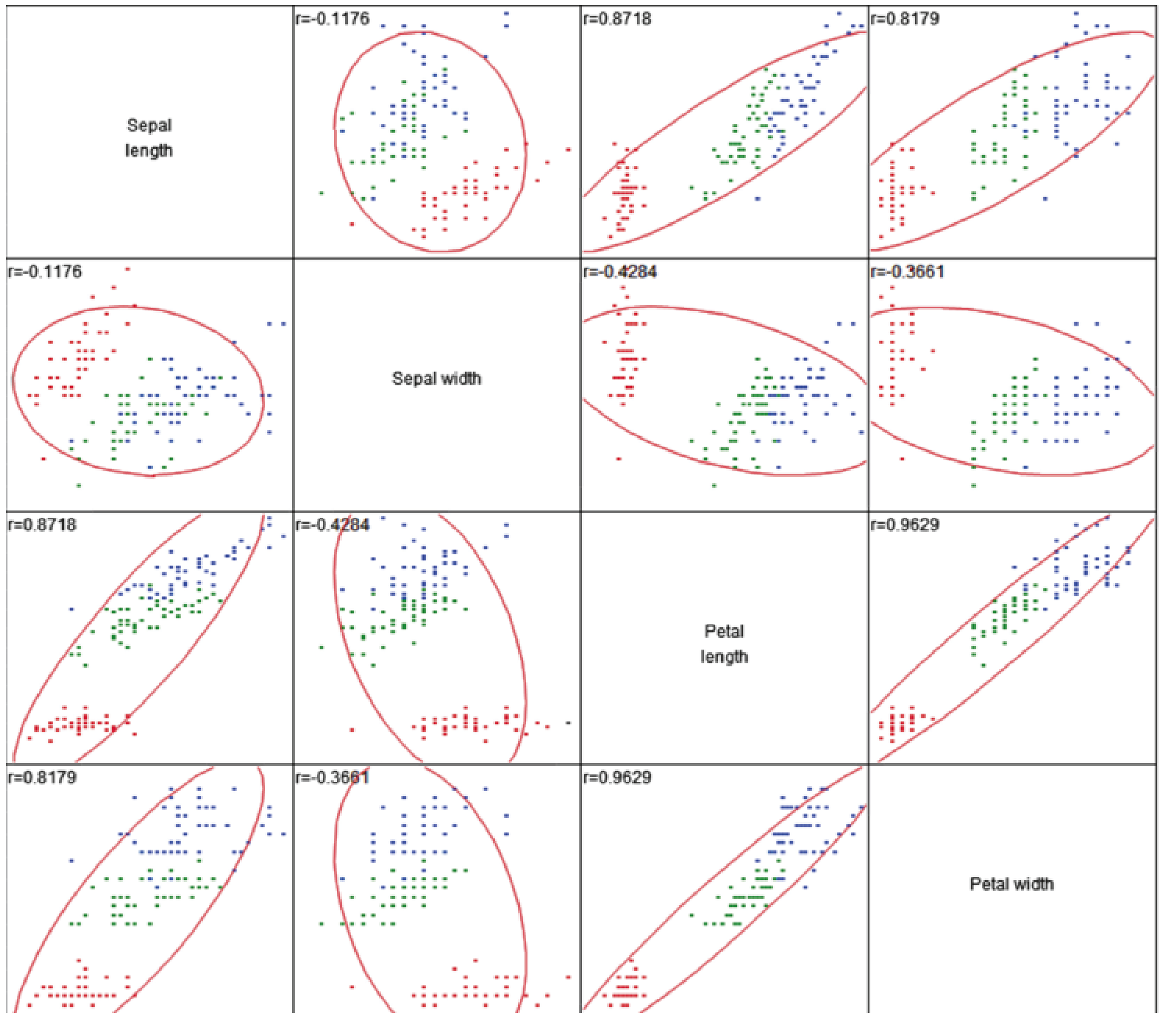
solve a system:

$$\mathbf{\Sigma x} = \lambda \mathbf{x}$$

$$(\mathbf{\Sigma} - \lambda \mathbf{I})\mathbf{x} = 0$$

Feature Extraction

Principal Component Analysis



Feature Extraction

Principal Component Analysis

Number	Eigenvalue	Percent
1	2.9185	72.962
2	0.9140	22.851
3	0.1468	3.669
4	0.0207	0.518

solve a system:

$$\Sigma \mathbf{x} = \lambda \mathbf{x}$$

$$(\Sigma - \lambda \mathbf{I})\mathbf{x} = 0$$

Cum Percent
72.962
95.813
99.482
100.000

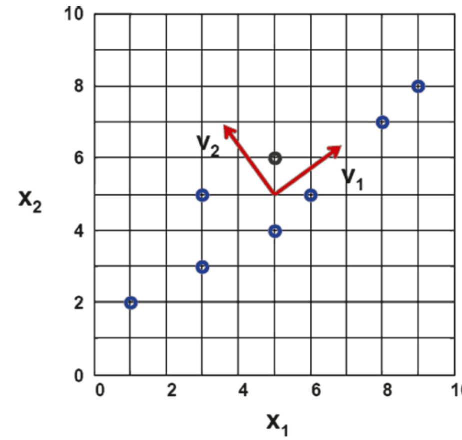
threshold $\sim \mathbf{95,8}\%$

Feature Extraction

Principal Component Analysis

Example:

$$\mathbf{x} = (x_1, x_2) = \{(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)\}$$



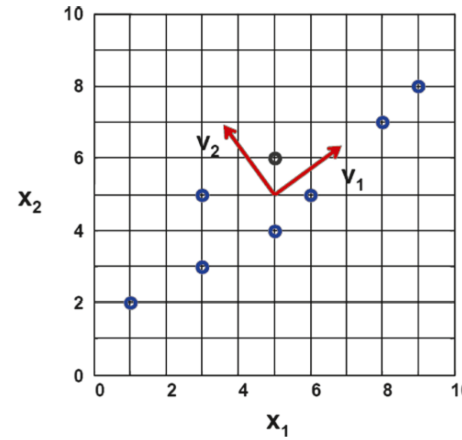
Feature Extraction

Principal Component Analysis

Example:

$$\mathbf{x} = (x_1, x_2) = \{(1, 2), (3, 3), (3, 5), (5, 4), (5, 6), (6, 5), (8, 7), (9, 8)\}$$

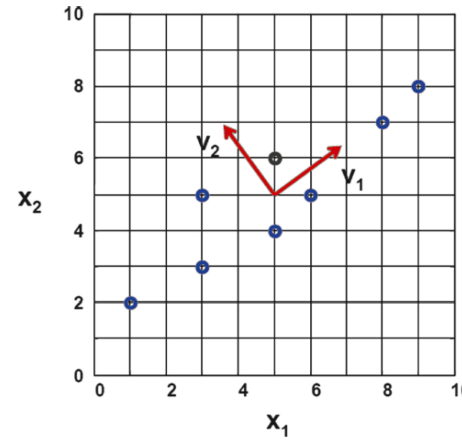
The covariance matrix, $\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$



Feature Extraction

Principal Component Analysis

Example:



$$\mathbf{x} = (x_1, x_2) = \{(1, 2), (3, 3), (3, 5), (5, 4), (5, 6), (6, 5), (8, 7), (9, 8)\}$$

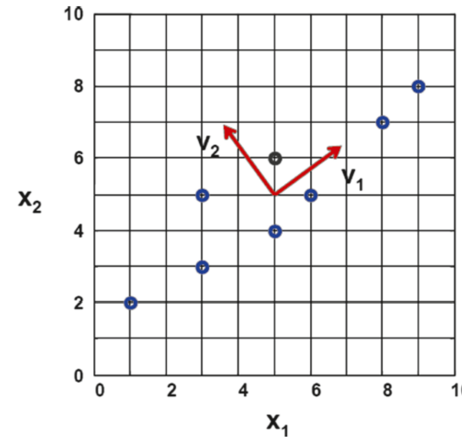
The covariance matrix, $\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$

The eigenvalues are the zeros of the characteristic equation

$$\sum v = \lambda v \Rightarrow |\Sigma - \lambda I| = 0 \Rightarrow \begin{vmatrix} 6.25 - \lambda 4.25 \\ 4.253.5 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda_1 = 9.34; \lambda_2 = 0.41$$

Feature Extraction

Principal Component Analysis



Example:

$\mathbf{x} = (x_1, x_2) = \{(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)\}$

The covariance matrix, $\Sigma = \begin{pmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{pmatrix}$

The eigenvalues are the zeros of the characteristic equation

$$\sum v = \lambda v \Rightarrow |\Sigma - \lambda I| = 0 \Rightarrow \begin{vmatrix} 6.25 - \lambda 4.25 \\ 4.253.5 - \lambda \end{vmatrix} = 0 \Rightarrow \lambda_1 = 9.34; \lambda_2 = 0.41$$

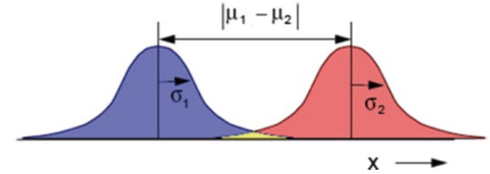
And the eigenvectors are the solutions of the system

$$\begin{bmatrix} 6.254.25 \\ 4.253.5 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$$

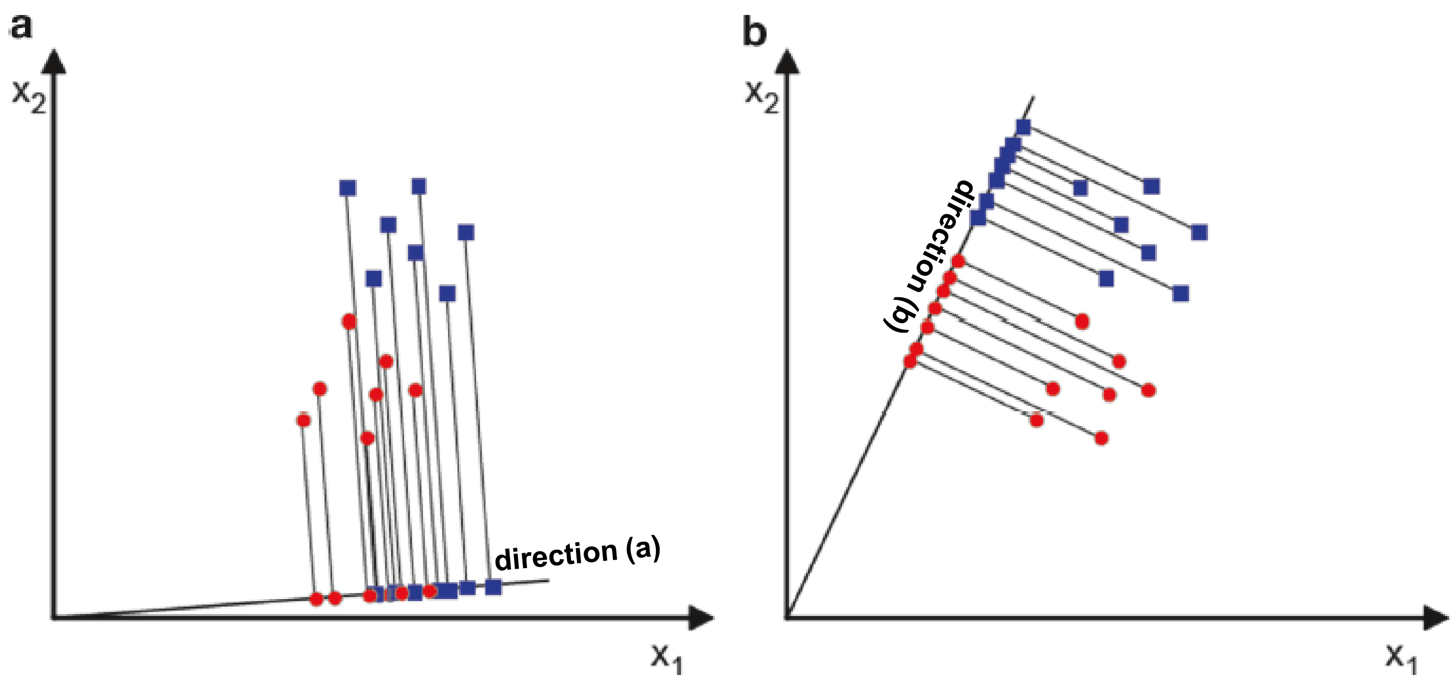
$$\begin{bmatrix} 6.254.25 \\ 4.253.5 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 v_{21} \\ \lambda_2 v_{22} \end{bmatrix} \Rightarrow \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix}$$

Feature Extraction

Linear Discriminant Analysis



$$\mathrm{FDR} = \frac{\mu_1{}^2 - \mu_2{}^2}{\sigma_1{}^2 + \sigma_2{}^2}$$

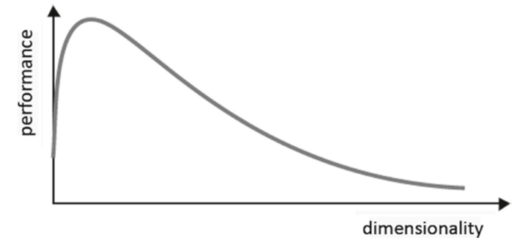


FDA direction (a) < FDA direction(b)

Find direction with **maximal FDA**

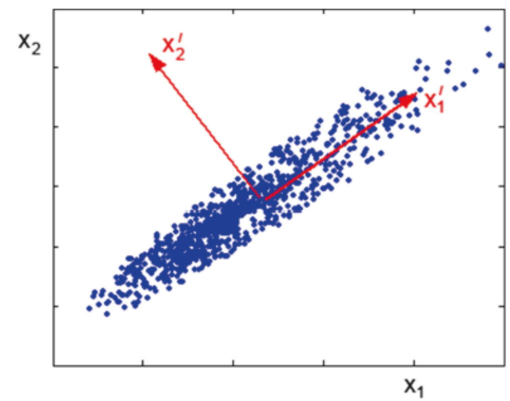
Summary

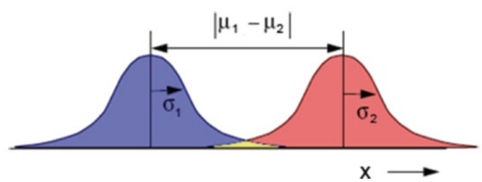
Feature Extraction and Feature Selection



Most important topics in the field:

- Dimensionality reduction
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)





$$\mathrm{FDR} = \frac{{\mu_1}^2 - {\mu_2}^2}{{\sigma_1}^2 + {\sigma_2}^2}$$

Homework: Exercises and Labs

for the next week prepare practical exercises and labs from
Exercises Lec 6 (you will find it in the donwload area)