Classification and Pattern Recognition

Slides 3rd Year Signal and Image Processing - Master

Jean-Yves Tourneret[(1)]

(1) University of Toulouse, ENSEEIHT-IRIT-TéSA
Thème 1 : Analyse et Synthèse de l'Information
jyt@n7.fr

October 2018

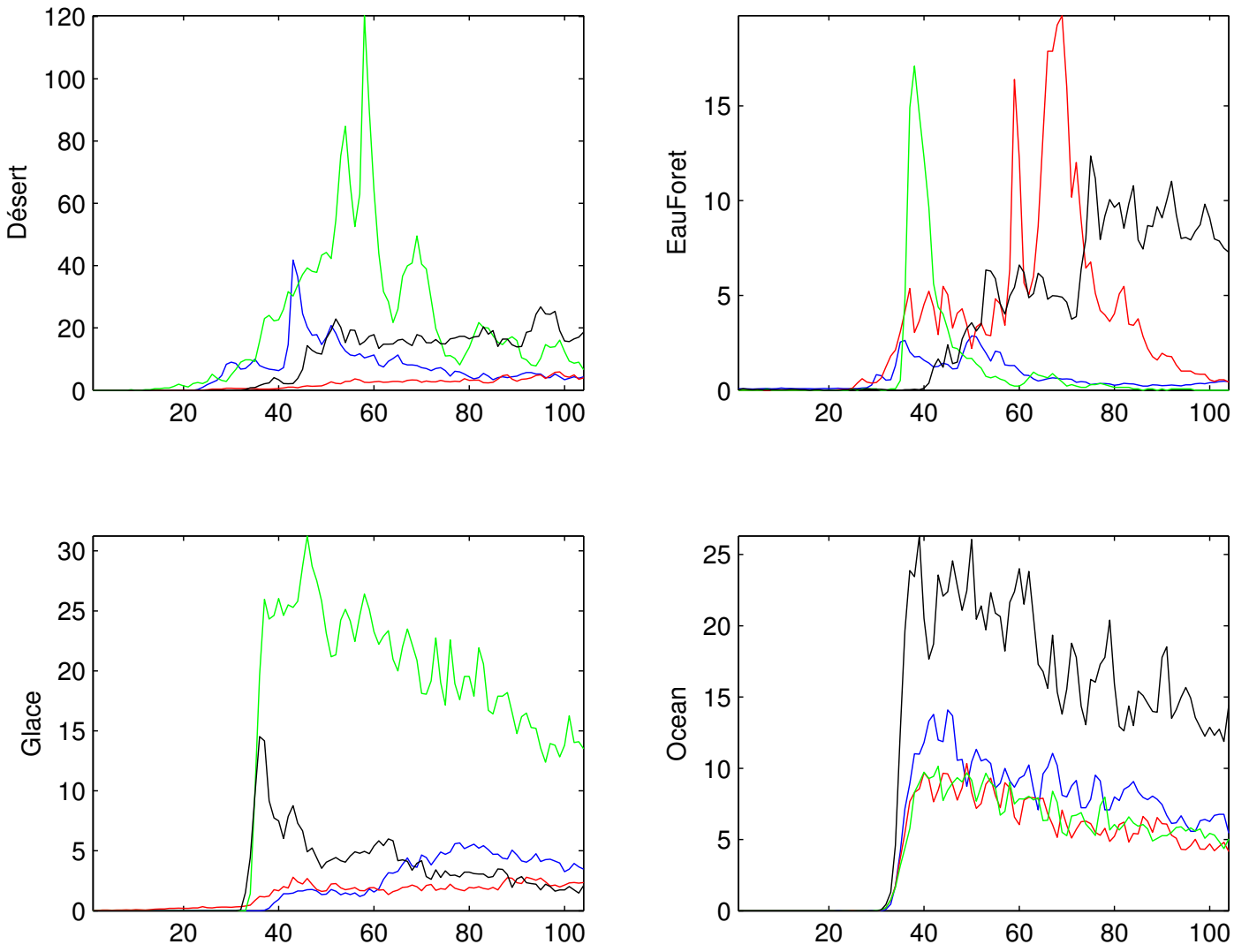
Summary

Summary

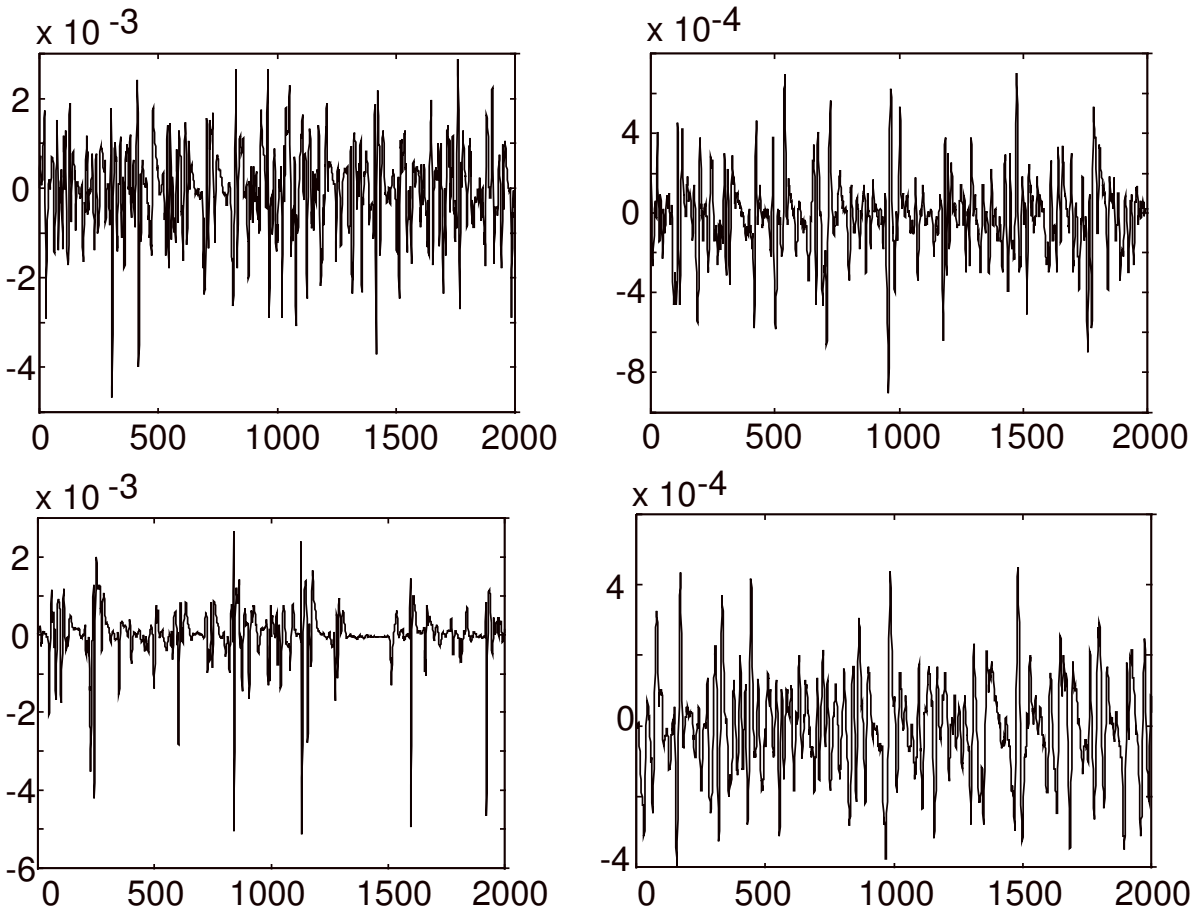
References

- R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd edition, Wiley, 2000.
- S. Theodoridis and K. Koutroumbas, Pattern Recognition, 4th edition, Academic Press, 2008.
- A. Jain, R. Duin and J. Mao, Statistical Pattern Recognition: A Review, IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, Jan. 2000.

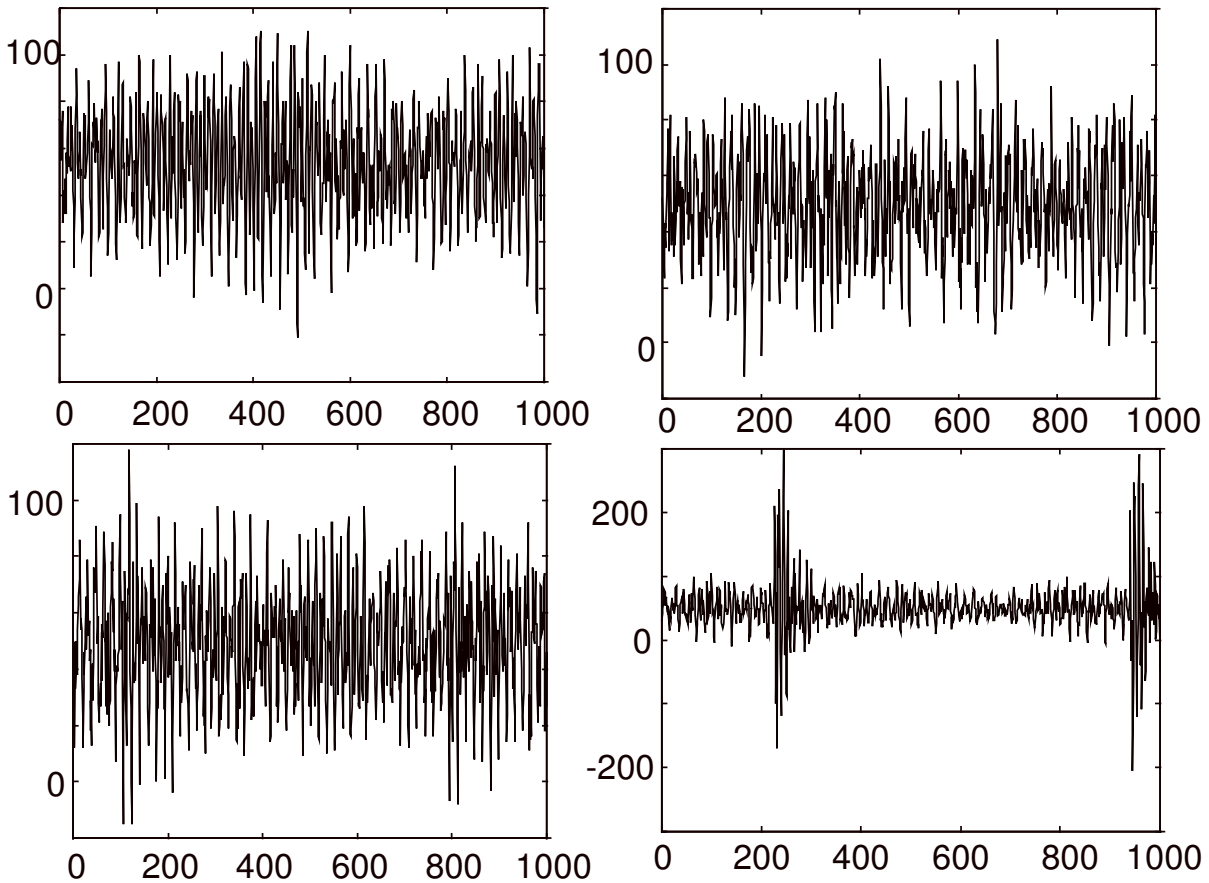
Example 1 : Altimetric Signals



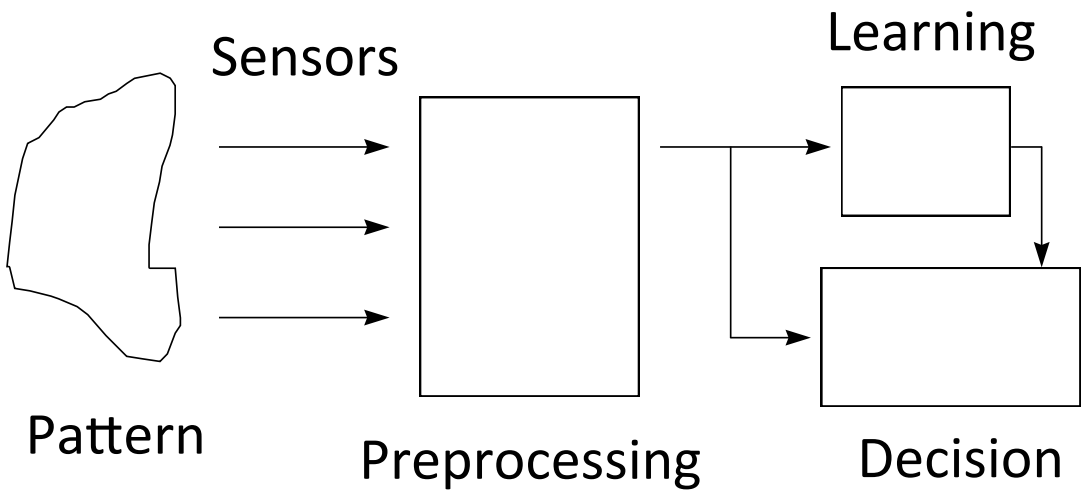
Example 2 : EMG Signals



Example 3 : Accelerometric Signals



Classification Model



Summary

Modeling and Representation

Stationary Signals

- Linear models (MA, AR, ARMA)
- Non-linear models (Volterra time-series, bilinear models, ...)
- Features (Reflexion coefficients, cepstral coefficients,...)
- ...

Non-stationary signals

- Wavelet analysis
- Time-frequency distributions
- ...

Summary

PCA of samples

First step: define a norm for $\boldsymbol{x} \in \mathbb{R}^p$

$$\|\boldsymbol{x}\|_{\boldsymbol{M}}^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{M}}^2 = \boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}$$

where $\boldsymbol{M}$ is a symmetric positive definite matrix of size $p \times p$

- $\boldsymbol{M} = I_p$

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{p} [\boldsymbol{x}(j) - \boldsymbol{y}(j)]^2$$

- $\boldsymbol{M} = \text{diag}\left(\frac{1}{\sigma_1^2}, ..., \frac{1}{\sigma_p^2}\right)$

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{p} [\boldsymbol{x}^*(j) - \boldsymbol{y}^*(j)]^2$$

where $\boldsymbol{x}^*(j) = \frac{x(j) - m(j)}{\sigma_j}$ (centered and reduced data). In what follows, data are centered, reduced and $\boldsymbol{M} = I_p$.

PCA of samples

Second step: minimize the difference between the samples and their projections

$$\text{Minimize } I_q = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \widetilde{\boldsymbol{x}}_i\|^2 \Leftrightarrow \text{Maximize } J_q = \frac{1}{n}\sum_{i=1}^{n} \|\widetilde{\boldsymbol{x}}_i\|^2$$

Property: solutions can be found recursively with $J(\boldsymbol{u}) = u^t \boldsymbol{T} \boldsymbol{u}$

- Determination of principal components
- Number of principal components
  $\boldsymbol{T}$ of size $p \times p$ invertible $\Longrightarrow$ p principal components
- How to choose the number of principal components?

$$I_q = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \left[1 - \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}\right]$$

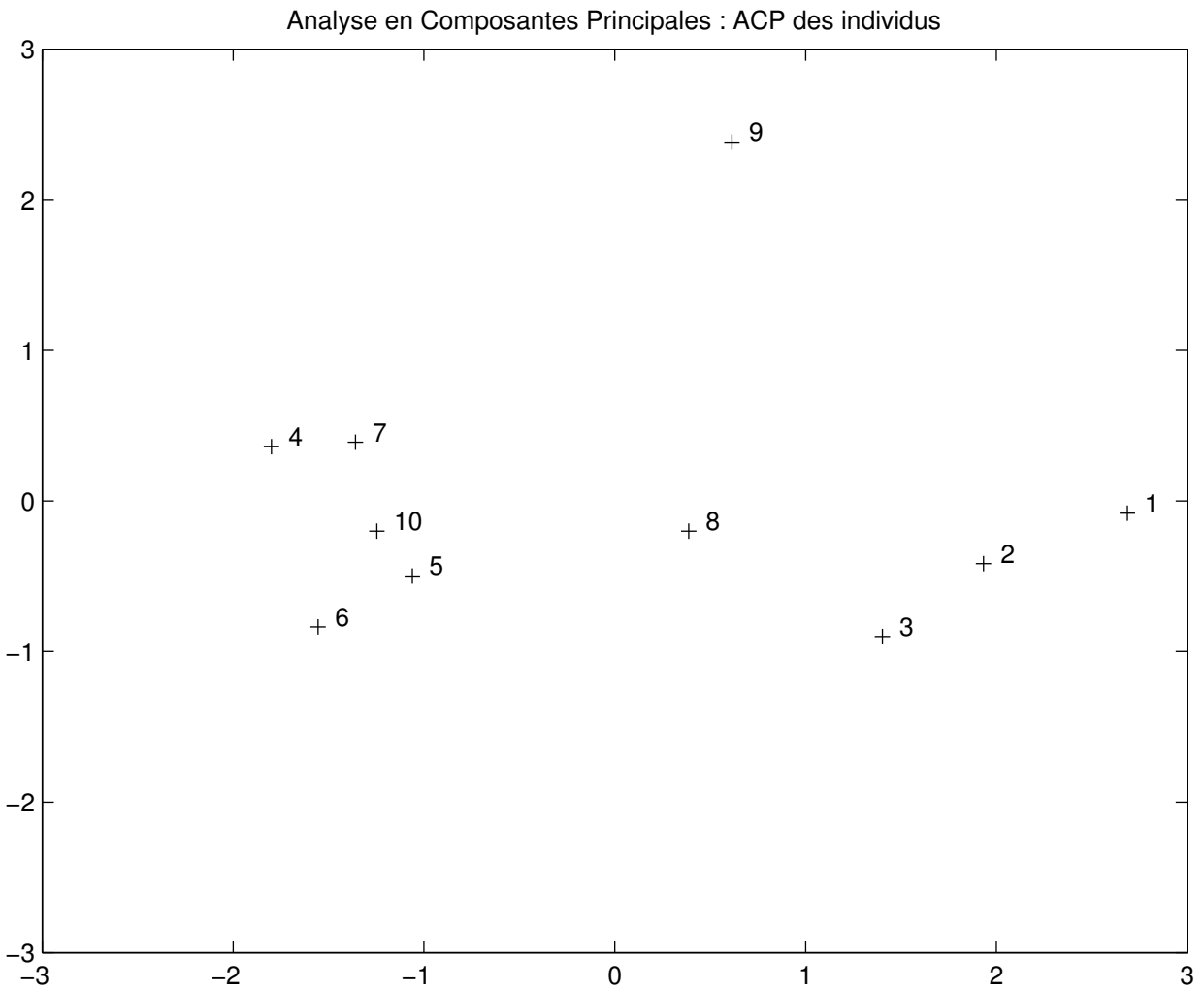
PCA of variables

- Number of principal components
  Covariance matrix of size $N \times N$ with p principal components
- Normalization
$$v'_j(i) = \frac{x_i(j) - m(j)}{\sqrt{n}\sigma(j)}$$
- Interpretation
  Representation in the correlation circle
  Correlation between variables
$$\|\boldsymbol{v}'_k - \boldsymbol{v}'_j\|^2 = 2(1 - r_{jk})$$

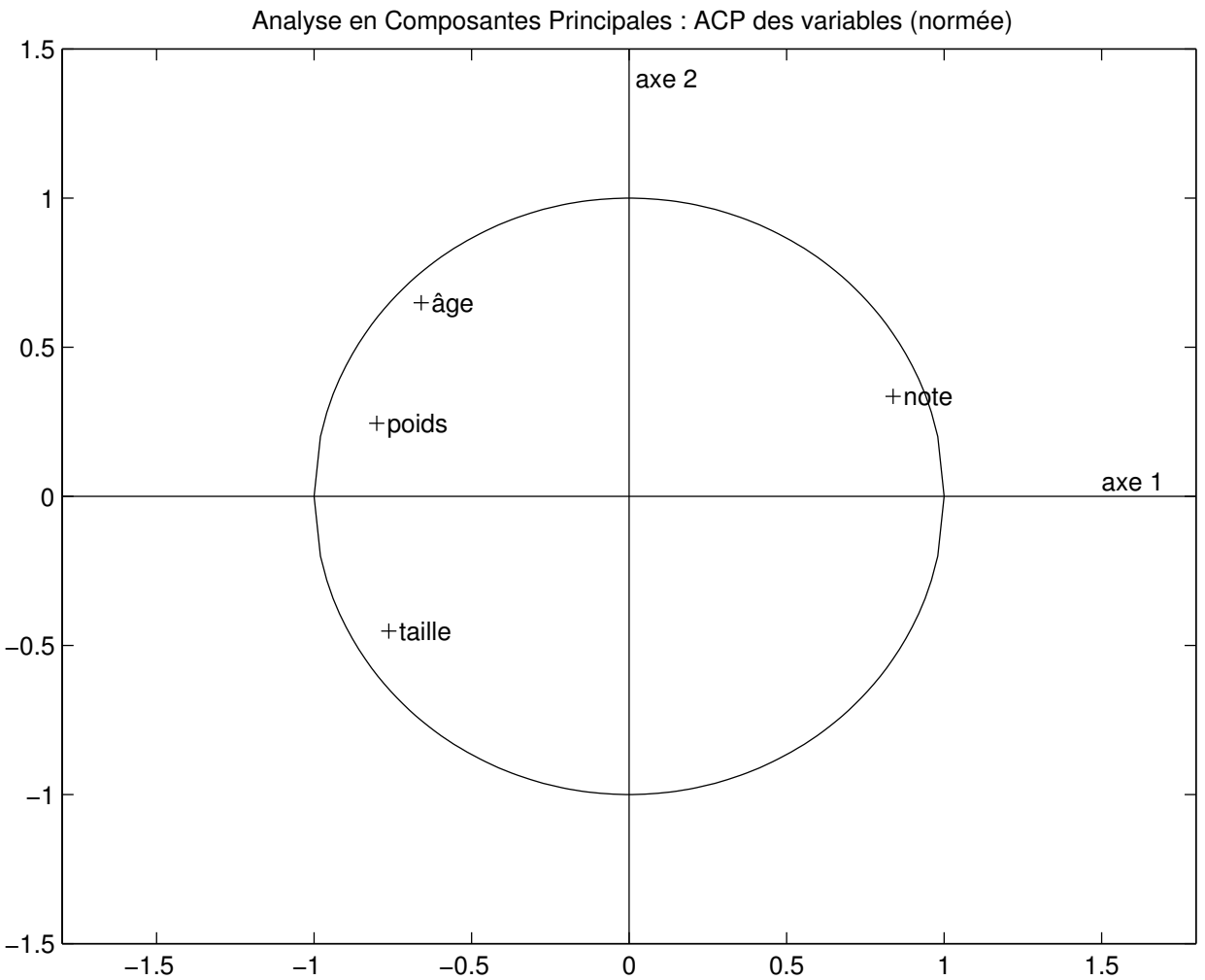
Toy Example

	Weight	Size	Age	Grade		Weight	Size	Age	Grade
$\boldsymbol{x}_1$	45	1.50	13	14	$\boldsymbol{x}_6$	60	1.70	14	7
$\boldsymbol{x}_2$	50	1.60	13	16	$\boldsymbol{x}_7$	70	1.60	14	8
$\boldsymbol{x}_3$	50	1.65	13	15	$\boldsymbol{x}_8$	65	1.60	13	13
$\boldsymbol{x}_4$	60	1.70	15	9	$\boldsymbol{x}_9$	60	1.55	15	17
$\boldsymbol{x}_5$	60	1.70	14	10	$\boldsymbol{x}_{10}$	65	1.70	14	11

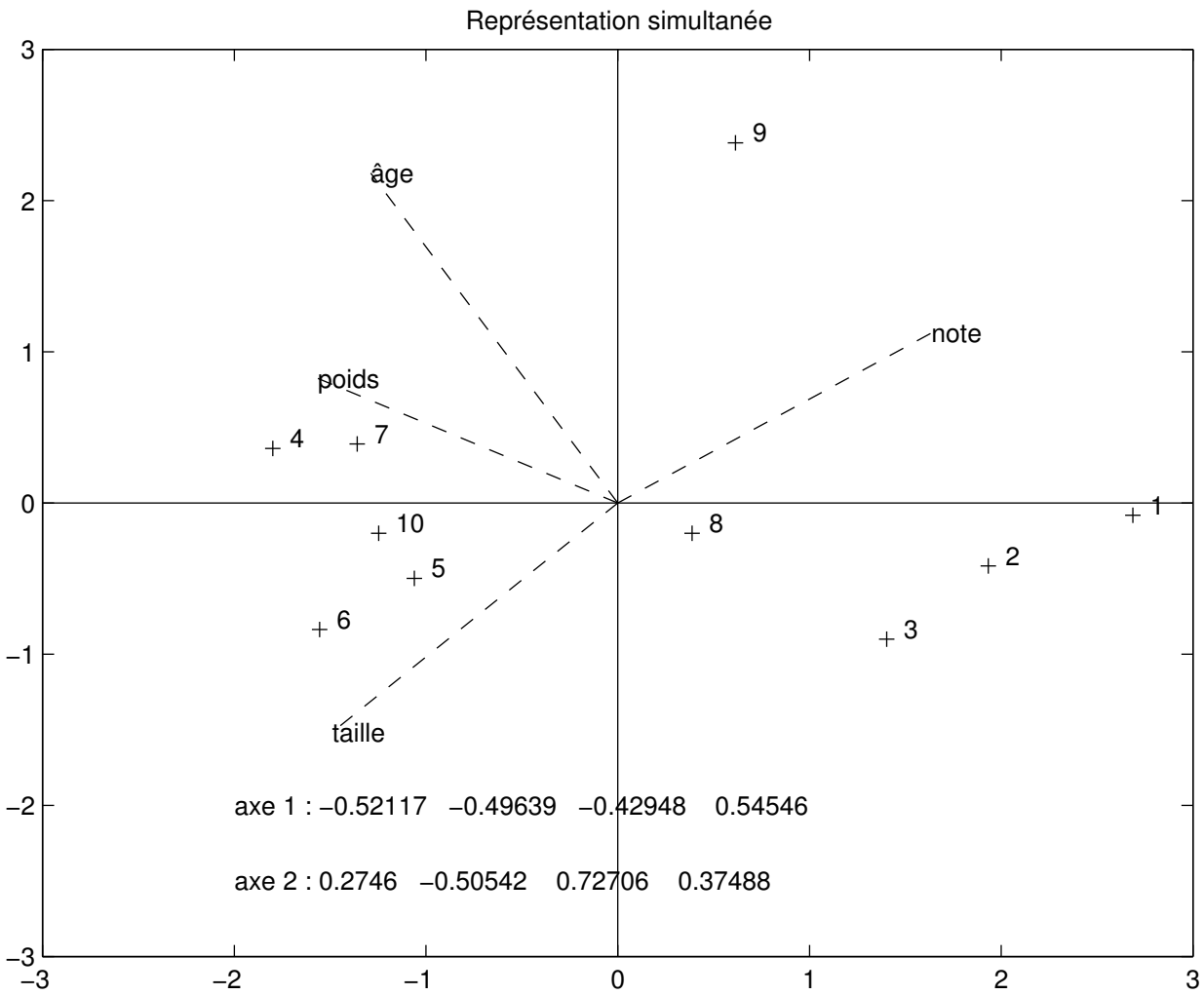
PCA of samples



PCA of variables

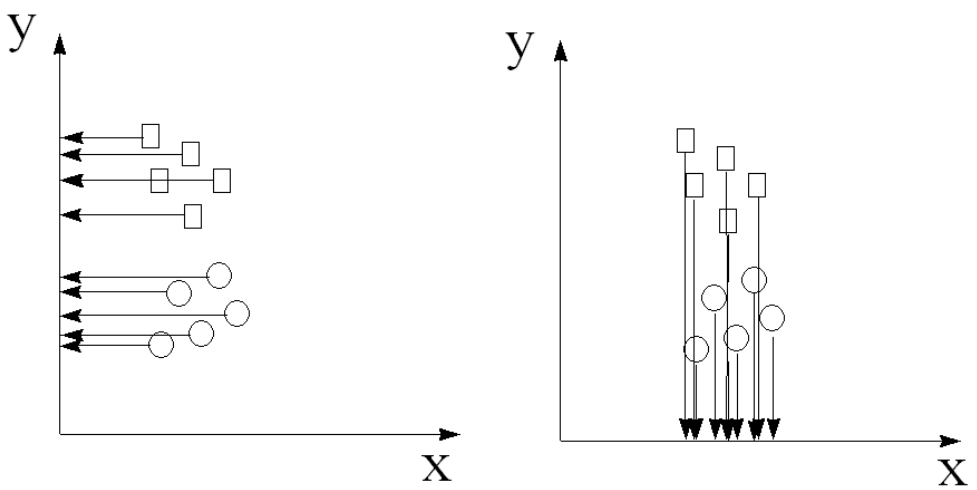


Joint PCA


Représentation simultanée
âge
poids
note
taille
9
4
7
10
8
1
5
2
6
3
axe 1 : −0.52117 −0.49639 −0.42948 0.54546
axe 2 : 0.2746 −0.50542 0.72706 0.37488
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Summary

Fisher criterion



Fisher criterion

Two classes

$$J(\boldsymbol{u}) = \frac{(\widetilde{m}_1 - \widetilde{m}_2)^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2} = \frac{\boldsymbol{u}^t \boldsymbol{B} \boldsymbol{u}}{\boldsymbol{u}^t \boldsymbol{S} \boldsymbol{u}}$$

where $\widetilde{m}_i = \boldsymbol{u}^t \boldsymbol{m}_i$, $\widetilde{x} = \boldsymbol{u}^t \boldsymbol{x}$ and $\widetilde{s}_i^2 = \sum_{\boldsymbol{x} \in W_i} (\widetilde{x} - \widetilde{m}_i)^2$.

B between-class scatter matrix

$$\boldsymbol{B} = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^t$$

S within-class scatter matrix

$$\begin{aligned} \boldsymbol{S} &= \boldsymbol{S}_1 + \boldsymbol{S}_2 \\ &= \sum_{\boldsymbol{x} \in W_1} (\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^t + \sum_{\boldsymbol{x} \in W_2} (\boldsymbol{x} - \boldsymbol{m}_2)(\boldsymbol{x} - \boldsymbol{m}_2)^t \end{aligned}$$

Fisher criterion

Two classes

- Property

$$\boldsymbol{T} = \boldsymbol{S} + \frac{n_1 n_2}{n} \boldsymbol{B}$$

since

$$\boldsymbol{T} = \sum_{x \in W_1 \cup W_2} (\boldsymbol{x} - \boldsymbol{m}) (\boldsymbol{x} - \boldsymbol{m})^t$$

- Determination of the discriminant axis

$$\boldsymbol{u} = k \boldsymbol{S}^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

Fisher criterion

Generalization to $K \geq 2$ classes

$\boldsymbol{S}$: within-class scatter matrix

$$\boldsymbol{S} = \sum_{i=1}^{K} \boldsymbol{S}_i = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in W_i} (\boldsymbol{x} - \boldsymbol{m}_i) (\boldsymbol{x} - \boldsymbol{m}_i)^t$$

$\boldsymbol{T}$: covariance matrix

$$\boldsymbol{T} = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in W_i} (\boldsymbol{x} - \boldsymbol{m}) (\boldsymbol{x} - \boldsymbol{m})^t = \boldsymbol{S} + \sum_{i=1}^{K} n_i (\boldsymbol{m}_i - \boldsymbol{m}) (\boldsymbol{m}_i - \boldsymbol{m})^t$$

hence the following between-class scatter matrix

$$\boldsymbol{B} = \sum_{i=1}^{K} n_i (\boldsymbol{m}_i - \boldsymbol{m}) (\boldsymbol{m}_i - \boldsymbol{m})^t$$

Properties

Determination of the discriminant axes

Maximization of $I(\boldsymbol{u}) = \frac{\boldsymbol{u}^t \boldsymbol{B} \boldsymbol{u}}{\boldsymbol{u}^t \boldsymbol{T} \boldsymbol{u}}$ $\Leftrightarrow$ Maximization of $J(\boldsymbol{u}) = \frac{\boldsymbol{u}^t \boldsymbol{B} \boldsymbol{u}}{\boldsymbol{u}^t \boldsymbol{S} \boldsymbol{u}}$

Remarks

- the eigenvalues of $\boldsymbol{T}^{-1}\boldsymbol{B}$ belong to $[0, 1]$
- limit cases $\mu = 0$ and $\mu = 1$
- Number of discriminant axes: $K - 1$
- Discriminant power of variable $\#i$: $J(\boldsymbol{e}_i)$ or $I(\boldsymbol{e}_i)$

Example (Matlab)

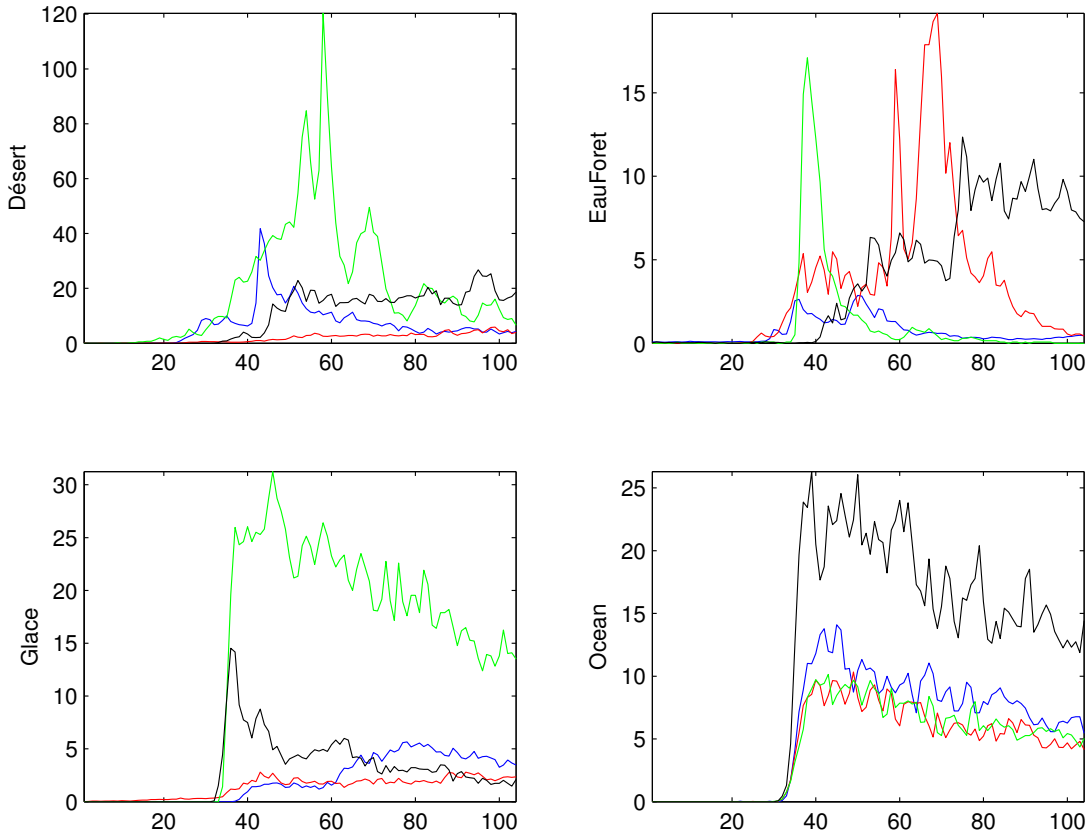
- $\boldsymbol{X} = [10; 11; -1 - 1; 1 - 1; 0 - 1; 00;$ **0 1; -1 0; -1 1; 0 2**$]$
- $\boldsymbol{T} = \text{cov}(\boldsymbol{X}) * (n-1)$ $\quad \dashrightarrow \boldsymbol{T} = \begin{pmatrix} 6.0000 & 0 \\ 0 & 9.6000 \end{pmatrix}$
- $\boldsymbol{m} = [5/6; -4/3]$
- $\boldsymbol{B} = \boldsymbol{m} * \boldsymbol{m}'$
- $[\boldsymbol{U}\ \boldsymbol{V}] = \text{eig}(\text{inv}(\boldsymbol{T}) * \boldsymbol{B})$

$$\dashrightarrow \boldsymbol{U} = \begin{pmatrix} -0.8480 & 0.7071 \\ -0.5300 & -0.7071 \end{pmatrix}, \boldsymbol{V} = \begin{pmatrix} 0 & 0 \\ 0 & 0.7222 \end{pmatrix}$$

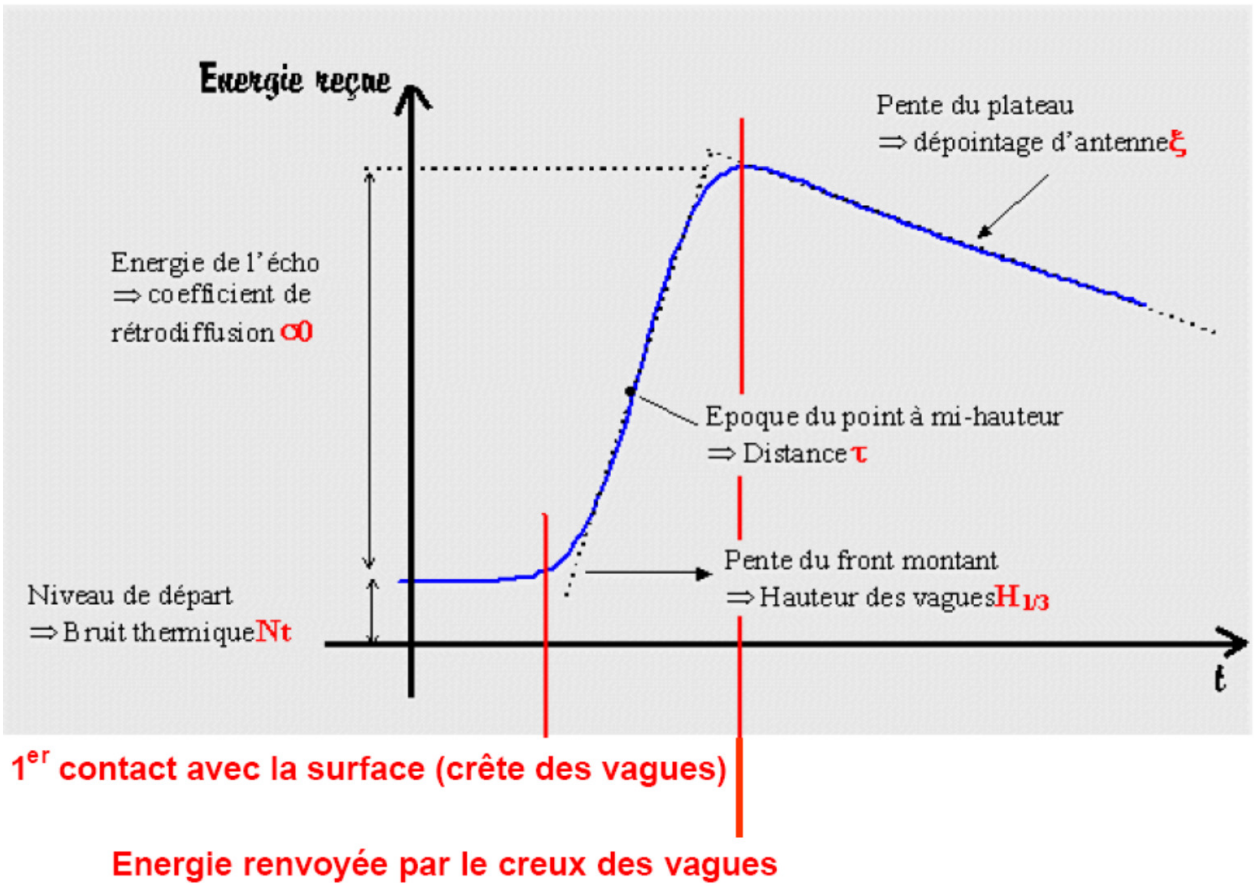
Thus, the only discriminant axis is defined by the vector $(1, -1)$:

Straight line of equation $y = -x$.

Waveforms



Analytical model



Classification strategy

Waveform-based classification

- Preprocessing: linear discriminant analysis
- Classification: Bayesian classifier, centroid distance rule, k nearest neighbors,...

Parameter-based classification

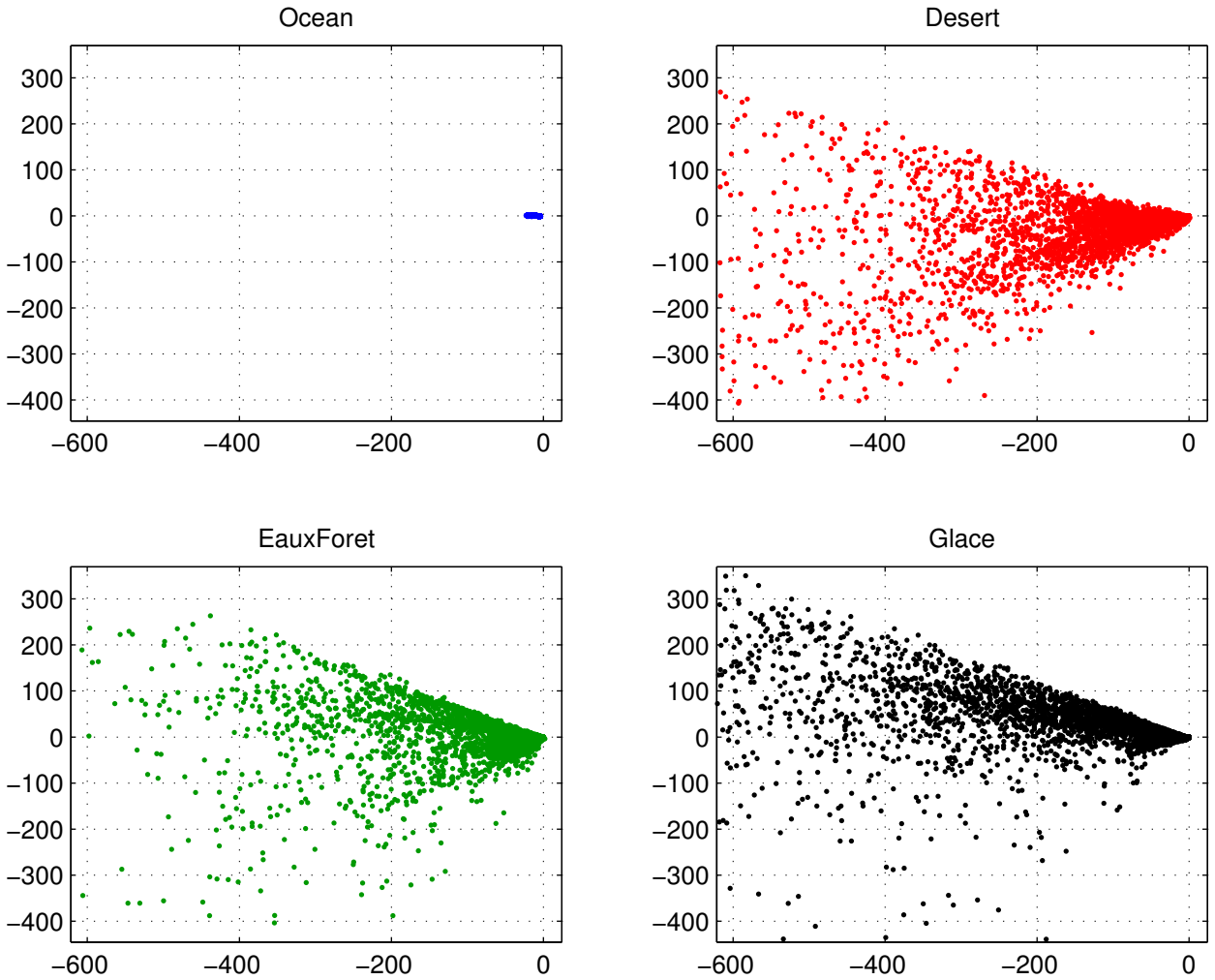
- Preprocessing: linear discriminant analysis
- Classification: Bayesian classifier, centroid distance rule, k nearest neighbors,...

Real Data

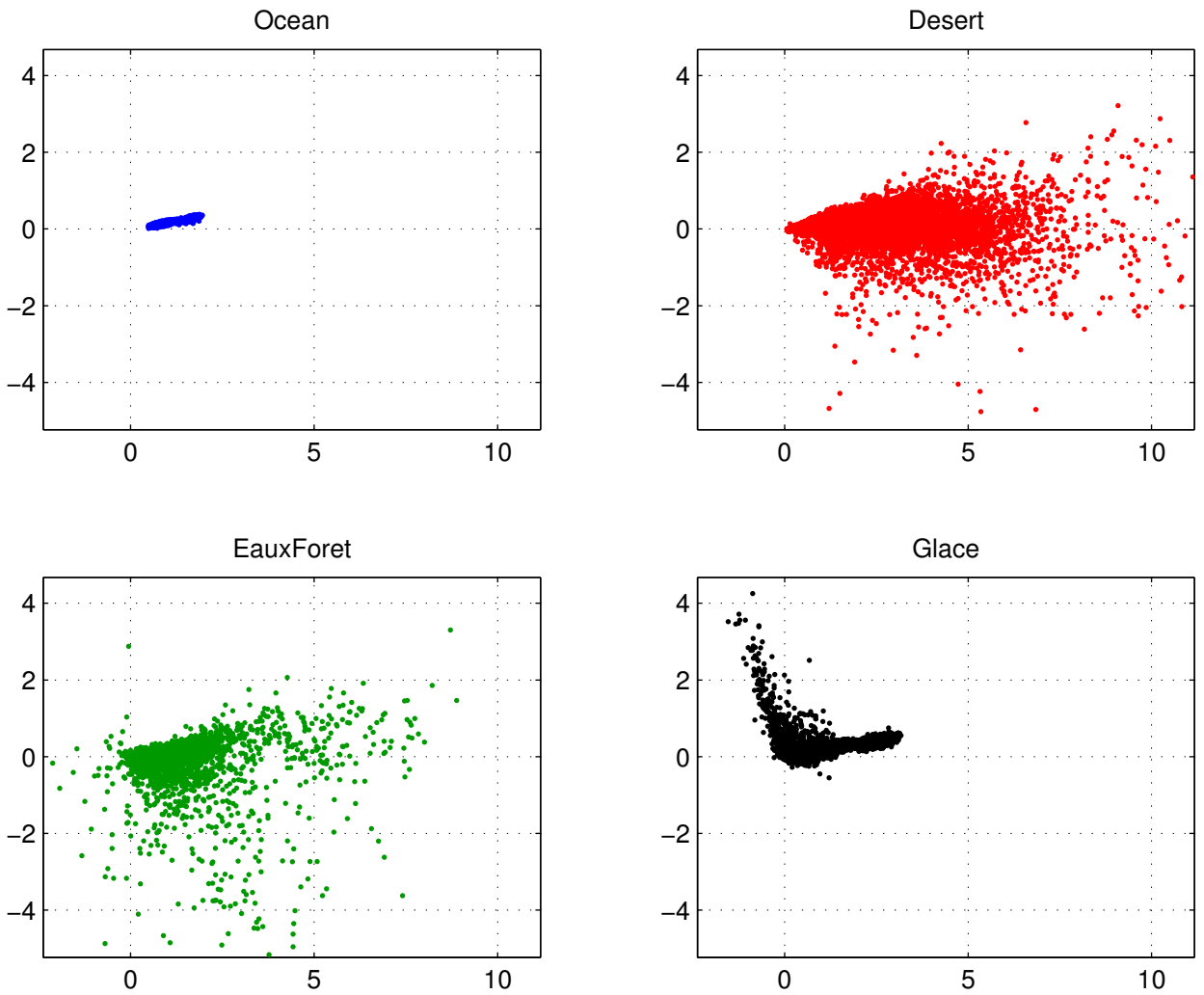
Waveform-based classification

- Class ω_1: Oceans
- Class ω_2: Deserts (Algeria, Lybia, South Africa)
- Class ω_3: Water and Forests (Amazonia, Canada, Congo, Russia)
- Class ω_4: Ice (Arctic continental ice, Groenland continental ice, Antarctic sea ice, Arctic sea ice)

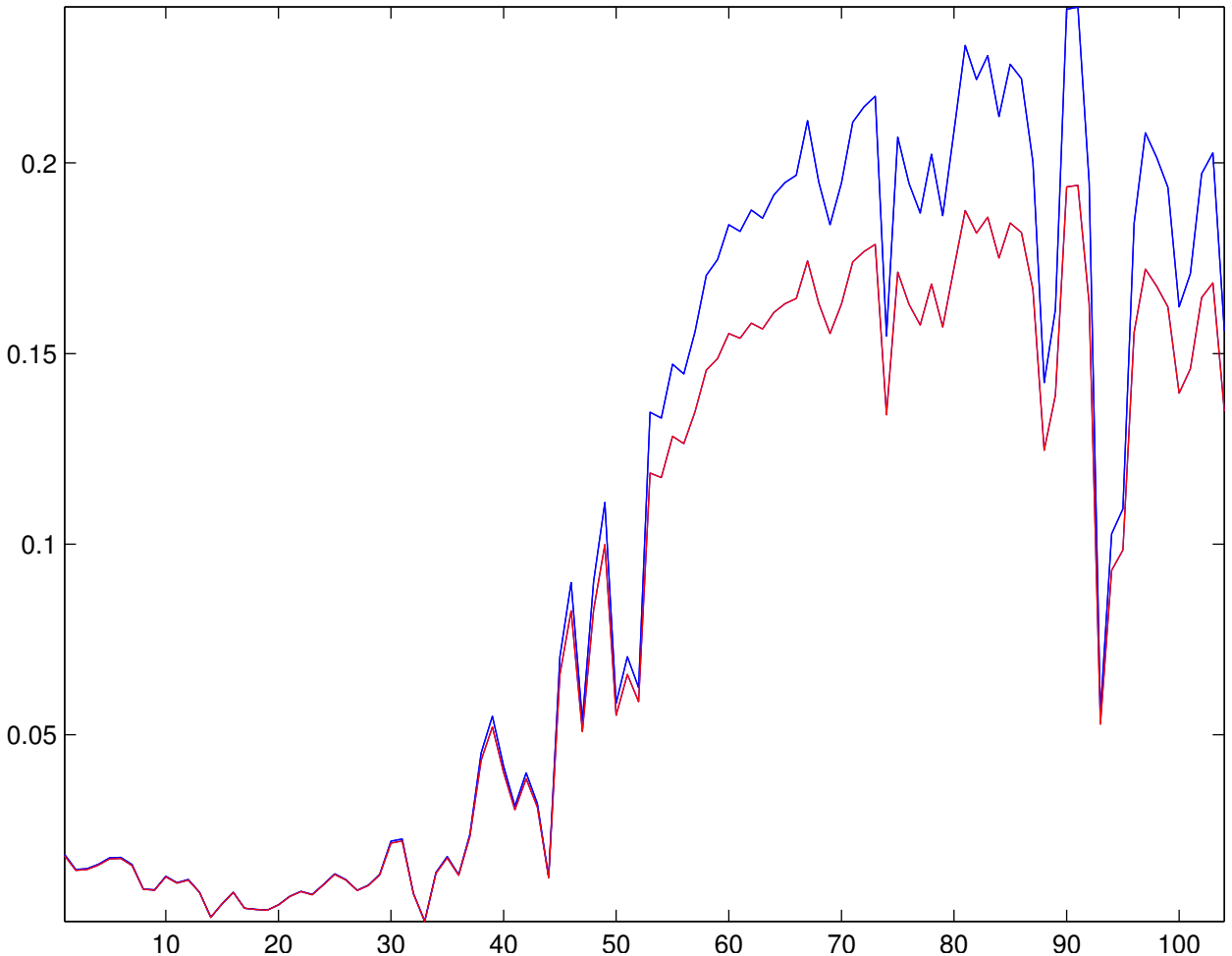
PCA of Samples (Waveforms)



Discriminant Analysis (Waveforms)



Fisher criterion



Features

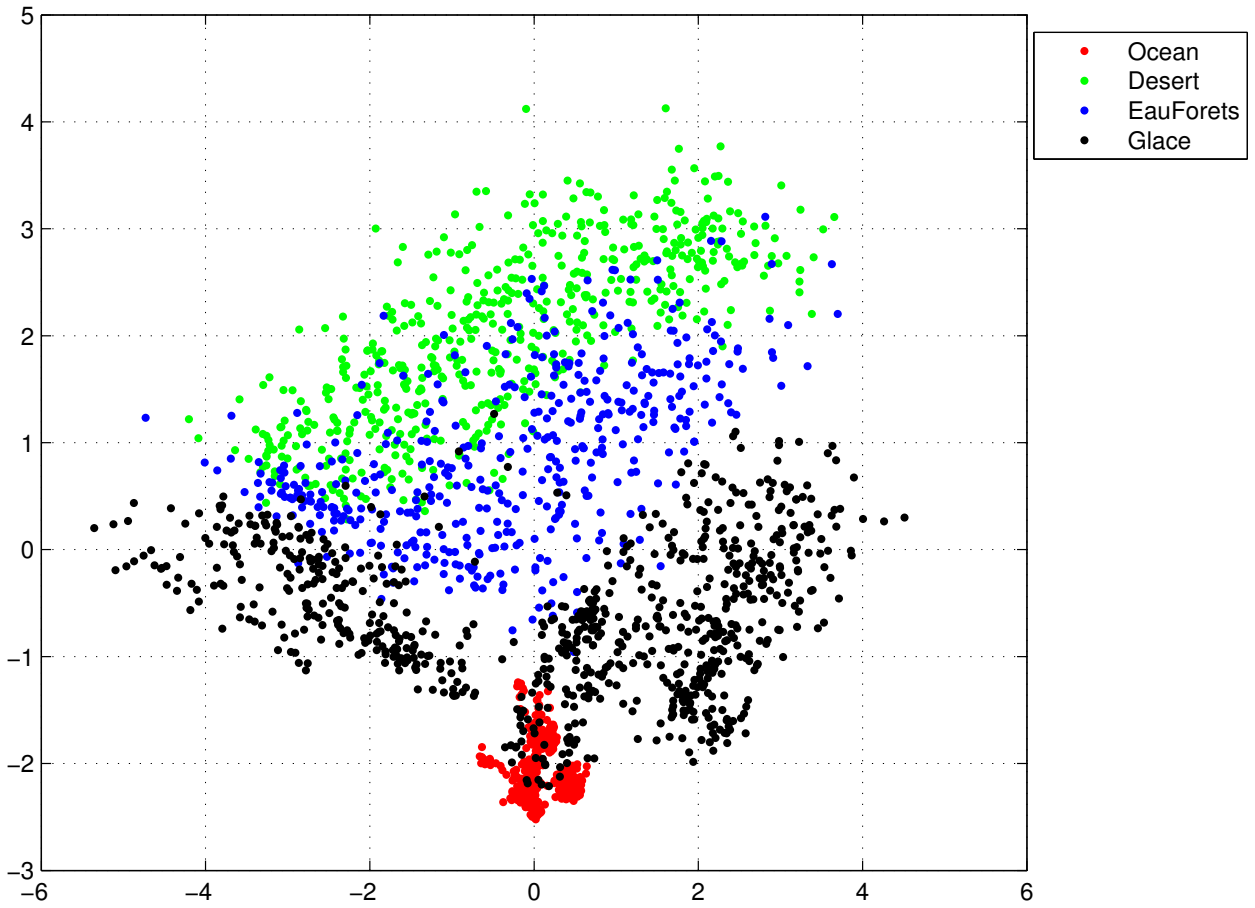
Parameters estimated from waveforms (Retracking)

- **ku_swh_1Hz** : Significant Wave Height of ku band
- **ku_std_swh_1Hz** : Standard deviation of swh 18 Hz
- **ku_sigma0_1Hz** : Backscatter coefficient ku band
- **s_sigma0_1Hz** : Backscatter coefficient s band
- **elevation_1Hz** : Elevation above the reference ellipsoid
- **ku_peakiness_1Hz** : Peakiness of waveforms in ku band
- **ku_ice2_width_1Hz** : Width of the leading edge

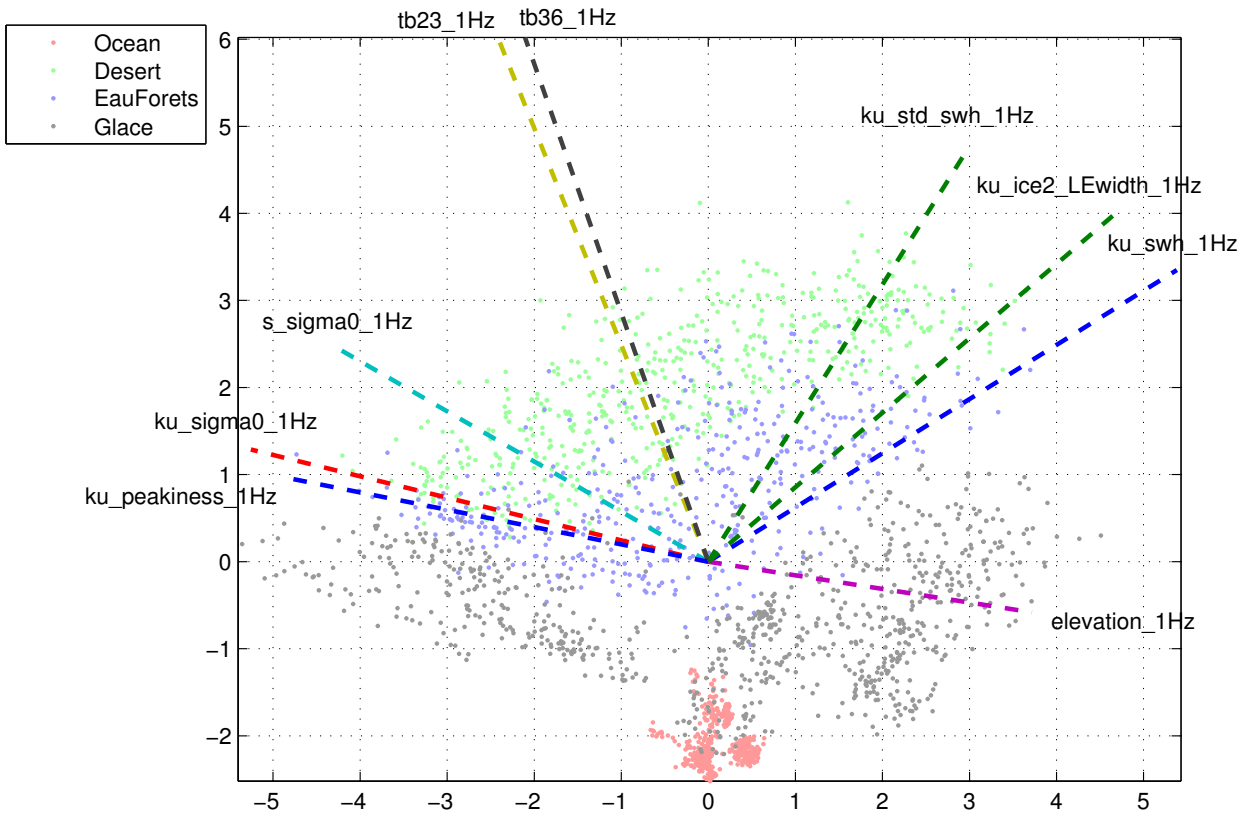
Parameters from radiometers

- **tb23_1Hz** : Brightness Temperature at 23 GHz
- **tb36_1Hz** : Brightness Temperature at 36 GHz

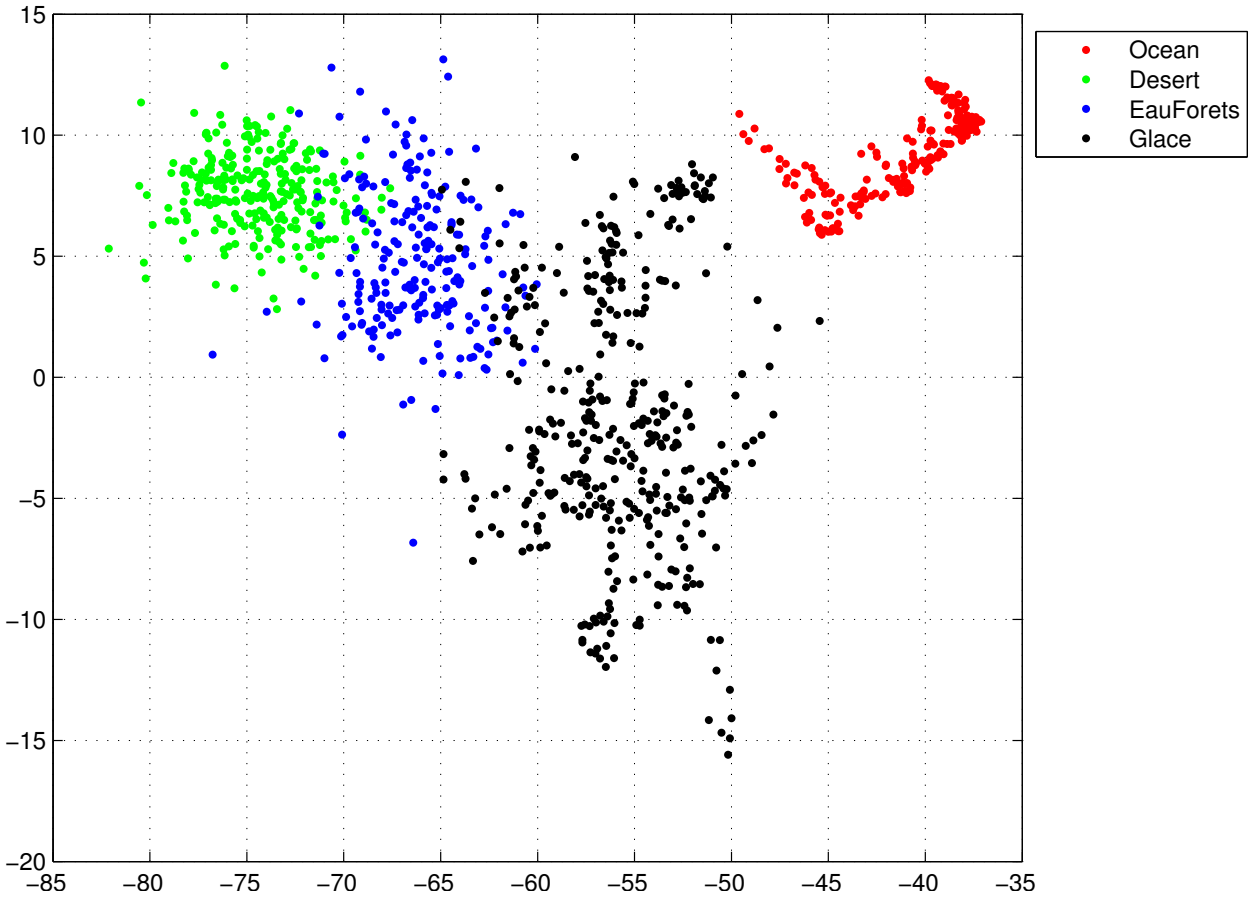
PCA of Samples (Features)



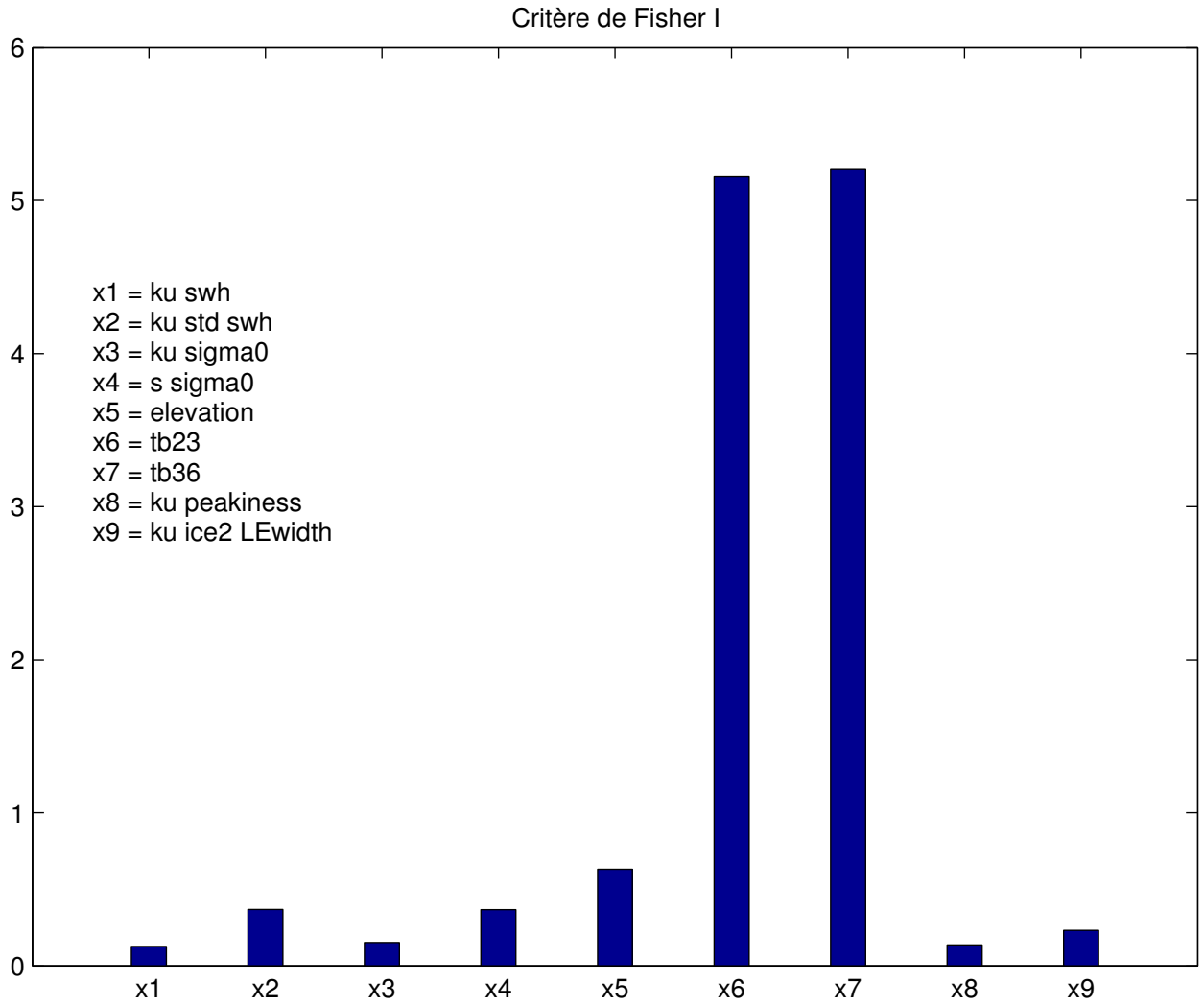
Joint PCA (features)



Discriminant Analysis (features)



Fisher Criterion (features)



Discriminant power of each parameter for classification

Summary

Classification

Notations

- K classes $\omega_1, ..., \omega_K$
- $\boldsymbol{x} = [x(1), ..., x(p)]^T$ measurements $\in X = \mathbb{R}^p$
- A: set of possible actions $a_1, ..., a_q$ where $a_i =$ "assign the vector $\boldsymbol{x}$ to the class ω_i", $\forall i = 1, ..., K$

Definition

$$d : \begin{array}{l} X \to A \\ \boldsymbol{x} \mapsto d(\boldsymbol{x}) \end{array}$$

Remark

Classification with reject option: $A = \{a_0, a_1, ..., a_K\}$ where $a_0 =$ "do not classify the vector $\boldsymbol{x}$"

Bayesian Rule

Hypothesis: Probabilistic Model

- *A priori* probability of class ω_i

$$P(\omega_i)$$

- Probability density function of the observation vector $\boldsymbol{x}$ conditionally to the class ω_i

$$f(\boldsymbol{x}|\omega_i)$$

Conclusion

- *A posteriori* probability that $\boldsymbol{x}$ belongs to class ω_i

$$P(\omega_i|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\omega_i)\,P(\omega_i)}{f(\boldsymbol{x})}$$

with $f(\boldsymbol{x}) = \sum_{i=1}^{K} f(\boldsymbol{x}|\omega_i)\,P(\omega_i)$.

MAP Classifier

Definition

$$d^*(\boldsymbol{x}) = a_j \Leftrightarrow P(\omega_j | \boldsymbol{x}) \geq P(\omega_k | \boldsymbol{x}), \forall k \in \{1, ..., K\}$$

Equiprobable Classes: Maximum Likelihood Classifier

$$d^*(\boldsymbol{x}) = a_j \Leftrightarrow f(\boldsymbol{x} | \omega_j) \geq f(\boldsymbol{x} | \omega_k), \forall k \in \{1, ..., K\}$$

Property

The MAP classifier minimizes the probability of error

Proof (2 classes)

$$\begin{aligned} P_e &= P\left[d\left(\boldsymbol{x}\right) = a_1 \cap \boldsymbol{x} \in \omega_2\right] + P\left[d\left(\boldsymbol{x}\right) = a_2 \cap \boldsymbol{x} \in \omega_1\right] \\ &= P\left[d\left(\boldsymbol{x}\right) = a_1 \mid \boldsymbol{x} \in \omega_2\right] P\left(\omega_2\right) + P\left[d\left(\boldsymbol{x}\right) = a_2 \mid \boldsymbol{x} \in \omega_1\right] P\left(\omega_1\right) \end{aligned}$$

Let $R_i = \{\boldsymbol{x} \in \mathbb{R}^p / d(\boldsymbol{x}) = a_i\}$ be the acceptance region for class ω_i

$$\begin{aligned} P_e &= \int_{R_2} P\left(\omega_1\right) f\left(\boldsymbol{x} \left|\omega_1\right.\right) d\boldsymbol{x} + \int_{R_1} P\left(\omega_2\right) f\left(\boldsymbol{x} \left|\omega_2\right.\right) d\boldsymbol{x} \\ &= P\left(\omega_2\right)\left[1 - \int_{R_2} f\left(\boldsymbol{x} \left|\omega_2\right.\right) d\boldsymbol{x}\right] + \int_{R_2} P\left(\omega_1\right) f\left(\boldsymbol{x} \left|\omega_1\right.\right) d\boldsymbol{x} \\ &= P\left(\omega_2\right) + \int_{R_2} \left[P\left(\omega_1\right) f\left(\boldsymbol{x} \left|\omega_1\right.\right) - P\left(\omega_2\right) f\left(\boldsymbol{x} \left|\omega_2\right.\right)\right] d\boldsymbol{x} \\ &= P\left(\omega_2\right) - \int_{R_2} \left[P\left(\omega_2 \left|\boldsymbol{x}\right.\right) - P\left(\omega_1 \left|\boldsymbol{x}\right.\right)\right] f\left(\boldsymbol{x}\right) d\boldsymbol{x} \end{aligned}$$

P_e is minimum when $R_2 = \{\boldsymbol{x} / P\left(\omega_2 \left|\boldsymbol{x}\right.\right) > P\left(\omega_1 \left|\boldsymbol{x}\right.\right)\}$

Probability of Error

- Definition (K classes)

$$P_e = \sum_{i=1}^{K} P\left[d\left(x\right) = a_i \cap x \notin \omega_i\right]$$

- Proof
  admitted

Cost Functions

Definitions

- Cost of making the decision a_j knowing that x belongs to class ω_i

$$c\left(a_j, \omega_i\right) \geq 0$$

- Average cost of making the decision a_j given x

$$R\left(a_j \middle| x\right) = \sum_{i=1}^{K} c\left(a_j, \omega_i\right) P\left(\omega_i \middle| x\right)$$

- Average cost of making the decision $d(x)$ given x

$$R_d\left(x\right) = \sum_{i=1}^{K} c\left(d(x), \omega_i\right) P\left(\omega_i \middle| x\right)$$

- Average overall cost associated with the decision rule d

$$R_d = \int_{\mathbb{R}^p} R_d\left(x\right) f\left(x\right) dx$$

Bayesian classification

Definition

Determine a decision function d which minimizes the average overall cost R_d. Since $f(\boldsymbol{x}) \geq 0$, the function d which minimizes R_d denoted as d^* is defined by

$$R_{d^*}(\boldsymbol{x}) \leq R_d(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \mathbb{R}^p$$

Classification without reject option

$$d^*(\boldsymbol{x}) = a_j \Longleftrightarrow R_j(\boldsymbol{x}) \leq R_k(\boldsymbol{x}) \qquad \forall k$$

with

$$R_j(\boldsymbol{x}) = \sum_{j=1}^{K} c_{ji} P(\omega_i | \boldsymbol{x}) = R(a_j | \boldsymbol{x}) \text{ and } c_{ji} = c(a_j, \omega_i)$$

Particular case: 2 classes

Definition of costs

$$\begin{aligned} R_1(x) &= c_{11}P(\omega_1|\boldsymbol{x}) + c_{12}P(\omega_2|\boldsymbol{x}) \\ R_2(\boldsymbol{x}) &= c_{21}P(\omega_1|\boldsymbol{x}) + c_{22}P(\omega_2|\boldsymbol{x}) \end{aligned}$$

with $c_{11} < c_{12}$ and $c_{22} < c_{21}$.

Bayesian Classifier

$$\begin{aligned} d^*(\boldsymbol{x}) &= a_1 \Leftrightarrow R_1(\boldsymbol{x}) \leq R_2(x) \\ d^*(\boldsymbol{x}) &= a_1 \Leftrightarrow \frac{f(\boldsymbol{x}|\omega_1)}{f(\boldsymbol{x}|\omega_2)} \geq \frac{c_{12}-c_{22}}{c_{21}-c_{11}} \frac{P(\omega_2)}{P(\omega_1)} \end{aligned}$$

Vocabulary

$\frac{f(\boldsymbol{x}|\omega_1)}{f(\boldsymbol{x}|\omega_2)}$ is the likelihood ratio.

Non-informative cost function

Definition

$$c_{ji} = 1 - \delta_{ij} = \begin{cases} 0 & \text{si } i = j \\ 1 & \text{si } i \neq j \end{cases}$$

Conditional Risks

$$R_j(\boldsymbol{x}) = \sum_{i=1}^{K} c_{ji} P(\omega_i | \boldsymbol{x}) = \sum_{i \neq j}^{K} P(\omega_i | \boldsymbol{x}) = 1 - P(\omega_j | \boldsymbol{x})$$

MAP Classifier

$$d^*(\boldsymbol{x}) = a_j \Leftrightarrow P(\omega_j | \boldsymbol{x}) \geq P(\omega_k | \boldsymbol{x}), \forall k \in \{1, ..., K\}$$

Maximum Likelihood Classifier

$$d^*(\boldsymbol{x}) = a_j \Leftrightarrow f(\boldsymbol{x} | \omega_j) \geq f(\boldsymbol{x} | \omega_k), \forall k \in \{1, ..., K\}$$

Gaussian Case

Densities

$$f\left(\boldsymbol{x}\left|\omega_i\right.\right) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma_i}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{m}_i\right)^t \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}-\boldsymbol{m}_i\right)\right]$$

General Case

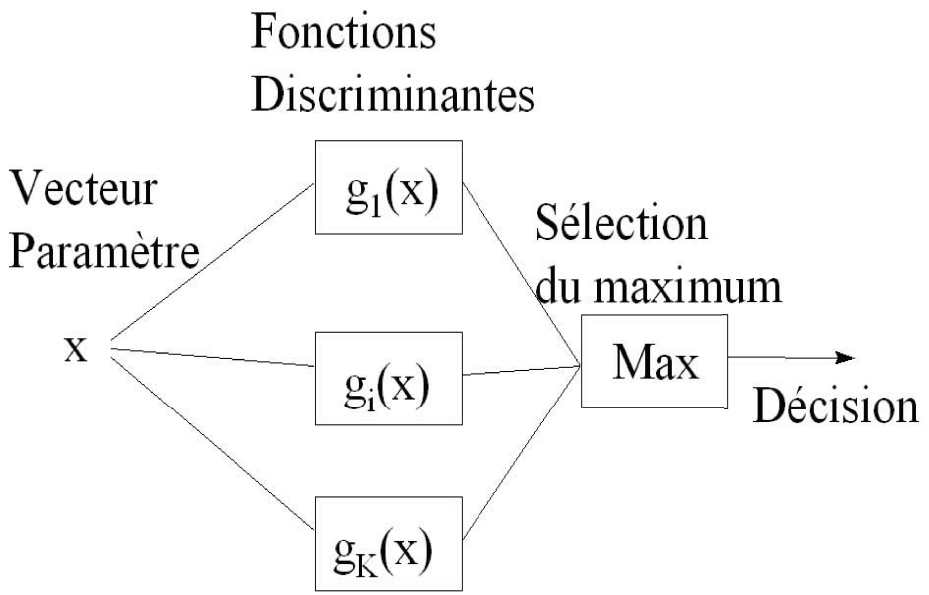
$$d^*\left(\boldsymbol{x}\right) = a_i \Leftrightarrow g_i\left(\boldsymbol{x}\right) \geq g_k\left(k\right) \qquad \forall k$$

with

$$g_i\left(\boldsymbol{x}\right) = -\left(\boldsymbol{x}-\boldsymbol{m}_i\right)^t \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{x}-\boldsymbol{m}_i\right) - \ln\det \boldsymbol{\Sigma}_i + 2\ln P\left(\omega_i\right)$$

Gaussian Case

Classifier



Discriminant functions

Quadratic term + linear term + constant

Identical covariance matrices ($\Sigma_i = \Sigma$)

Centroid Distance Rule

$$d^*(\boldsymbol{x}) = a_i \Leftrightarrow d_M(\boldsymbol{x}, \boldsymbol{m}_i) \leq d_M(\boldsymbol{x}, \boldsymbol{m}_k) \qquad \forall k$$

where d_M is the Mahalanobis distance

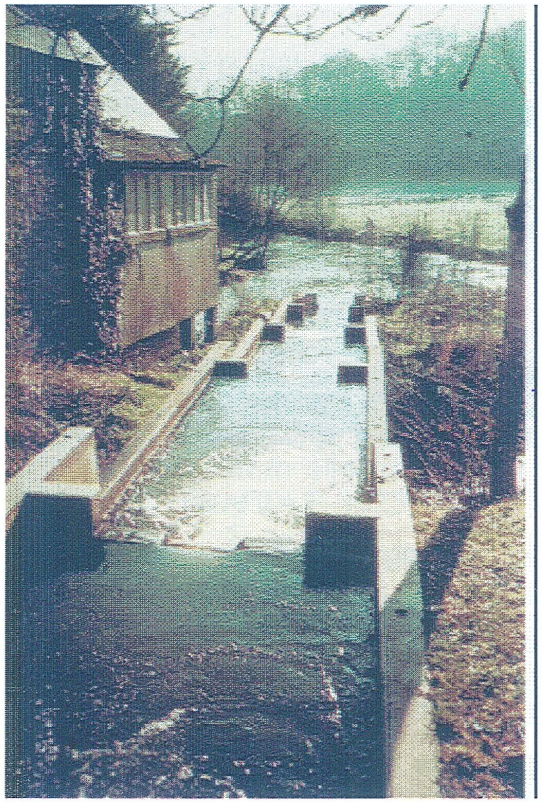
$$d_M(\boldsymbol{x}, \boldsymbol{m}_k) = \sqrt{(\boldsymbol{x} - \boldsymbol{m}_k)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{m}_k)}$$

Affine discriminant functions

$$d^*(\boldsymbol{x}) = a_i \Leftrightarrow \left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{m}_i + \boldsymbol{m}_k)\right)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}_i - \boldsymbol{m}_k) \geq \ln \frac{P(\omega_k)}{P(\omega_i)}$$

Contexte de l'étude

Les passes à poissons



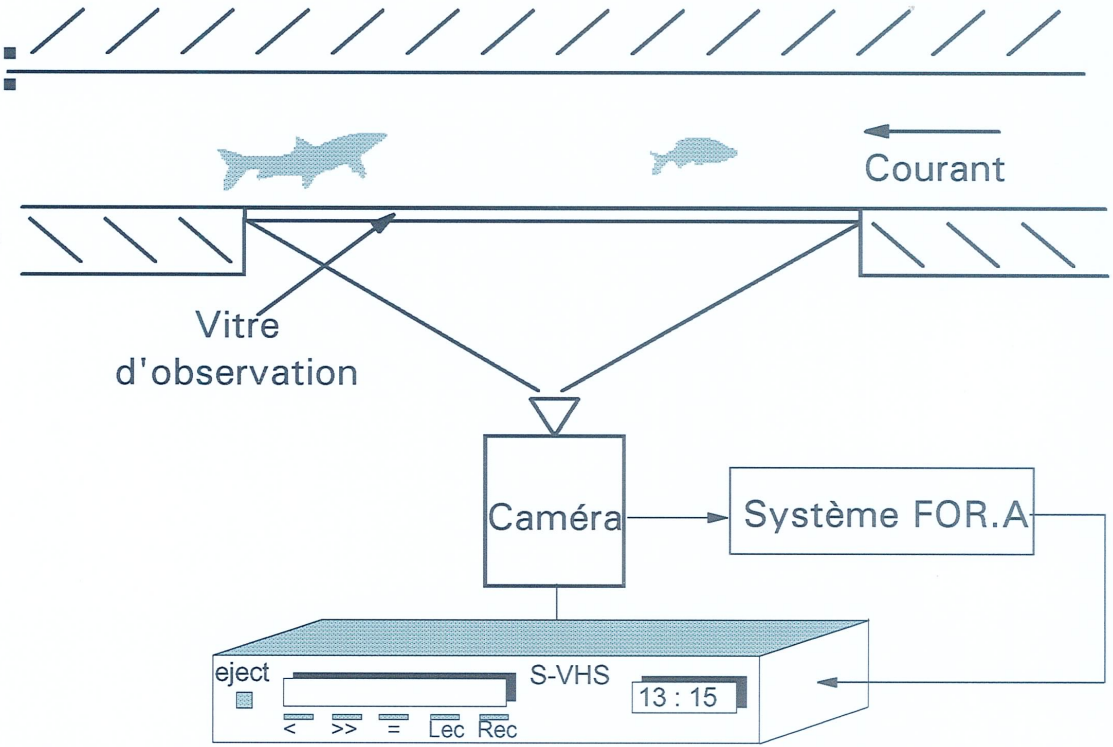
à bassins successifs

à ralentisseurs

ascenseur

Suivi des passes à poissons

- Dispositif vidéo actuel de comptage par espèces :

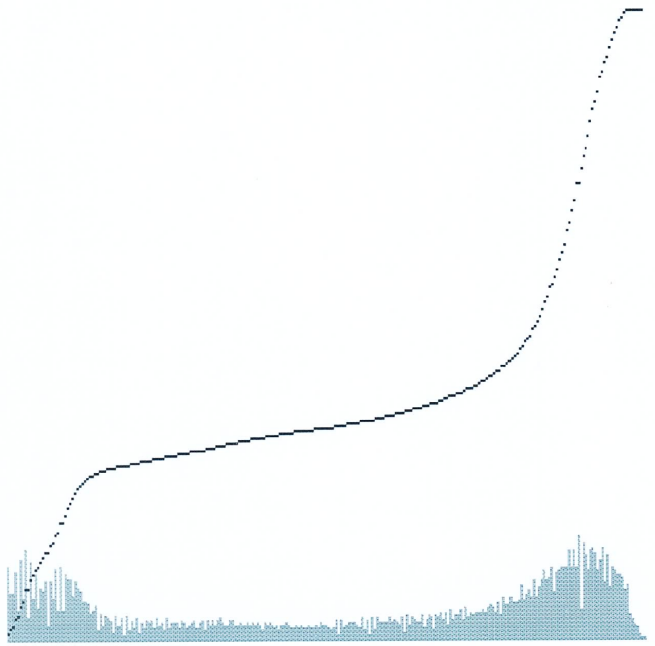


- Dépouillement des bandes vidéo enregistrées par un opérateur humain

Acquisition des images

Nouvelles conditions d'acquisition

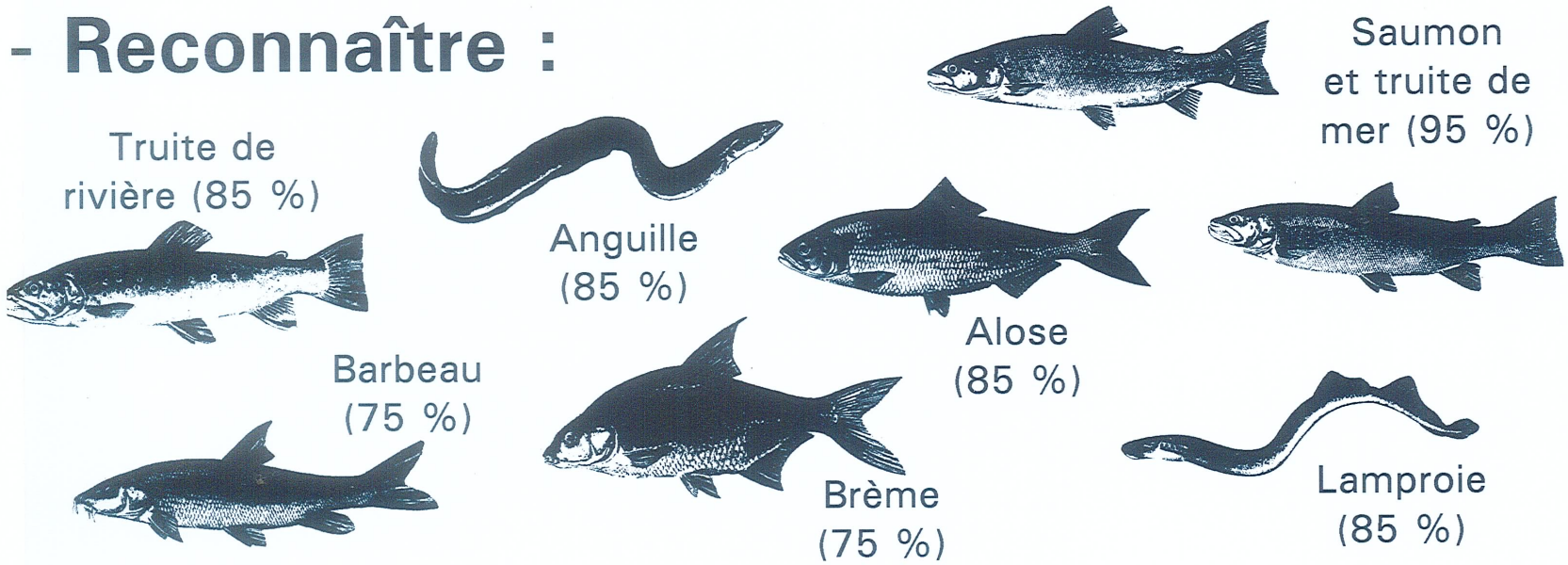
Image



Histogramme : bimodal

Cahier des charges du système de vision automatique

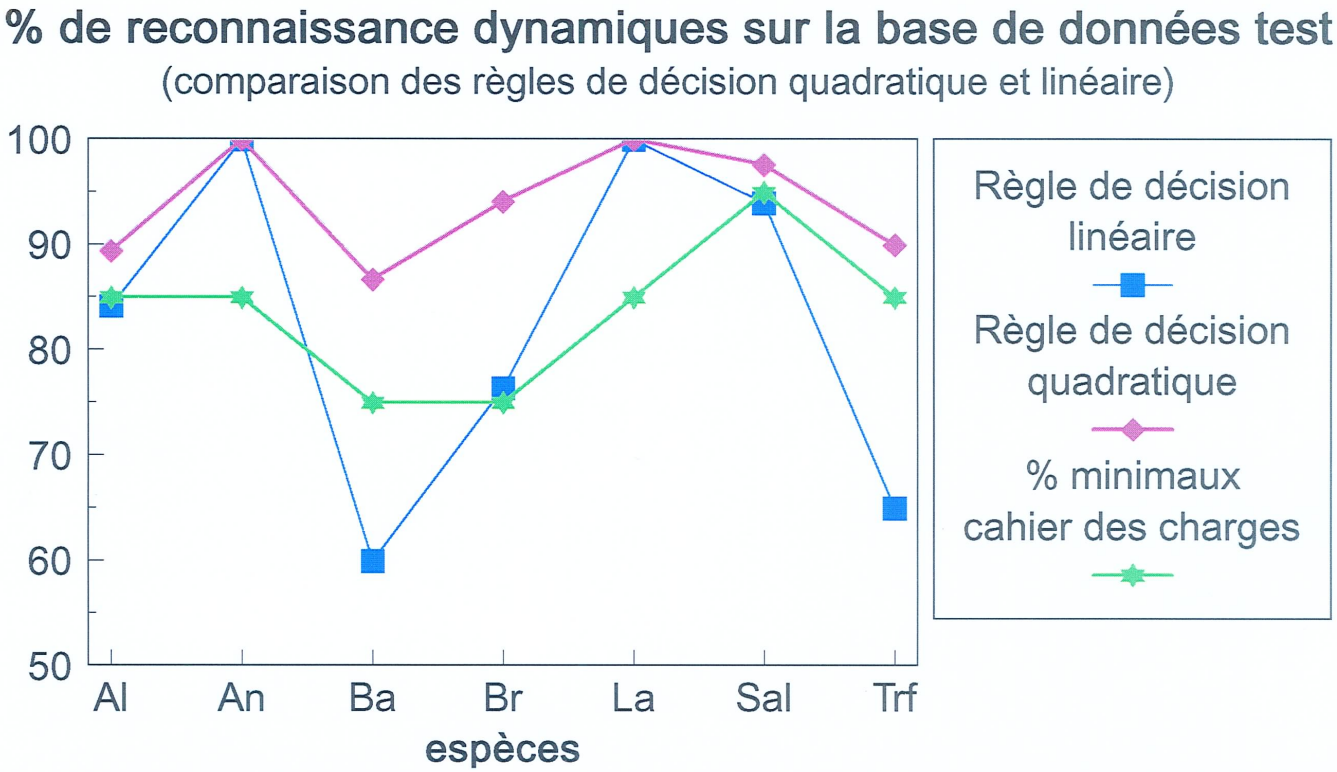
- **Compter les poissons, toutes espèces confondues, au delà d'une certaine taille**
- **Reconnaître :**



- **Fournir :** La comptabilisation des différentes espèces
  Les dates et heures de passage

Mesure d'efficacité de la règle de classement

Résultats obtenus sur la BDT



Summary

MMSE Estimator

$\theta \in \mathbb{R}$

$$E\left[\left(\theta - \widehat{\theta}_{MMSE}(\boldsymbol{x})\right)^2\right] = \min_{\pi} E\left[(\theta - \pi(\boldsymbol{x}))^2\right]$$

with $\boldsymbol{x} = (x_1, ..., x_n)$, hence the MMSE estimator

$$\widehat{\theta}_{MMSE}(\boldsymbol{x}) = E[\theta \,|\boldsymbol{x}]$$

$\theta \in \mathbb{R}^p$

The MMSE estimator $\widehat{\boldsymbol{\theta}}_{MMSE}(x) = E[\boldsymbol{\theta} \,|\boldsymbol{x}]$ minimizes the quadratic cost

$$E\left[(\boldsymbol{\theta} - \pi(\boldsymbol{x}))^t Q (\boldsymbol{\theta} - \pi(\boldsymbol{x}))\right]$$

for any symmetric definite positive matrix $\boldsymbol{Q}$ (and in particular for $\boldsymbol{Q} = I_p$, the identity matrix).

MAP Estimator

$\theta \in \mathbb{R}$

The MAP estimator $\widehat{\theta}_{\mathrm{MAP}}(\boldsymbol{x})$ minimizes the average of a "uniform" cost function

$$c(\theta - \pi(\boldsymbol{x})) = \begin{cases} 0 \text{ si } |\theta - \pi(\boldsymbol{x})| \leq \frac{\Delta}{2} \\ 1 \text{ si } |\theta - \pi(\boldsymbol{x})| > \frac{\Delta}{2} \end{cases}$$

and is defined by

$$c\left(\theta - \widehat{\theta}_{\mathrm{MAP}}(x)\right) = \min_{\pi} c(\theta - \pi(\boldsymbol{x}))$$

If Δ is arbitrary small, $\widehat{\theta}_{MAP}(\boldsymbol{x})$ is the value of $\pi(\boldsymbol{x})$ which maximizes the posterior $p(\theta|\boldsymbol{x})$ hence its name MAP estimator. The MAP estimator is computed by setting to zero the derivative of $p(\theta|\boldsymbol{x})$ (or of its logarithm) with respect to θ.

$\theta \in \mathbb{R}^p$

Determine the values of θ_i which make the partial derivatives of $p(\boldsymbol{\theta}|\boldsymbol{x})$ (or of its logarithm) with respect to θ_i equal to zero.

Summary

Histogram

Density estimation

- Probability: P_R
  Let $\boldsymbol{x}$ be a random vector of $\mathbb{R}^p$

$$P_R = P\left[\boldsymbol{x} \in R\right] = \int_R f\left(\boldsymbol{u}\right) d\boldsymbol{u}$$

- Number of vectors $\boldsymbol{x}_i$ belonging to the region R: N_R

$$P\left[N_R = k\right] = \mathcal{C}_n^k P_R^k \left(1 - P_R\right)^{n-k} \qquad k \in \{0, ..., n\}$$

$$E\left(\frac{N_R}{n}\right) = P_R, \quad \text{Var}\left(\frac{N_R}{n}\right) = \frac{P_R\left(1 - P_R\right)}{n}$$

N_R/n is an unbiased and convergent estimator of P_R.

- Estimator of the probability density function: $f\left(\boldsymbol{x}\right)$
  If f is approximately constant on R

$$\int_R f\left(\boldsymbol{u}\right) d\boldsymbol{u} \simeq f\left(\boldsymbol{x}\right) V_R \Rightarrow \boxed{\widehat{f}\left(\boldsymbol{x}\right) = \frac{N_R}{nV_R}}$$

Remarks

- If the volume V_R is fixed and the number of data n increases, N_R/n converges to P_R but $\frac{N_R}{nV_R}$ converges to the mean of f on R and not to $f(\boldsymbol{x})$.
- If the volume V decreases (n being fixed), the region R can contain zero observation, hence $\widehat{f}(\boldsymbol{x}) \simeq 0$.

$\Rightarrow n$ and R should vary simultaneously

We build a sequence of regions R_k containing $\boldsymbol{x}$ and N_k observations

- let V_k be the "volume" of the region R_k
- let $\widehat{f_k}(\boldsymbol{x}) = \frac{N_k}{kV_k}$ be the kth estimation of $f(\boldsymbol{x})$.

How can we define the regions R_k in order to ensure $\widehat{f_k}(\boldsymbol{x})$ is a good estimator of $f(\boldsymbol{x})$?

Convergence

To ensure the convergence of $\widehat{f}_k(\boldsymbol{x})$ to $f(\boldsymbol{x})$ when $k \to \infty$, the following three conditions have to be satisfied:

$$\begin{aligned} \lim_{k\to\infty} V_k &= 0 \\ \lim_{k\to\infty} N_k &= \infty \\ \lim_{k\to\infty} \frac{N_k}{k} &= 0. \end{aligned}$$

The problem is to choose the sequence of pairs (N_k, V_k) which satisfies these conditions. Two methods:

- Parzen windows
- k nearest neighbors

Parzen windows

Definition of a kernel ϕ

Let R_k be an hypercube of side h_k and dimension p. Thus $V_k = h_k^p$ and

$$N_k = \sum_{i=1}^{k} \phi\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h_k}\right)$$

with

$$\begin{aligned} &\phi(\boldsymbol{u}) = 1 \text{ if } |u(j)| \leq 1/2 \qquad \forall j \in \{1, ..., p\} \\ &\phi(\boldsymbol{u}) = 0 \text{ else} \end{aligned}$$

Indeed, $\phi\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h_k}\right) = 1$ if $\boldsymbol{x}_i$ is in the region R_k.

Density estimator

$$\widehat{f}_k(x) = \frac{1}{k h_k^p} \sum_{i=1}^{k} \phi\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h_k}\right)$$

Effect of parameter h_k

Define

$$\delta_k(\boldsymbol{x}) = \frac{1}{V_k}\phi\left(\frac{\boldsymbol{x}}{h_k}\right)$$

- When h_k is "large", $\delta_k(\boldsymbol{x})$ varies slowly
  - $\widehat{f}_k(\boldsymbol{x})$ is the superimposition of k slowly varying functions
  - $\widehat{f}_k(\boldsymbol{x})$ is an estimator of $f(\boldsymbol{x})$ with limited resolution.
- When h_k is "small", $\delta_k(\boldsymbol{x})$ varies rapidly
  - $\widehat{f}_k(\boldsymbol{x})$ is the superimposition of k narrow functions centered on the observed samples
  - $\widehat{f}_k(\boldsymbol{x})$ is a noisy estimation of $f(x)$.

Generalization to other kernels

$\widehat{f_k}(\boldsymbol{x})$ is a probability density function if ϕ satisfies the following conditions

$$\begin{array}{l} \phi(\boldsymbol{u}) \geq 0 \qquad \forall \boldsymbol{u} \\ \int_{\mathbb{R}^p} \phi(\boldsymbol{u})\, d\boldsymbol{u} = 1. \end{array}$$

In order to ensure that $\widehat{f_k}(\boldsymbol{x})$ is an unbiased and convergent estimator of $f(\boldsymbol{x})$, ϕ has to satisfy the following (necessary and sufficient) conditions

$$\begin{array}{rcl} \sup_{\boldsymbol{u}} \phi(\boldsymbol{u}) & < & +\infty, \quad \lim_{h_k \to \infty} \frac{1}{V_k} \phi\left(\frac{\boldsymbol{u}}{h_k}\right) = \delta(u) \\ \lim_{k \to \infty} V_k & = & 0, \quad \lim_{k \to \infty} k V_k = +\infty \end{array}$$

These conditions are for instance verified for $V_k = \frac{V_1}{\sqrt{k}}$ or $V_k = \frac{V_1}{\ln k}$

k-nearest neighbor method

Assumptions

N_k is fixed (e.g., N_k is the integer part of $\sqrt{k}$)
V_k is estimated

Density estimator

$$\widehat{f}_k(\boldsymbol{x}) = \frac{N_k}{kV_k}$$

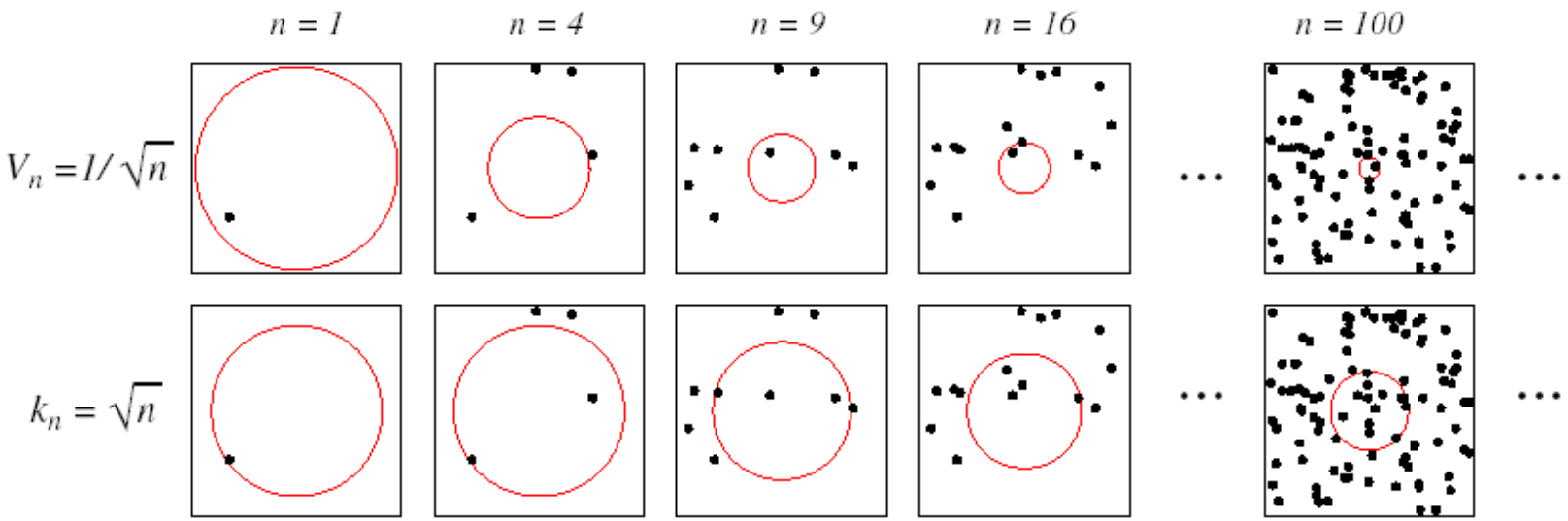
Convergence

The estimator $\widehat{f}_k(\boldsymbol{x})$ converges to $f(\boldsymbol{x})$ if and only if

$$\begin{aligned} \lim_{k\to\infty} N_k &= \infty \\ \lim_{k\to\infty} \frac{N_k}{k} &= 0 \end{aligned}$$

Fenêtres de Parzen vs. k-PPV (2)

- **Illustration des méthodes**
  1) **des fenêtres de Parzen**
  2) **des k plus proches voisins**



The k-nearest neighbor rule

The nearest neighbor rule

$$\boxed{d(x) = a_j \text{ if the nearest neighbor of } x \text{ belongs to } \omega_j}$$

The observed vector $\boldsymbol{x}$ is affected to the class of its nearest neighbor.

Inequality of Cover and Hart

$$P^* \leq P_1 \leq P^* \left(2 - \frac{K}{K-1} P^*\right)$$

The k-nearest neighbor rule

$\boldsymbol{x}$ is assigned to the class most common amongst its k-nearest neighbors (with a given distance measure)

Probability of error

Inequalities

$$P^* \leq P_k \leq P^* + \frac{1}{\sqrt{ke}} \text{ or } P^* \leq P_k \leq P^* + \sqrt{\frac{2P_1}{k}}$$

Approximations

When P^* is small, the following results can be obtained

$$P_1 \approx 2P^* \text{ and } P_3 \approx P^* + 3\left(P^*\right)^2$$

Summary

Unsupervised learning

N unlabelled data vectors of $\mathbb{R}^p$ denoted as $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ which should be split into K classes $\omega_1, ..., \omega_K$.

Motivations

- Supervised learning is costly
- The classes can change with time
- Provide some information about the data structure

Optimal solution

Number of partitions of $\boldsymbol{X}$ in K subsets

$$P(N, K) = \frac{1}{K!} \sum_{i=0}^{K} i^N (-1)^{K-i} C_K^i \qquad K < N$$

Example: $P(100, 5) \approx 10^{68}$!

Partition with minimum mean square error

Mean square error of a partition

Mean square error (MSE) of the class ω_i and of the partition $\boldsymbol{X}$

$$E_i^2 = \sum_{k=1}^{N_i} d^2\left(\boldsymbol{x}_k, \boldsymbol{g}_i\right), \qquad E^2 = \sum_{i=1}^{K} E_i^2$$

where $\boldsymbol{g}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} \boldsymbol{x}_k$ is the centroid of the class ω_i

Properties

$$\sum_{k=1}^{N_i} d^2\left(\boldsymbol{x}_k, \boldsymbol{y}\right) = \sum_{k=1}^{N_i} d^2\left(\boldsymbol{x}_k, \boldsymbol{g}_i\right) + N_i d^2\left(\boldsymbol{g}_i, \boldsymbol{y}\right)$$

In particular, for $\boldsymbol{y} = \boldsymbol{g}$ (data centroid), we obtain

$$\sum_{i=1}^{K}\sum_{k=1}^{N_i} d^2\left(\boldsymbol{x}_k, \boldsymbol{g}\right) = \underbrace{\sum_{i=1}^{K}\sum_{k=1}^{N_i} d^2\left(\boldsymbol{x}_k, \boldsymbol{g}_i\right)}_{E^2} + \sum_{i=1}^{K} N_i d^2\left(\boldsymbol{g}_i, \boldsymbol{g}\right)$$

MSE of $\boldsymbol{X}$ = within-class MSE + between-class MSE

ISODATA algorithm

Search a partition of $\boldsymbol{X}$ ensuring a local minimum of E^2

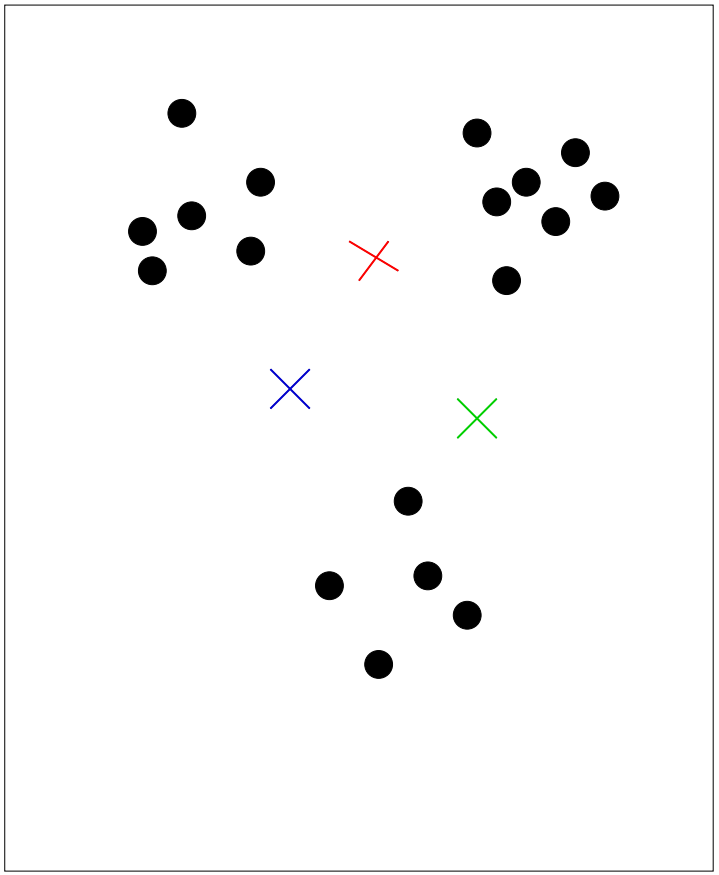
1. Initial choice of the number of classes and the class centroids
2. Assign each vector $\boldsymbol{x}_i$ to ω_j (using the centroid distance rule) such that

$$d\left(\boldsymbol{x}_i, \boldsymbol{g}_j\right) = \inf_k d\left(\boldsymbol{x}_i, \boldsymbol{g}_k\right)$$

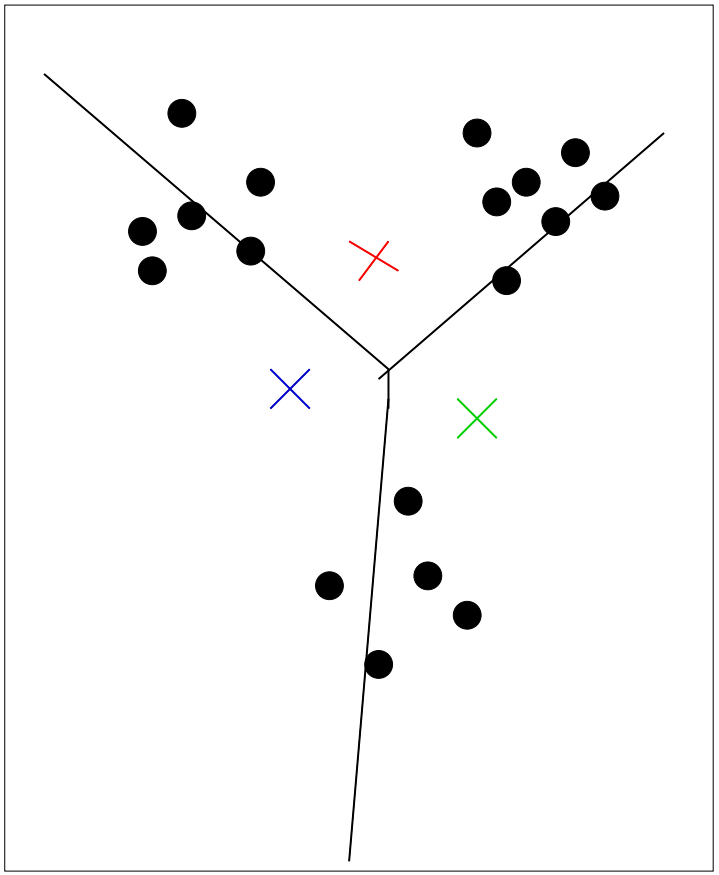
3. Compute the centroids $\boldsymbol{g}_k^*$ of the new classes ω_k^*
4. Repeat steps 2) and 3) until convergence

- *Improved version of ISODATA*
  - Two classes are merged if their centroids are close
  - A class is split if it contains too many vectors x_i or if its mean square error is too large
- *Convergence*: see notes or textbooks

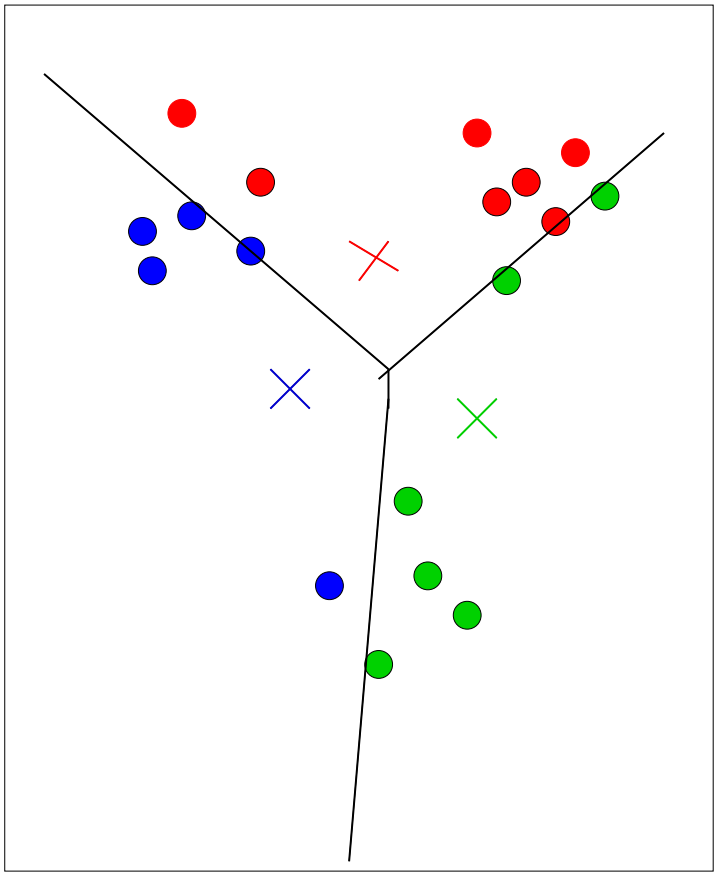
K-means



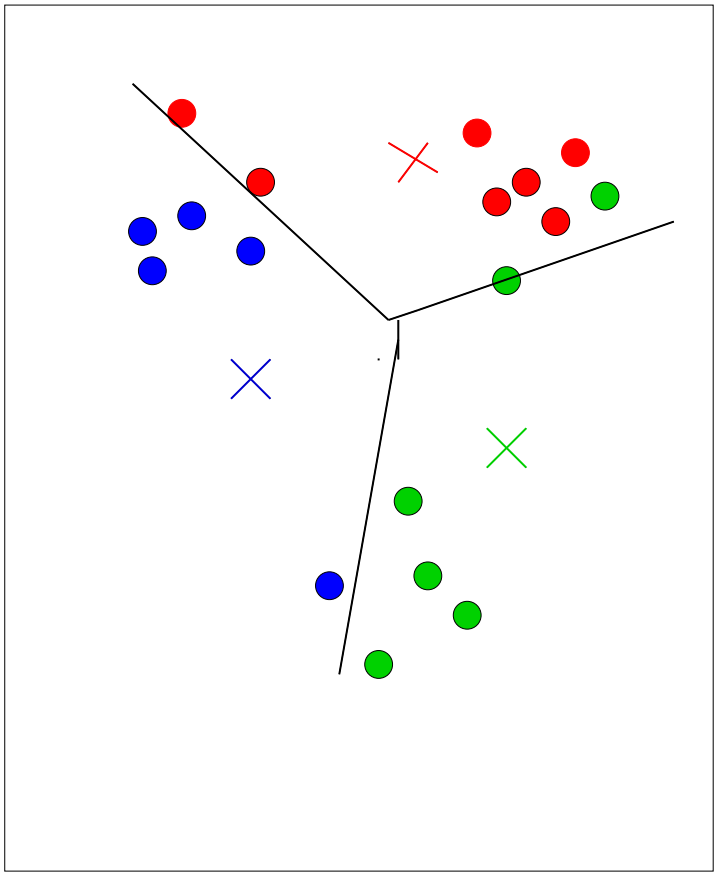
K-means



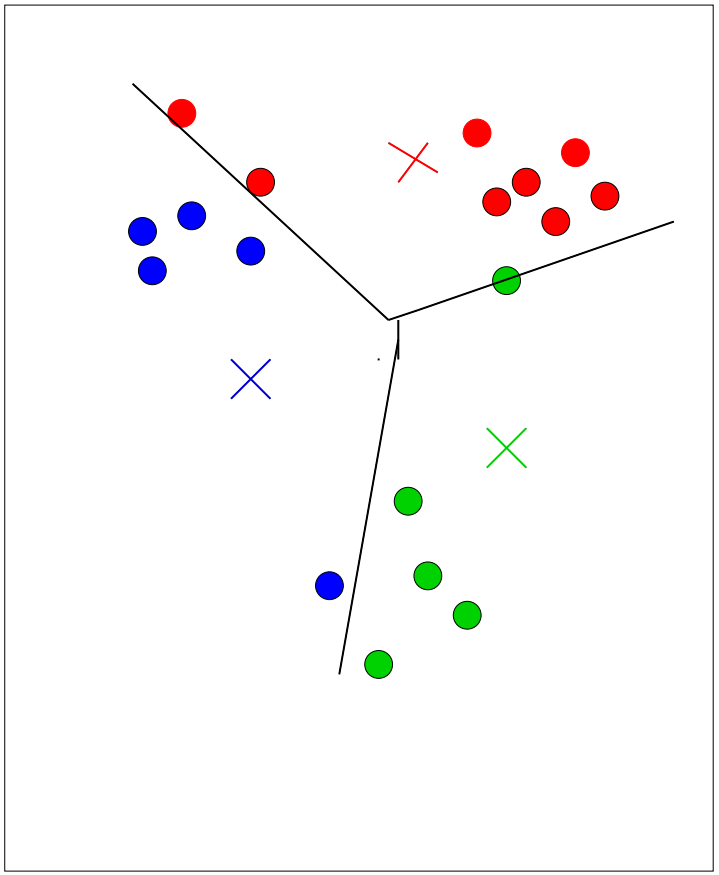
K-means



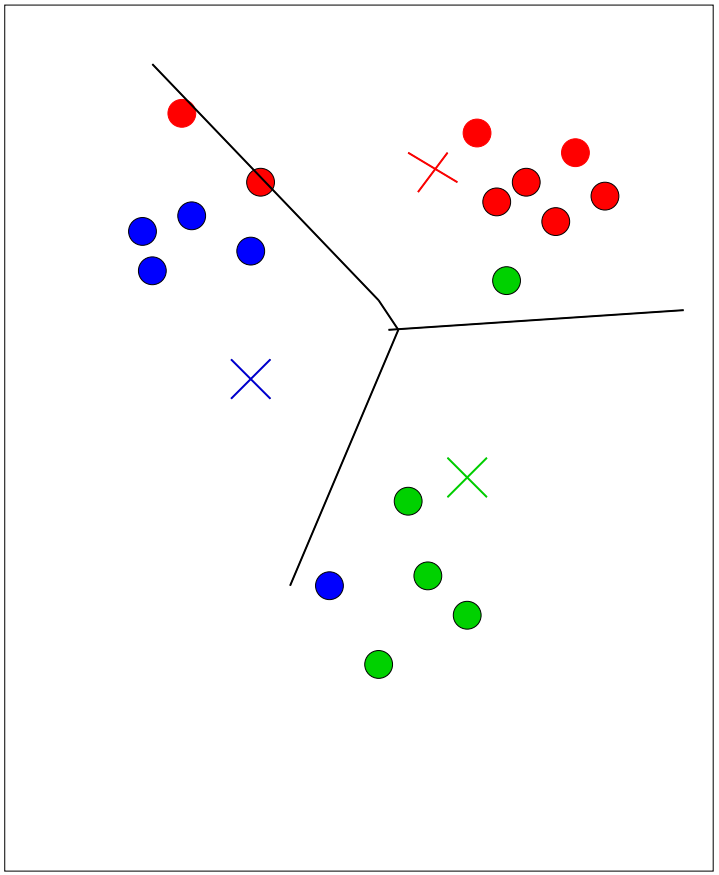
K-means



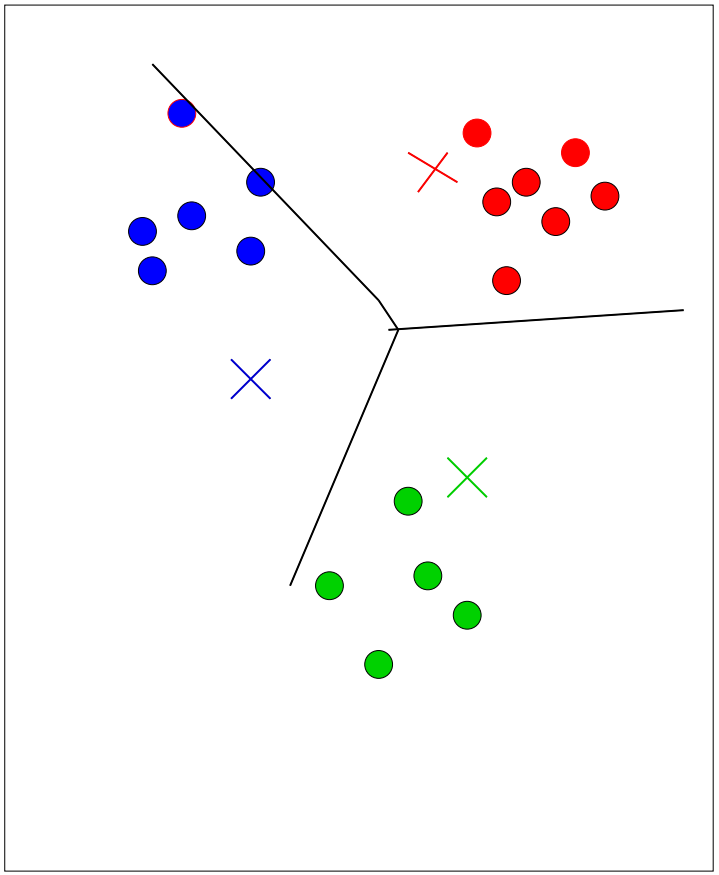
K-means



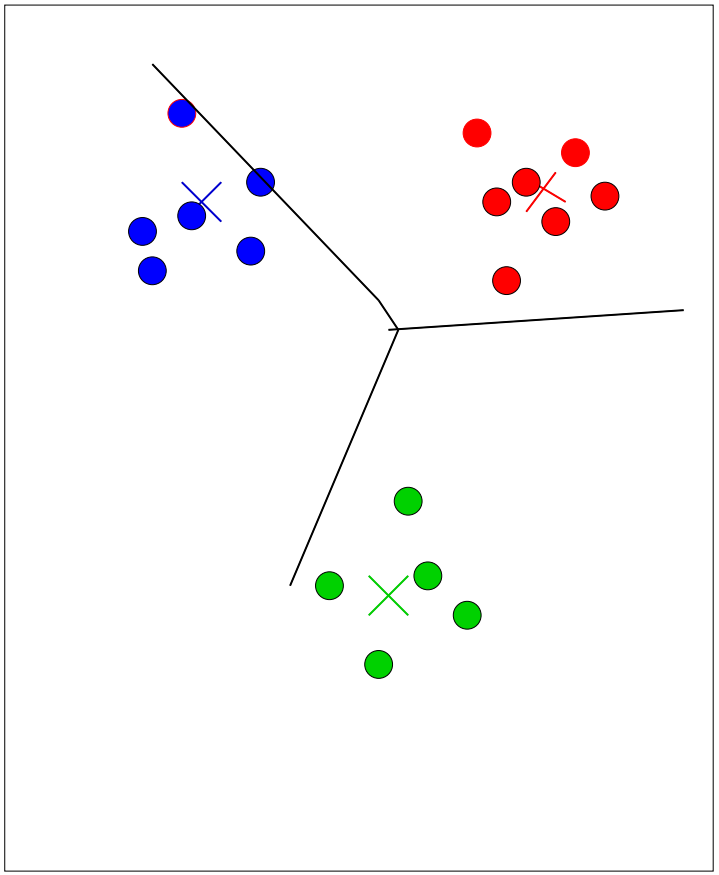
K-means



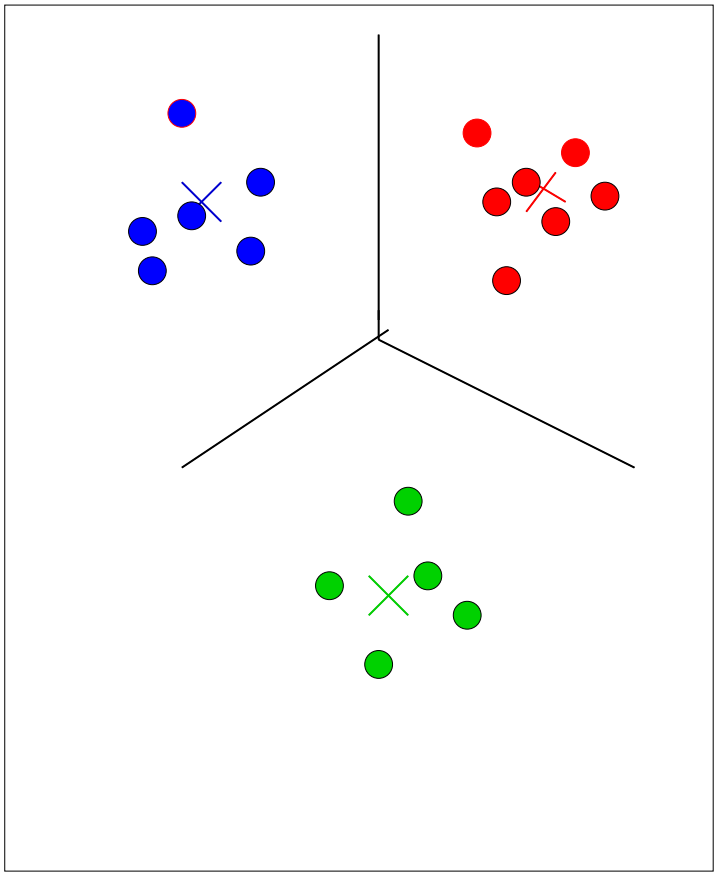
K-means



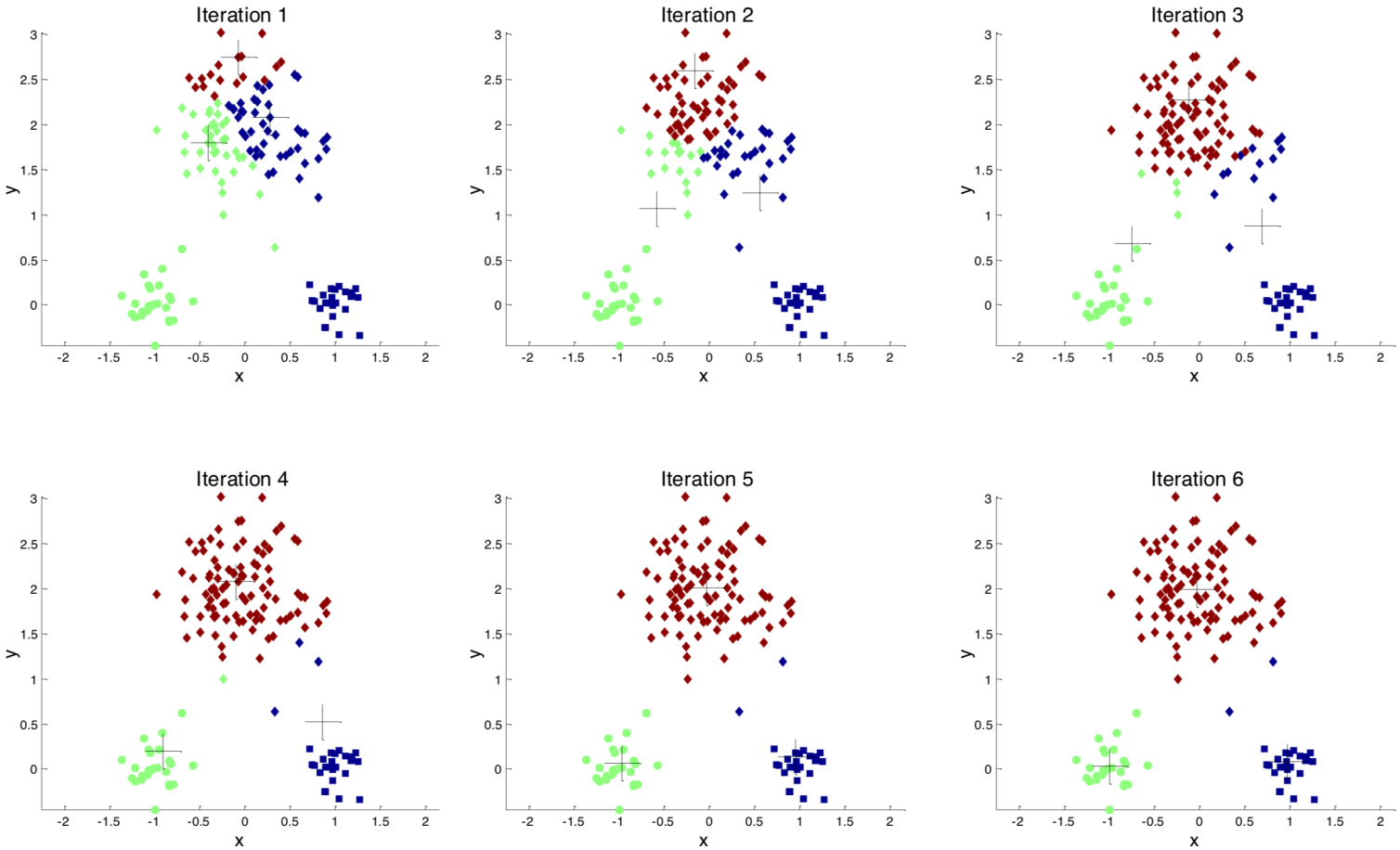
K-means



K-means

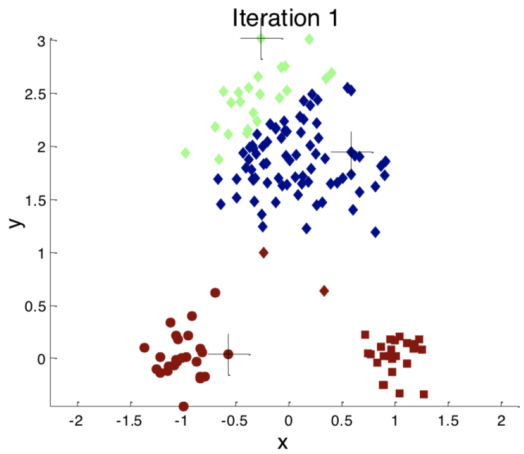


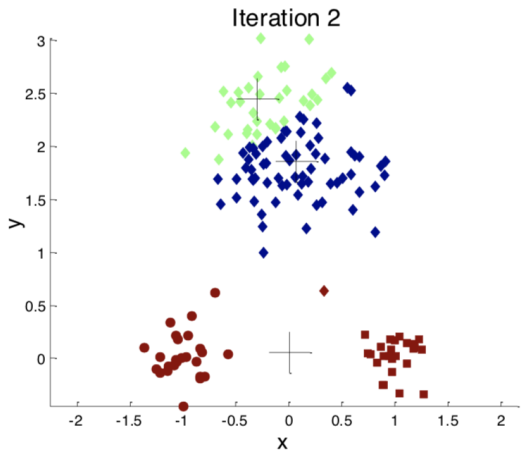
Initialization Problems

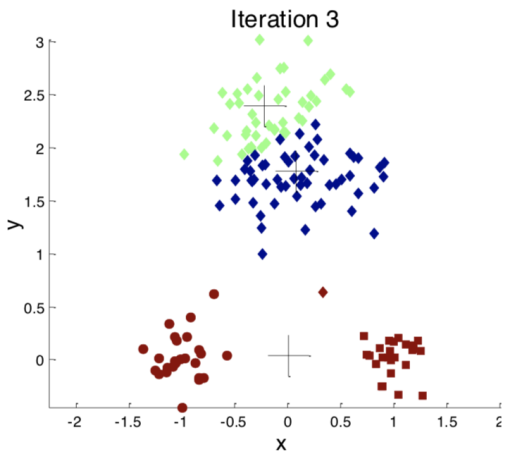


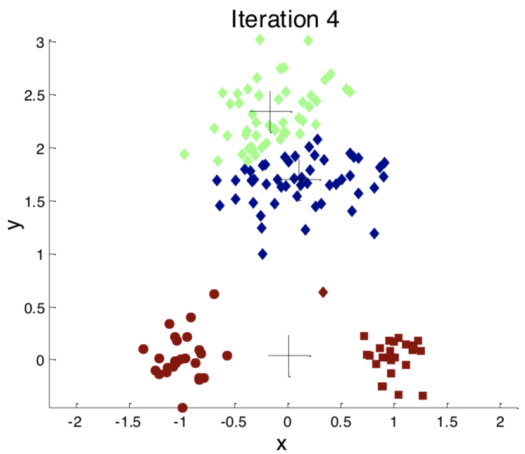
Thanks to Jing Gao from SUNY Buffalo university for her slides!!

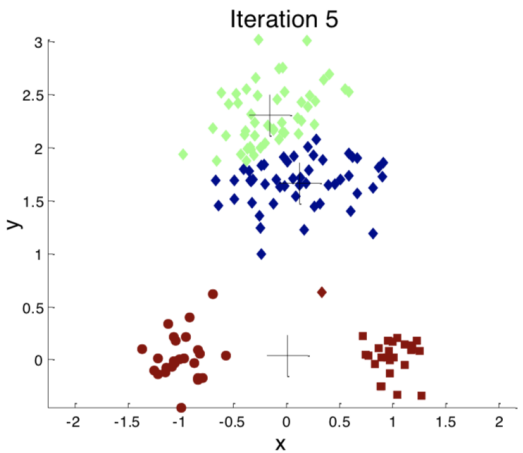
Initialization Problems



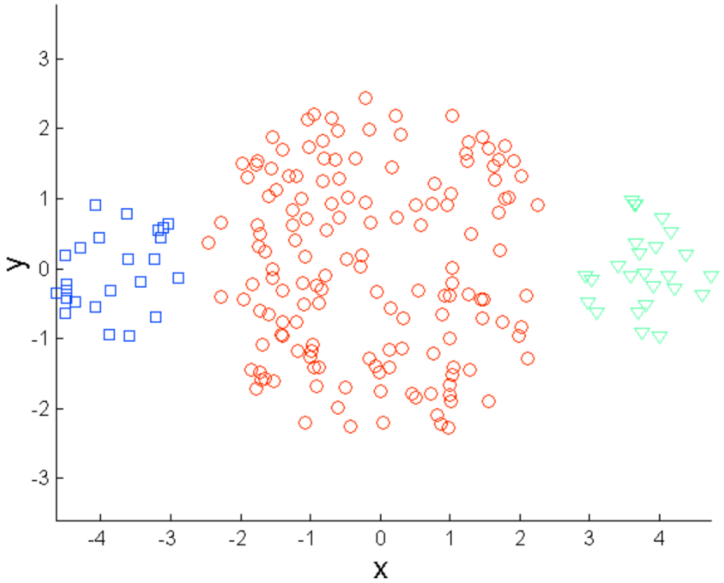




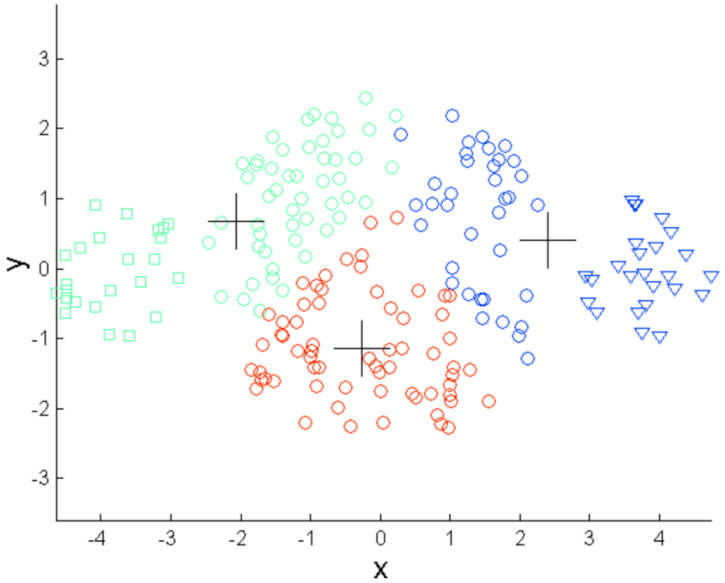




Limitations of K-means: Differing Sizes

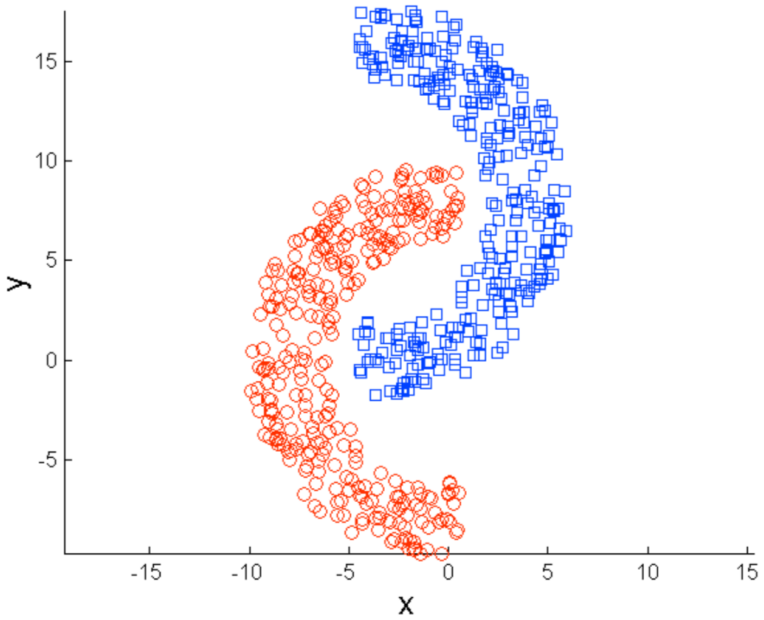


Original Points

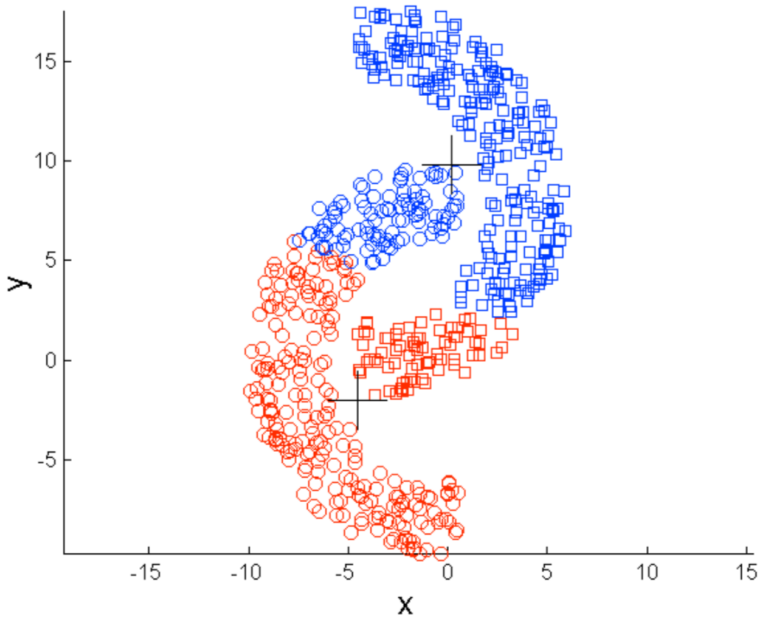


K-means (3 Clusters)

Limitations of K-means: Irregular Shapes

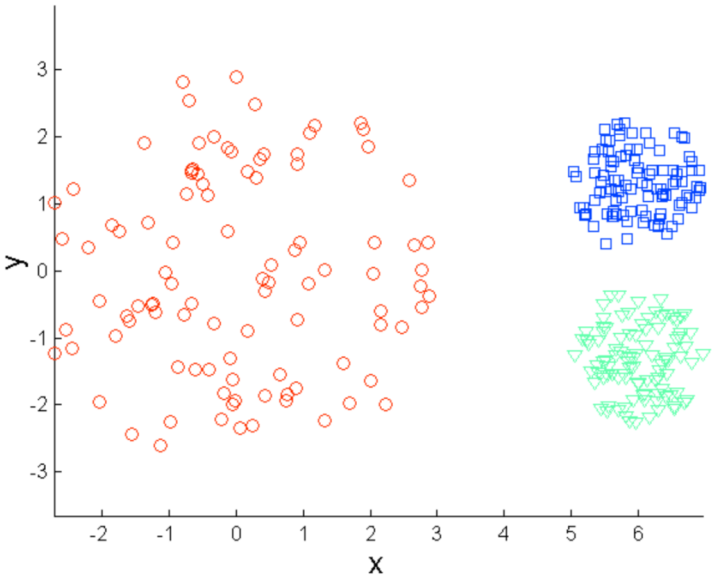


Original Points

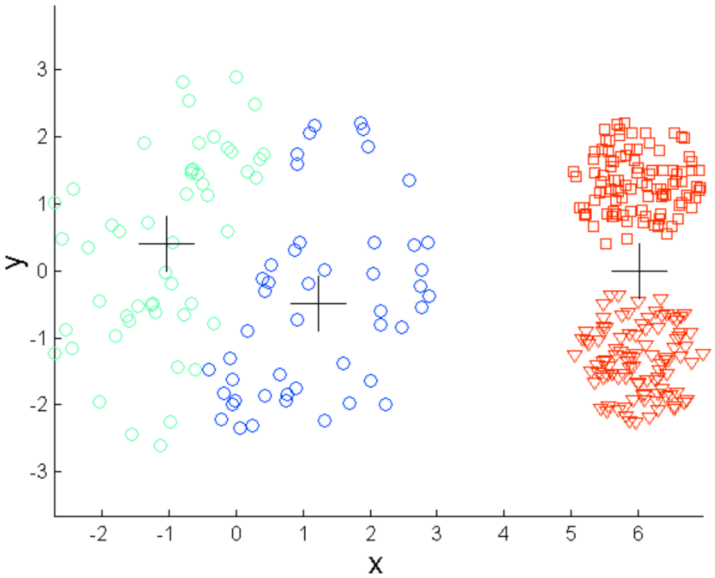


K-means (2 Clusters)

Limitations of K-means: Differing Density



Original Points



K-means (3 Clusters)

Summary

Hierarchical classification

Ascending hierarchy: method of distances

- Distance *Min* (single linkage algorithm)

$$d(X_i, X_j) = \min d(\boldsymbol{x}, \boldsymbol{y}) \qquad \boldsymbol{x} \in X_i, \boldsymbol{y} \in X_j$$

  This distance favors elongated classes

- Distance *Max* (complete linkage algorithm)

$$d(X_i, X_j) = \max d(\boldsymbol{x}, \boldsymbol{y}) \qquad \boldsymbol{x} \in X_i, \boldsymbol{y} \in X_j$$

- Average linkage algorithm

$$d(X_i, X_j) = \frac{1}{N_i N_j} \sum_{\boldsymbol{x} \in X_i, \boldsymbol{y} \in X_j} d(\boldsymbol{x}, \boldsymbol{y})$$

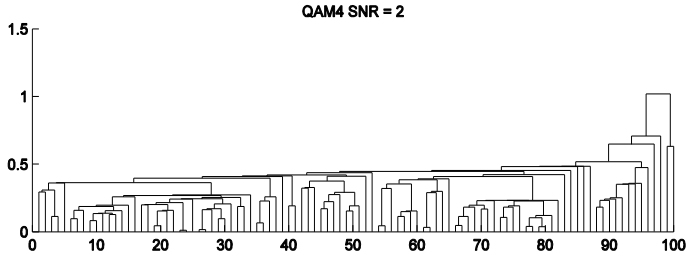
- Distance between the means

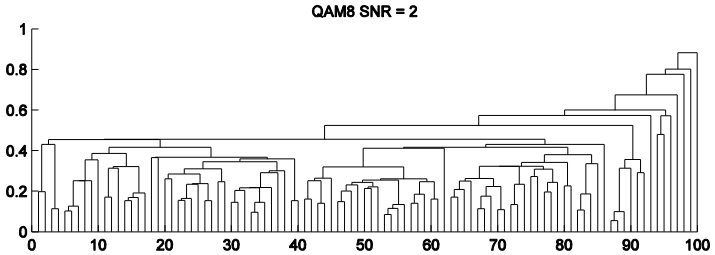
$$d(X_i, X_j) = d\left(\boldsymbol{g}_i, \boldsymbol{g}_j\right)$$

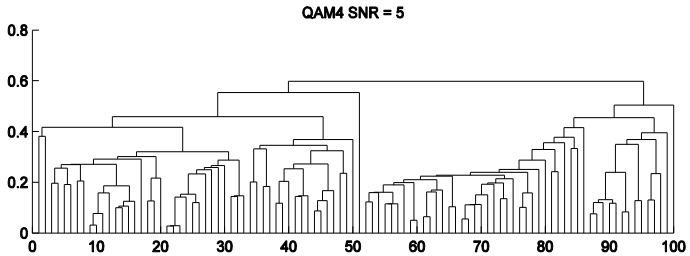
  where X_i and X_j have cardinals N_i and N_j and centroids $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$.

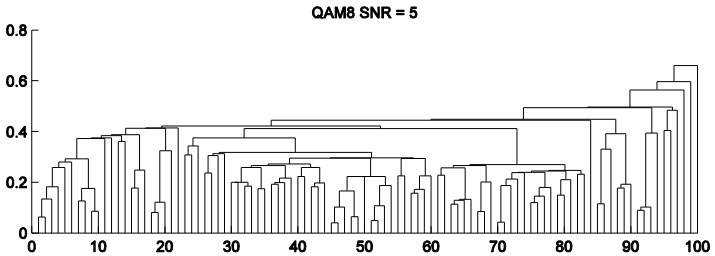
Representation using a tree whose nodes indicate the different groups

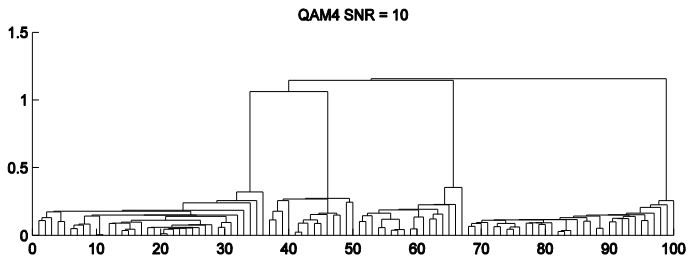
Classification of Modulations

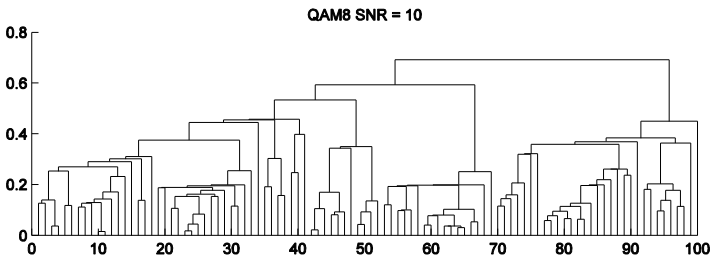


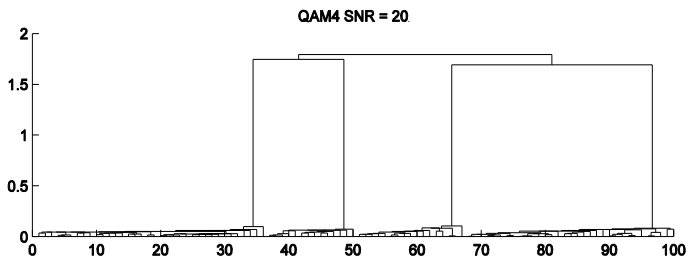


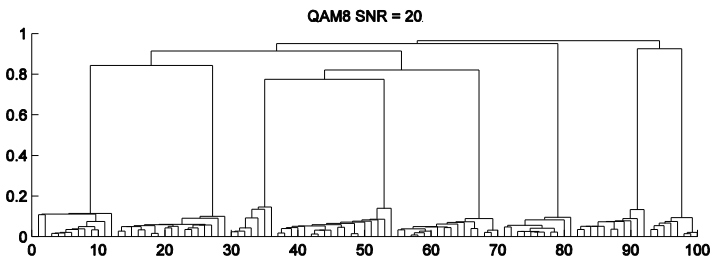




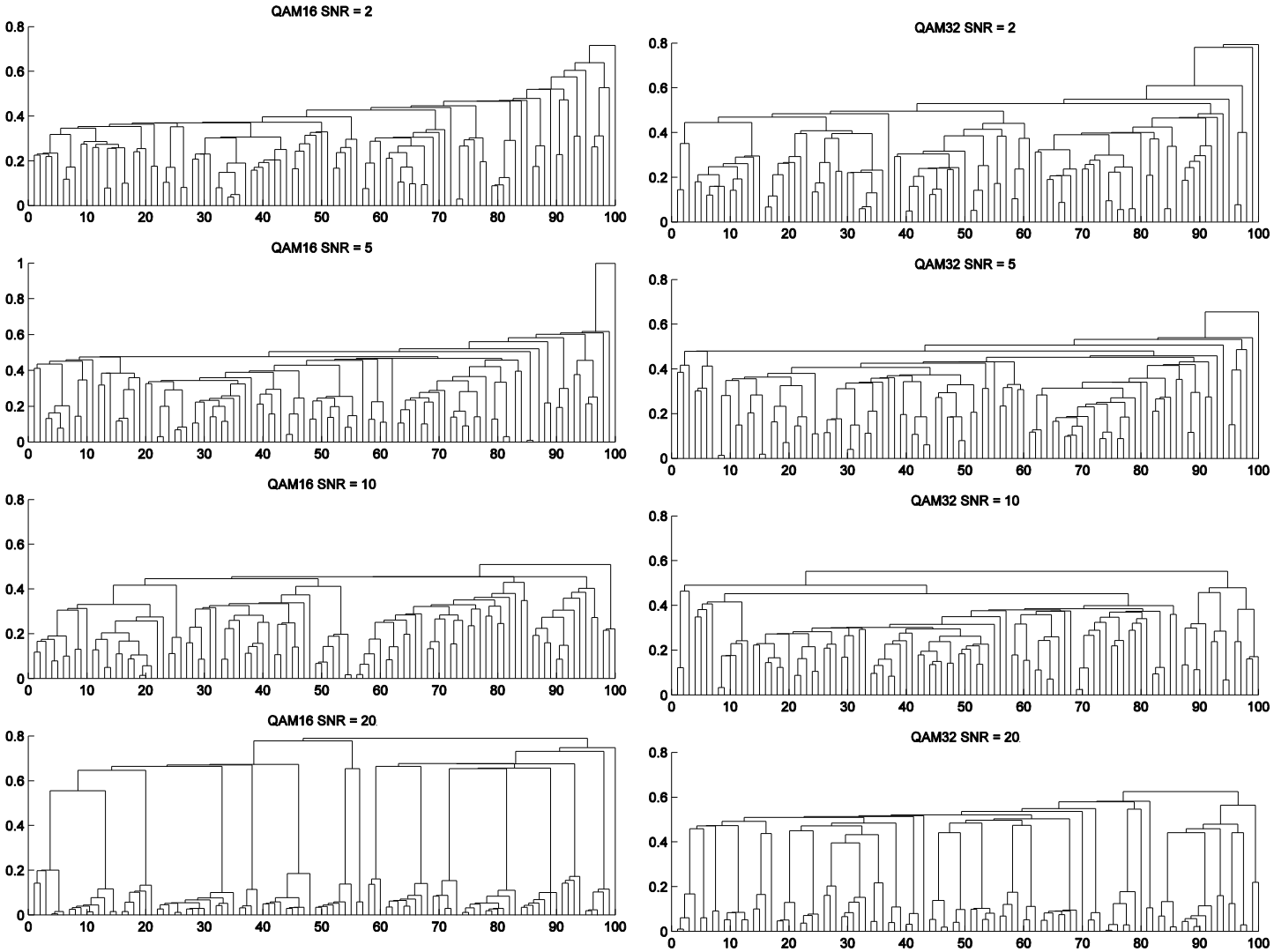








Classification of Modulations



Summary

Density-based clustering methods

Definitions

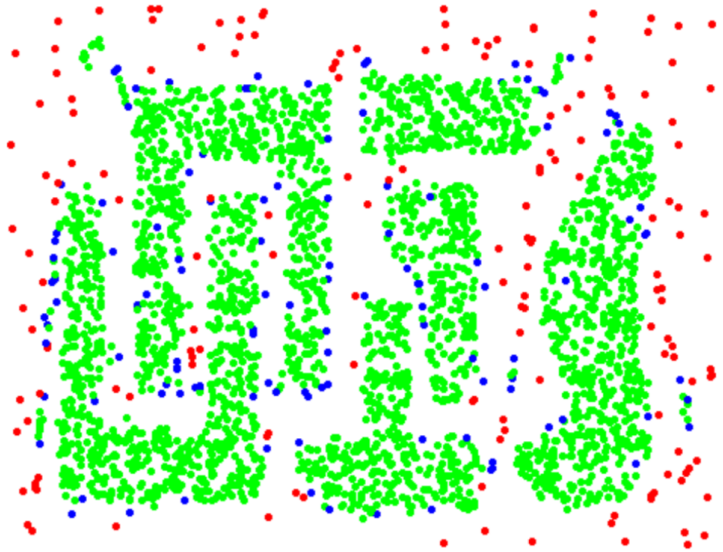
- Neighborhood of a point $p \in \boldsymbol{X}$

$$N_\epsilon(p) = \{q \in \boldsymbol{X} | d(p,q) \leq \epsilon\}$$

- Core point: a point p is a core point if it has more than MinPts neighbors in $N_\epsilon(p)$.
- Border point: a point p is a border point if it has less than MinPts neighbors in $N_\epsilon(p)$ and if it is in the neighborhood of a core point
- Noise point: a noise point is a point that is not a core point nor a border point

Core, border and noise points

Original Points



Point types: core, border and outliers

$\varepsilon = 10$, MinPts = 4

Thanks to Jing Gao from SUNY Buffalo university for her slides!!

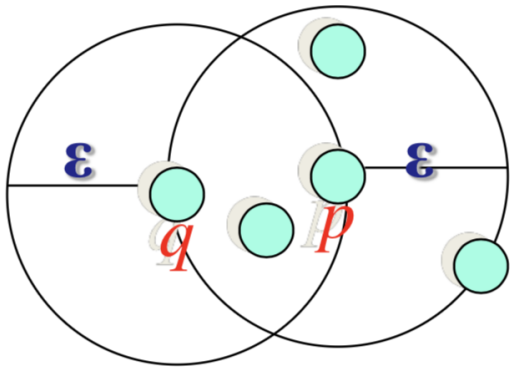
Relationships between different points

Directly density-reachable points

A point q is directly reachable from p for $(\epsilon, \mathrm{MinPts})$ if

- $q \in N_\epsilon(p)$
- p is a core point, i.e., $N_\epsilon(p)$ contains more than MinPts points.

Example



$\mathrm{MinPts} = 4$

- q is directly density-reachable from p
- p is not directly density-reachable from q (q is not a core point)

It is an asymmetric relationship!

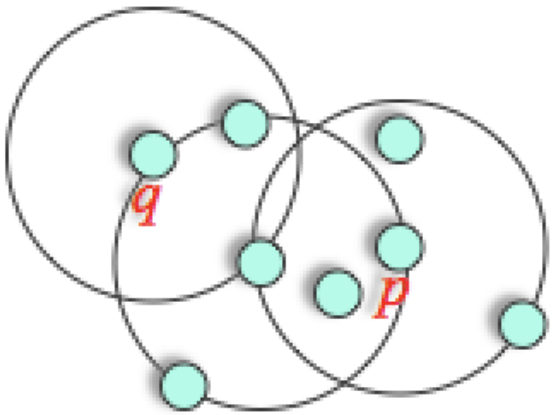
Relationships between different points

Density-reachable points

A point q is density-reachable from p for $(\epsilon, \mathrm{MinPts})$ if there is chain of points p_i such that

- $p_1 = p$ and $p_N = q$
- each point p_{i+1} is directly density-reachable from p_i.

Example



- q is density-reachable from p
- p is not density-reachable from q (q is not a core point)

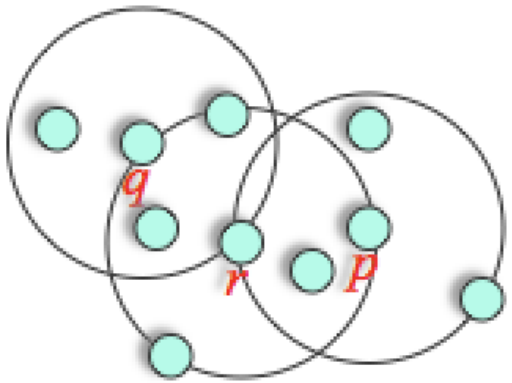
It is an asymmetric relationship!

Relationships between different points

Density-connectivity

Points p and q are density-connected for $(\epsilon, \mathrm{MinPts})$ if there is an object r such that both p and q are density reachable from r.

Example



It is a symmetric relationship!

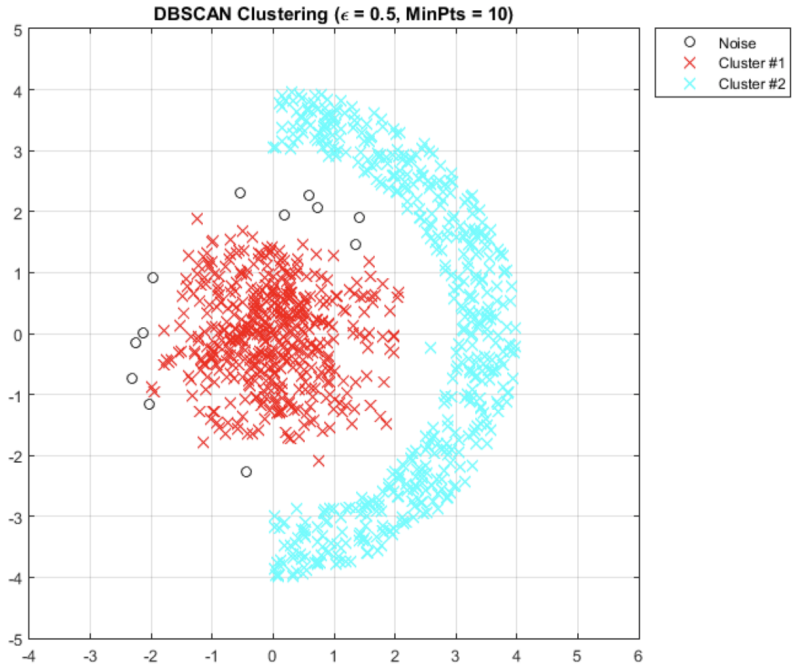
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Cluster definition

A cluster C is defined as a maximal set of density-connected points

- Maximality: if $p \in C$ and if q is density-reachable from p, then $q \in C$.
- Connectivity: for all $(p, q) \in C$, p and q are density-connected.

Example



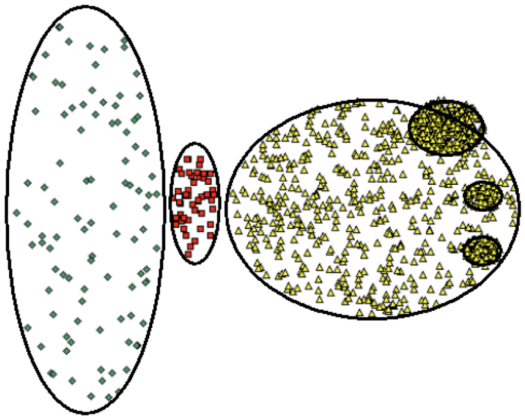
DBSCAN

A simple algorithm

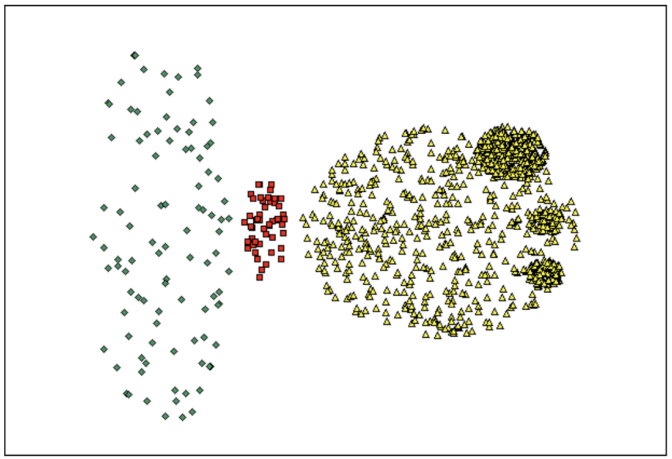
- Select a point p
- Determine all density-reachable points from p for $(\epsilon, \mathrm{MinPts})$.
- If p is a core point, i.e., if the cardinal of $N_\epsilon(p)$ is larger than MinPts, a cluster is formed
- If p is a border point, DBSCAN visits the next point
- Continue the procedure until all points have been visited

The DBSCAN algorithm generally provides a result independent of the order the points have been processed.

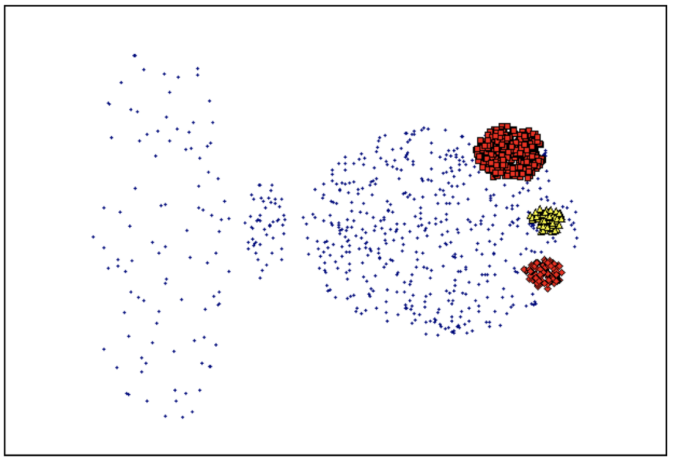
How can we choose ϵ and MinPts?



Original Points



(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)

- **Cannot handle varying densities**
- **sensitive to parameters—hard to determine the correct set of parameters**

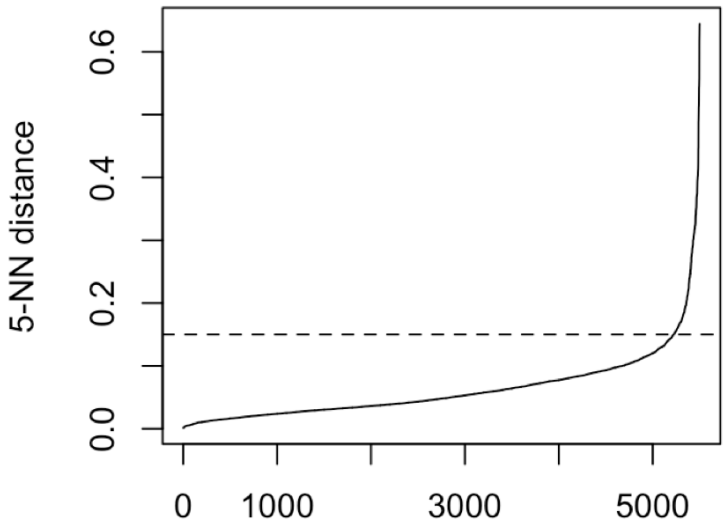
Thanks to Jing Gao from SUNY Buffalo university for her slides!!

How can we choose ϵ and MinPts?

Some ideas

- Points belonging to a cluster have a distance to their k-nearest neighbor (denoted as k-dist) smaller than noise points
- Compute all the k-dist values and sort them in increasing order
- Choose ϵ as the value of k-dist associated with a sharp change in the curve (this value does not vary significantly with the value of k)

Example



Conclusions

Pros

- Clusters of arbitrary shapes
- Robustness to noise

Cons

- Problems with clusters of different densities
- Parameter determination can be difficult

Some references (related to DBSCAN)

- M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, Aug. 1996.
- J. Gao, Slides on Data Mining and Bioinformatics, University at Buffalo, https://www.cse.buffalo.edu/ jing/cse601/fa13/.

Summary

Gaussian mixture (unsupervised learning)

Idea: N unlabelled data vectors of $\mathbb{R}^p$ denoted as $\boldsymbol{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ emerge from K Gaussian components/classes denoted as $\omega_1, \ldots, \omega_K$.

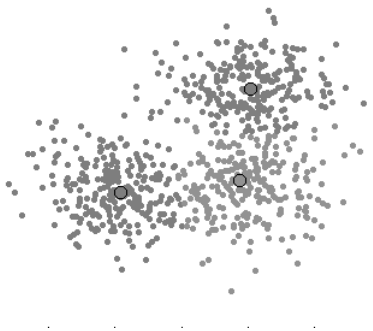
Gaussian mixture model (GMM)

- Definition

$$p(\mathbf{x}|\theta) = \sum_{j=1}^{K} \pi_j p(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{1}$$

- $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_j, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_j, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_j, \ldots, \boldsymbol{\Sigma}_K)$ contains all the parameters of the mixture model.

Example: $K = 3$ seems reasonable ($\boldsymbol{\theta}$ is unknown)



Gaussian mixture (supervised learning)

When the N data $\boldsymbol{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ are *complete* (i.e. labelled, assigned to classes), the parameter estimation problem is straightforward (each Gaussian can be estimated separately). Besides the data, we also have their labels: $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ where $y_i \in \{\omega_1, \ldots, \omega_K\}$.

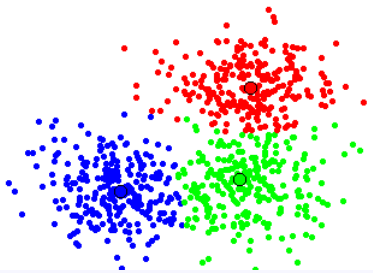
Gaussian mixture model (GMM)

- Definition

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j p(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{2}$$

- Assignments: binary variable $\delta(j|i)$ assigns data $\mathbf{x}_i$ to the jth Gaussian (class ω_j) if $\delta(j|i) = 1$ ($\delta(j|i) = 0$ otherwise).

Example: $K = 3$ is now given (along with data assignments).



Gaussian mixture (supervised learning)

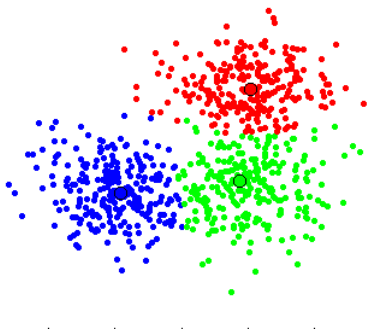
When the N data $\boldsymbol{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ are *complete* (i.e., labelled or assigned to classes), the parameter estimation problem is straightforward (each Gaussian density can be estimated separately).

Estimation of the Gaussian mixture model

- $\widehat{\pi}_j \leftarrow \frac{\widehat{n}_j}{n}$ with $\widehat{n}_j = \sum_{i=1}^{n} \delta(j|i)$
- $\widehat{\boldsymbol{\mu}}_j \leftarrow \frac{1}{\widehat{n}_j} \sum_{i=1}^{n} \delta(j|i) \boldsymbol{x_i}$
- $\widehat{\boldsymbol{\Sigma}}_j \leftarrow \frac{1}{\widehat{n}_j} \sum_{i=1}^{n} \delta(j|i)(\mathbf{x_i} - \widehat{\boldsymbol{\mu}}_j)(\mathbf{x_i} - \widehat{\boldsymbol{\mu}}_j)^T$

  where $\delta(j|i) = 1$ assigns data $\mathbf{x}_i$ to the jth Gaussian density

Example: $K = 3$ is now given (along with data assignments).



Gaussian mixture (back to unsupervised case)

Without assignment (unsupervised learning), we start from an initial setting of the parameters $\boldsymbol{\theta}$: $\boldsymbol{\theta}^0 = (\pi_1^0, \ldots, \pi_K^0, \boldsymbol{\mu}_1^0, \boldsymbol{\Sigma}_1^0, \ldots, \boldsymbol{\mu}_K^0, \boldsymbol{\Sigma}_K^0)$ and compute

$$P(y_i = \omega_j | \boldsymbol{x_i}, \boldsymbol{\theta}^0) = \frac{p(\boldsymbol{x_i} | y_i = \omega_j, \boldsymbol{\theta}^0) P(\omega_j)}{\sum_j p(\boldsymbol{x_i} | y_i = \omega_j, \boldsymbol{\theta}^0) P(\omega_j)} \tag{3}$$

$$P(y_i = \omega_j | \boldsymbol{x_i}, \boldsymbol{\theta}^0) = \frac{\pi_j^0 p(\boldsymbol{x_i} | \boldsymbol{\mu}_j^0, \boldsymbol{\Sigma}_j^0)}{\sum_j \pi_j^0 p(\boldsymbol{x_i} | \boldsymbol{\mu}_j^0, \boldsymbol{\Sigma}_j^0)} \tag{4}$$

Soft assignment

- $\widehat{\delta}(j|i) = P(y_i = \omega_j | \boldsymbol{x_i}, \boldsymbol{\theta}^0)$
- $\sum_{j=1}^K \widehat{\delta}(j|i) = 1.$

Expectation-Maximization (EM-Algorithm for GMM estimation)

Without assignment (unsupervised case), we start from $\boldsymbol{\theta}^0$ and iterate soft-assignments that "complete the incomplete data" (Expectation-step) and parameter refinement (Maximisation-step). We can show that the new setting of the parameters $\boldsymbol{\theta}^{(k+1)}$ maximises the log-likelihood of the "completed" data.

EM-algorithm:

1. **Initialization** Specify $\boldsymbol{\theta}^{(k=0)}$.
2. **Repeat**

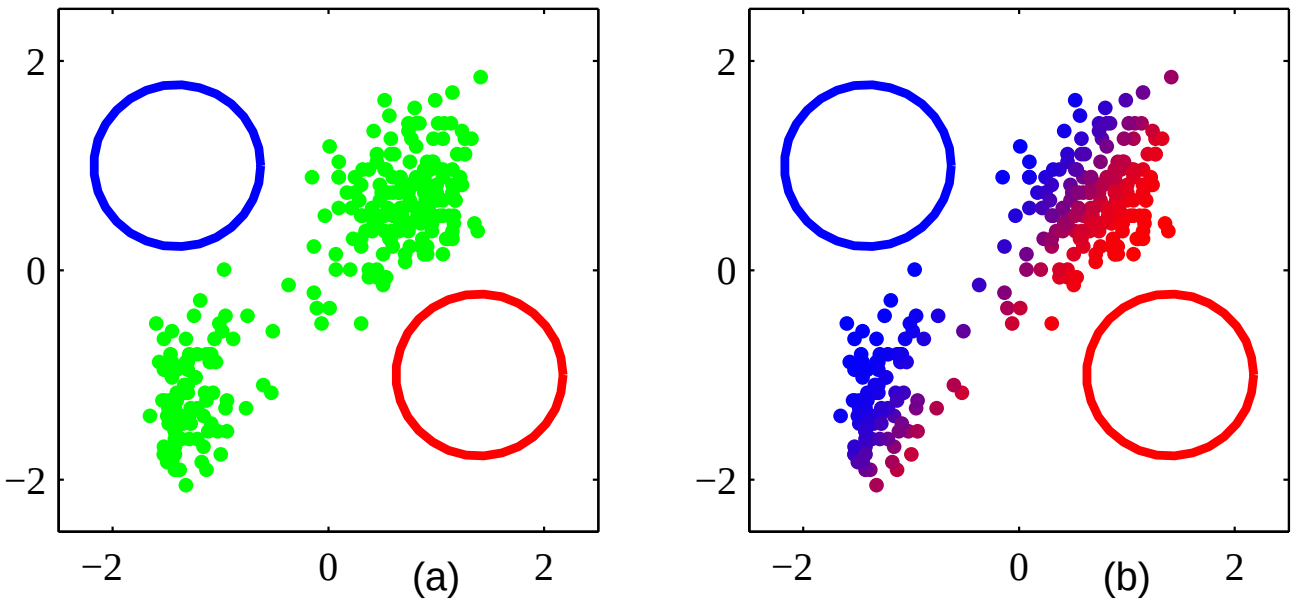
(E-step) soft-assignments of $\mathbf{x_i}$ $\forall i$

$$\widehat{\delta}(j|i) \leftarrow P(y_i = \omega_j | \mathbf{x_i}, \boldsymbol{\theta}^{(k)}) \tag{5}$$

(M-step) Refine $\theta^{(k+1)}$:

- $\widehat{\pi}_j \leftarrow \frac{\widehat{n}_j}{n}$ with $\widehat{n}_j = \sum_{i=1}^{n} \widehat{\delta}(j|i)$
- $\widehat{\boldsymbol{\mu}}_j \leftarrow \frac{1}{\widehat{n}_j} \sum_{i=1}^{n} \widehat{\delta}(j|i) \mathbf{x_i}$
- $\widehat{\boldsymbol{\Sigma}}_j \leftarrow \frac{1}{\widehat{n}_j} \sum_{i=1}^{n} \widehat{\delta}(j|i)(\mathbf{x_i} - \widehat{\boldsymbol{\mu}}_j)(\mathbf{x_i} - \widehat{\boldsymbol{\mu}}_j)^T$
- $k \leftarrow k + 1$

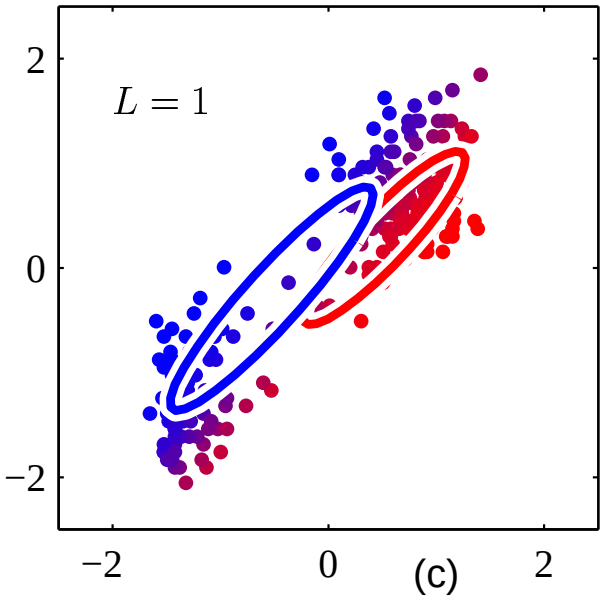
EM-Algorithm for GMM estimation (demo from book by Bishop*)



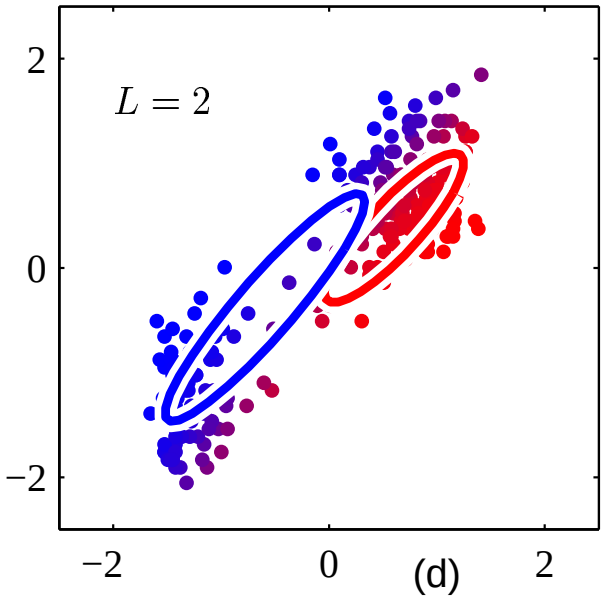
(Initialisation) (E)

* Christopher Bishop, *Pattern Recognition and Machine Learning*. New-York: Springer Verlag, 2006.

EM-Algorithm for GMM estimation (demo 2/3)

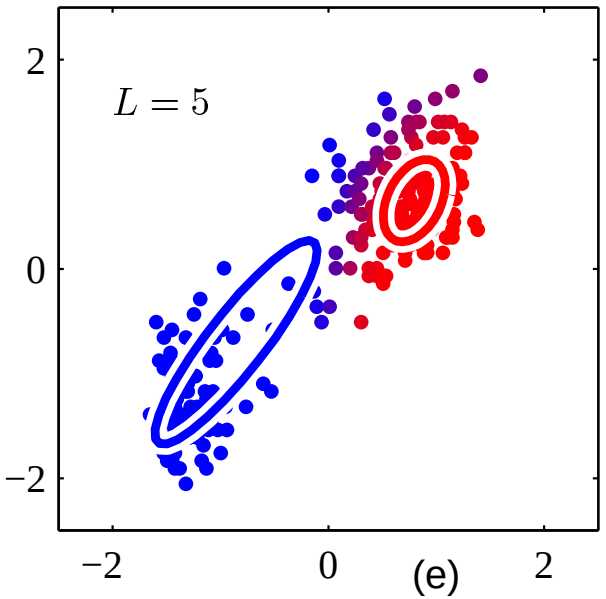


(M)

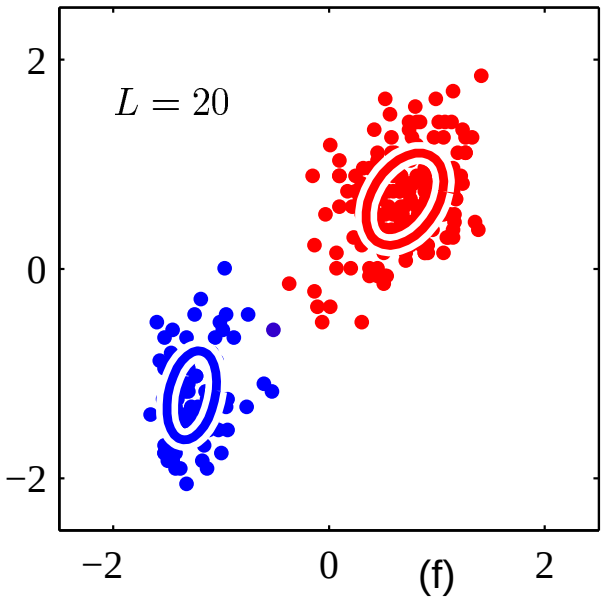


(Iteration $k = 2$)

EM-Algorithm for GMM estimation (demo 3/3)



(Iteration $k = 5$)



(Iteration $k = 20$)

Convergence

- The EM algorithm monotonically increases the log-likelihood of the data.
- We have $l(\boldsymbol{\theta}^0) < l(\boldsymbol{\theta}^1) < ... < l(\boldsymbol{\theta}^k)$ with $l(\boldsymbol{\theta}^k) = \sum_{i=1}^{n} \ln p(\boldsymbol{x_i}|\boldsymbol{\theta}^k)$

Mélange de lois AGGD

Vraisemblance du mélange

$$p(x|\theta) = \sum_{j=1}^{M} p_j p(x|\theta_j)$$

avec

$$p(x|\theta_j) = \begin{cases} \frac{\delta_j^{1/\lambda_j}}{\gamma_j^{1/\lambda_j}\Gamma(1+1/\lambda_j)} \exp\left(-\frac{\delta_j}{\gamma_j}\left(\frac{\mu_j - x}{\alpha_j}\right)^{\lambda_j}\right) & \text{si } x < \mu_j \\ \frac{\delta_j^{1/\lambda_j}}{\gamma_j^{1/\lambda_j}\Gamma(1+1/\lambda_j)} \exp\left(-\frac{\delta_j}{\gamma_j}\left(\frac{x - \mu_j}{1-\alpha_j}\right)^{\lambda_j}\right) & \text{si } x \geq \mu_j \end{cases}$$

- α_j : paramètre d'asymétrie
- μ_j : paramètre de décalage
- γ_j : paramètre d'échelle
- λ_j : paramètre de forme
- $\delta_j = \frac{2\alpha_j^{\lambda_j}(1-\alpha_j)^{\lambda_j}}{\alpha_j^{\lambda_j} + (1-\alpha_j)^{\lambda_j}}$
- M : nombre de composantes

Algorithme EM + Binarisation

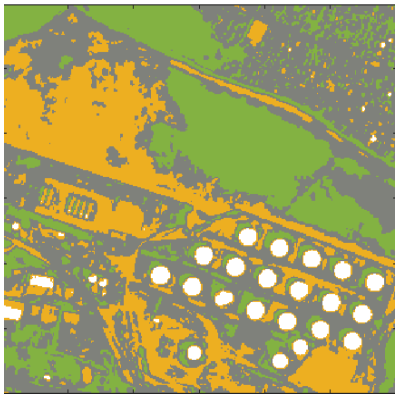
Augmentation des données avec des labels : $\mathbf{x} \rightarrow (\mathbf{x}, \mathbf{Z})$ où $\mathbf{x} = (x^1, \ldots, x^N)$ et $\mathbf{Z} = (\mathbf{z}^1, \ldots, \mathbf{z}^N)$ avec $\mathbf{z}^i$ de dimension M (correspondant aux M classes)
$\Rightarrow$ Log-vraisemblance augmentée

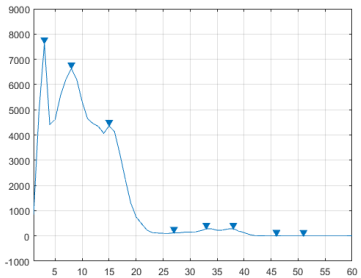
$$L(\mathbf{x}, \mathbf{Z}; \theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} z_m^i \log \left(p_m p(x^i; \theta_m) \right)$$

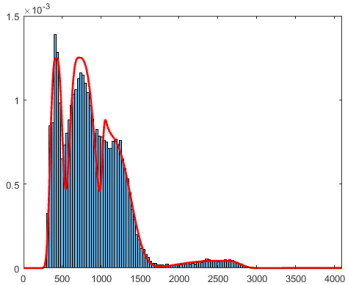
BUT : estimer $\mathbf{p} = (p_1, \ldots, p_M)$ et θ

- **Étape E** *Expectation* : calcul d'une espérance conditionnelle de la log-vraisemblance (fonction Q) afin de mettre à jour les labels $\mathbf{Z}$,
- **Étape B** *Binarization* : seuillage de la matrice des labels $\mathbf{Z}$ en utilisant la loi multinomiale,
- **Étape M** *Maximization* : mise à jour des paramètres par maximisation de la vraisemblance de chaque classe (ML).

Algorithme 2







19 / 78

Summary

Linear discriminant functions

Bayesian classifier (2 classes)

$$d(\boldsymbol{x}) = a_1 \Leftrightarrow \widehat{f}(\boldsymbol{x}\,|\omega_1)\,P(\omega_1) \geq \widehat{f}(\boldsymbol{x}\,|\omega_2)\,P(\omega_2)$$

i.e.

$$\begin{aligned} d(\boldsymbol{x}) &= a_1 \Leftrightarrow g(\boldsymbol{x}) > 0 \\ d(\boldsymbol{x}) &= a_2 \Leftrightarrow g(\boldsymbol{x}) < 0 \end{aligned}$$

where $g(\boldsymbol{x}) = \widehat{f}(\boldsymbol{x}\,|\omega_1)\,P(\omega_1) - \widehat{f}(\boldsymbol{x}\,|\omega_2)\,P(\omega_2)$ is a discriminant function (in general it is non-linear)

Linear discriminant functions (2 classes)

$$\begin{aligned} d(\boldsymbol{x}) &= a_1 \Leftrightarrow g(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} > 0 \\ d(\boldsymbol{x}) &= a_2 \Leftrightarrow g(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} < 0 \end{aligned}$$

with $\boldsymbol{w} = (w_1, ..., w_{p-1}, w_p)^t$ and $\boldsymbol{x} = (x_1, ..., x_{p-1}, 1)^t$

Pb: how can we choose $\boldsymbol{w}$?

The perceptron algorithm

Hypotheses

Two linearly separable classes: there exists $\boldsymbol{w}^*$ such that

$$\begin{aligned} (\boldsymbol{w}^*)^t \boldsymbol{x} &> 0 \qquad \forall \boldsymbol{x} \in \omega_1 \\ (\boldsymbol{w}^*)^t \boldsymbol{x} &< 0 \qquad \forall \boldsymbol{x} \in \omega_2 \end{aligned}$$

Determination of w^*

Optimization of a cost function

$$J(\boldsymbol{w}) = -\sum_{\boldsymbol{x} \in Y} \delta_{\boldsymbol{x}} \boldsymbol{w}^t \boldsymbol{x}$$

with

$$\delta_{\boldsymbol{x}} = 1 \text{ if } \boldsymbol{x} \in \omega_1, \quad \delta_{\boldsymbol{x}} = -1 \text{ if } \boldsymbol{x} \in \omega_2$$

and Y is the set of vectors of the training set that are misclassified by the following rule

$$d(\boldsymbol{x}) = a_1 \text{ if } \boldsymbol{w}^t \boldsymbol{x} > 0, \quad d(\boldsymbol{x}) = a_2 \text{ if } \boldsymbol{w}^t \boldsymbol{x} < 0$$

Properties

- $J(\boldsymbol{w}) \geq 0$ and $J(\boldsymbol{w}) = 0$ if there is no error
- The cost function is continuous piecewise linear and is not defined at discontinuity points

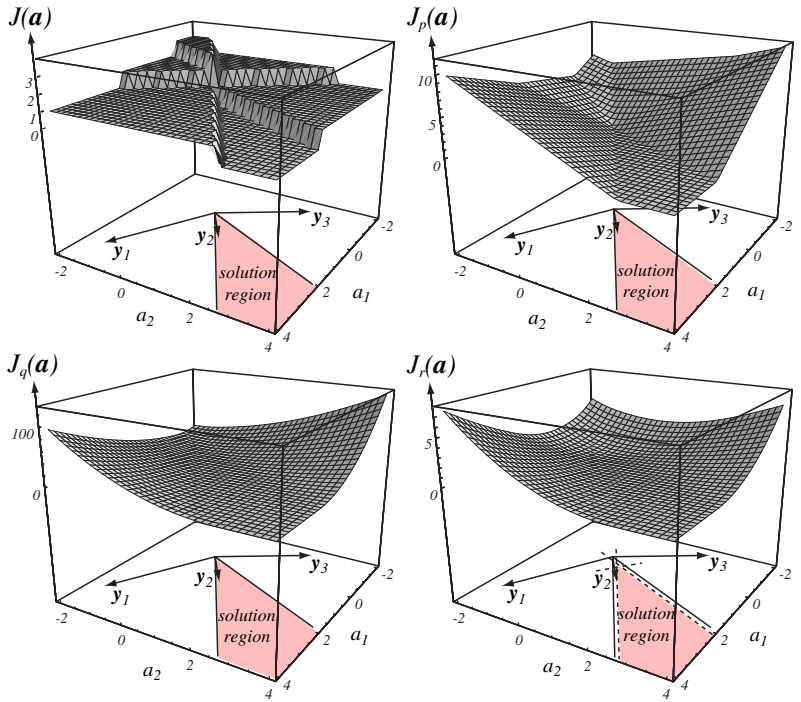


FIGURE 5.11. Four learning criteria as a function of weights in a linear classifier. At the upper left is the total number of patterns misclassified, which is piecewise constant and hence unacceptable for gradient descent procedures. At the upper right is the Perceptron criterion (Eq. 16), which is piecewise linear and acceptable for gradient descent. The lower left is squared error (Eq. 32), which has nice analytic properties and is useful even when the patterns are not linearly separable. The lower right is the square error with margin (Eq. 33). A designer may adjust the margin b in order to force the solution vector to lie toward the middle of the $b = 0$ solution region in hopes of improving generalization of the resulting classifier. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The steepest descent algorithm

Definition

$\boldsymbol{w}_k(n)$: value of the weight $\boldsymbol{w}_k$ at the nth iteration

$$\boldsymbol{w}_k(n+1) = \boldsymbol{w}_k(n) - \mu \left. \frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}_k} \right|_{\boldsymbol{w}_k = \boldsymbol{w}_k(n)}$$

$\mu > 0$: learning parameter

Update of the weights $\boldsymbol{w}_k$

$$\boldsymbol{w}(n+1) = \boldsymbol{w}(n) + \mu \sum_{\boldsymbol{x} \in Y} \delta_{\boldsymbol{x}} \boldsymbol{x}$$

Remarks

- The rule is repeated until convergence
- Convergence after a finite number of iterations n_0

$$\widehat{w}_k(n_0) = \widehat{\boldsymbol{w}}_k(n_0+1) = \widehat{\boldsymbol{w}}_k(n_0+2) = \ldots$$

when the learning classes ω_1 and ω_2 are linearly separable.

Convergence results

Fixed gain μ

$$\mu \in]0, 2]$$

Variable gain μ_n

$$\sum_{n=0}^{+\infty} \mu(n) = +\infty \text{ and } \sum_{n=0}^{+\infty} \mu^2(n) < +\infty$$

- Fast convergence if $\mu(n)$ is close to 2
- Small residual error if $\mu(n)$ is close to 0

Example: $\mu(n) = \frac{c}{n}$.

Sequential algorithm

The perceptron algorithm (Rosenblatt,1950)

1. Initialize the weights: $\boldsymbol{w}_k(0)$
2. Process a new vector from the learning set
   We choose an input vector $\boldsymbol{x}(n)$ with known class and we define the desired output $d(n)$ as follows

$$d(n) = 1 \text{ if } \boldsymbol{x}(n) \in \omega_1, \quad d(n) = -1 \text{ if } \boldsymbol{x}(n) \in \omega_2$$

3. Compute the network output (f_h is the "sign" function)

$$y(n) = f_h\left(\sum_{i=1}^{p} \boldsymbol{w}_i(n)\boldsymbol{x}_i(n)\right)$$

4. Weight update

$$\widehat{\boldsymbol{w}}_k(n+1) = \widehat{\boldsymbol{w}}_k(n) + \mu \boldsymbol{x}_k(n)\left[d(n) - y(n)\right]$$

   where $e(n) = d(n) - y(n) \in \{-2, 0, 2\}$ is the output error
5. Back to 2) until convergence

Convergence

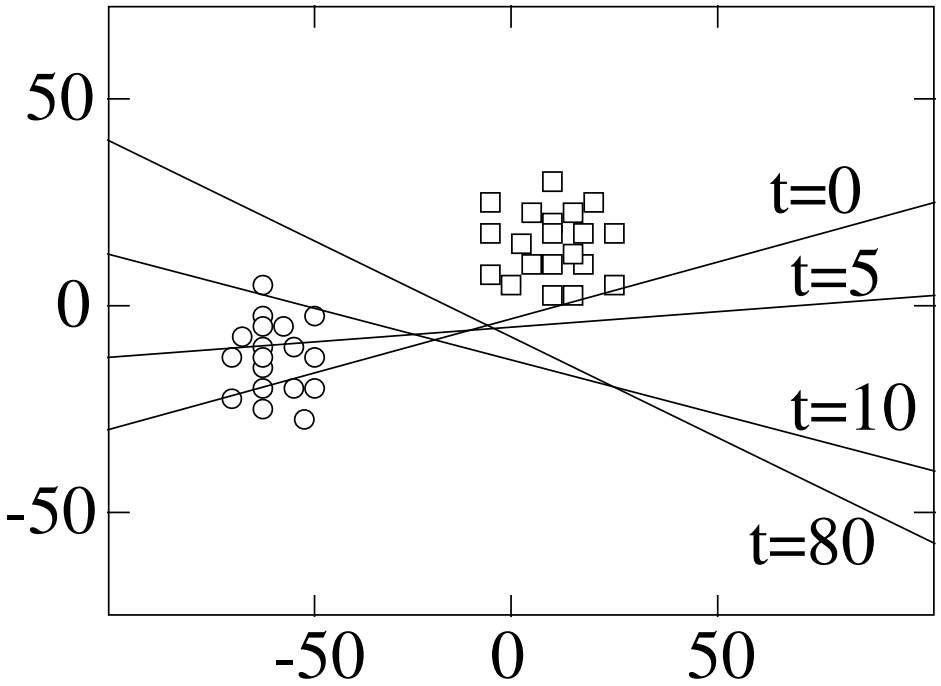
- Convergence after n_0 iterations $(e(n) = 0)$

$$\widehat{\boldsymbol{w}}_k\,(n_0) = \widehat{\boldsymbol{w}}_k\,(n_0+1) = \widehat{\boldsymbol{w}}_k\,(n_0+2) = ...$$

  when the two learning classes ω_1 and ω_2 are linearly separable (separated by a hyperplane)
- Variable learning parameter: $\mu(n)$
- Generalization to $K > 2$ classes: Kesler's construction

Toy example

Example: 1 layer, 80 input samples, $\mu = 0.01$



Other algorithms

Optimal Wiener-Hopf filter

$$J(\boldsymbol{w}) = \frac{1}{2}E\left[e^2(n)\right] = \frac{1}{2}E\left[\left(d(n) - \boldsymbol{w}^t\boldsymbol{x}(n)\right)^2\right]$$

Steepest descent method

$$\boldsymbol{w}_k(n+1) = \boldsymbol{w}_k(n) - \mu \left.\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}_k}\right|_{\boldsymbol{w}_k = \boldsymbol{w}_k(n)}$$

The LMS algorithm

$$\widehat{\boldsymbol{w}}_k(n+1) = \widehat{\boldsymbol{w}}_k(n) + \mu\boldsymbol{x}_k(n)\left[d(n) - y(n)\right]$$

Summary

Support vector machines (SVMs)

Learning set

$$\mathcal{B} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)\}$$

where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are n vectors of $\mathbb{R}^p$ and $y_1, ..., y_n$ are binary variables

$$y_i = 1 \text{ if } \boldsymbol{x}_i \in \omega_1, \quad y_i = -1 \text{ if } \boldsymbol{x}_i \in \omega_2$$

Hyperplane definition

$$g_{\boldsymbol{w},b}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} - b = 0$$

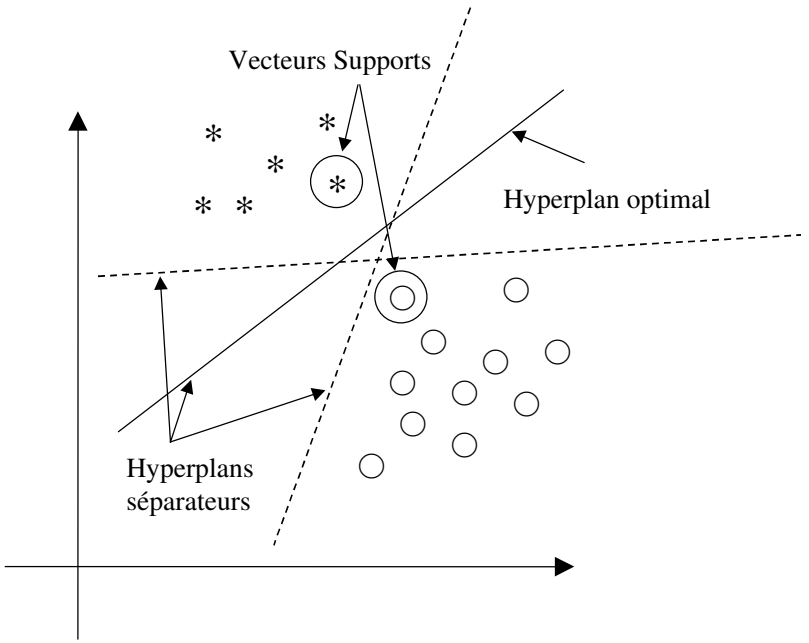
with

$$g_{\boldsymbol{w},b}(\boldsymbol{x}_i) > 0 \text{ if } \boldsymbol{x}_i \in \omega_1, \quad g_{\boldsymbol{w},b}(\boldsymbol{x}_i) < 0 \text{ if } \boldsymbol{x}_i \in \omega_2$$

Classification rule

$$f(\boldsymbol{x}) = \text{sign}\left[g_{\boldsymbol{w},b}(\boldsymbol{x})\right] \tag{6}$$

Illustration



Problem formulation

Margin of $\boldsymbol{x}_i$ with label y_i (algebraic distance to the hyperplane)

$$\gamma_i\left(\widetilde{\boldsymbol{w}}\right) = \frac{y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right)}{\|\boldsymbol{w}\|}$$

with $\widetilde{\boldsymbol{w}} = (\boldsymbol{w}, b)$ ($\boldsymbol{x}_i$ is correctly classified by (6) if $\gamma_i\left(\widetilde{\boldsymbol{w}}\right) > 0$)

Margin of the learning set

$$\gamma_{\mathcal{B}}\left(\widetilde{\boldsymbol{w}}\right) = \min_{i\in\{1,\ldots,n\}} \frac{y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right)}{\|\boldsymbol{w}\|}$$

Since $\gamma_{\mathcal{B}}\left(a\widetilde{\boldsymbol{w}}\right) = \gamma_{\mathcal{B}}\left(\widetilde{\boldsymbol{w}}\right)$, $\forall a > 0$, $\widetilde{\boldsymbol{w}}$ is not unique!

Constraints for the hyperplane: one forces the training samples that are the closest to the hyperplane to satisfy

$$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) = 1 \Rightarrow \min_{i\in\{1,\ldots,n\}} y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) = 1$$

The vectors $\boldsymbol{x}_i$ satisfying $y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) = 1$ are called support vectors.

Problem formulation

Canonical hyperplane

$$y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b \right) \geq 1, \quad \forall i = 1, ..., n$$

Classifier margin for a canonical hyperplane

$$\gamma_{\mathcal{B}} \left(\widetilde{\boldsymbol{w}} \right) = \frac{1}{\|\boldsymbol{w}\|}$$

New formulation

Minimize $\frac{1}{2} \|w\|^2$ **with the constraints** $y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b \right) \geq 1, \forall i$

Simple problem since the cost function to optimize is quadratic and the constraints are linear!

Optimization

Lagrangian

$$L\left(\widetilde{\boldsymbol{w}},\alpha\right)=\frac{1}{2}\boldsymbol{w}^{T}\boldsymbol{w}-\sum_{i=1}^{n}\alpha_{i}\left[y_{i}\left(\boldsymbol{w}^{T}\boldsymbol{x}_{i}-b\right)-1\right]$$

Set to zero the partial derivatives of L with respect to b and w

$$\sum_{i=1}^{n}\alpha_{i}y_{i}=0 \text{ and } \boldsymbol{w}=\sum_{i=1}^{n}\alpha_{i}y_{i}\boldsymbol{x}_{i}$$

Kühn and Tucker multipliers

For a convex optimization problem (convex function $f(x)$ to optimize and convex constraints $G_i\left(\boldsymbol{x}\right)\leq 0$), an optimality condition is the existence of parameters $\alpha_i\geq 0$ such that the Lagrangian derivative is zero, i.e.,

$$L'(\boldsymbol{x})=f'\left(\boldsymbol{x}\right)+\sum_{i=1}^{n}\alpha_{i}G_{i}'\left(\boldsymbol{x}\right)=0$$

with $\alpha_i=0$ if $G_i\left(\boldsymbol{x}\right)<0$ (i.e., $\alpha_i G_i\left(\boldsymbol{x}\right)=0$).

Dual problem

Solve $L'(\boldsymbol{x}) = 0$

$$\boldsymbol{w} = \sum_{\text{Support vectors}} \alpha_i y_i \boldsymbol{x}_i = \boldsymbol{x}^T \boldsymbol{Y} \alpha \tag{7}$$

with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$, $\boldsymbol{x} = (x_1, ..., x_n)^T$, $\boldsymbol{Y} = \text{diag}\,(y_1, ..., y_n)$ and

$$\begin{cases} \alpha_i = 0 \text{ if the constraint is a strict inequality} \\ \alpha_i > 0 \text{ if the constraint is an equality} \end{cases}$$

After replacing the expression of w in the Lagrangian, we obtain

$$U(\alpha) = -\frac{1}{2} \alpha^T \boldsymbol{Y} \left(\boldsymbol{x} \boldsymbol{x}^T \right) \boldsymbol{Y} \alpha + \sum_{i=1}^{n} \alpha_i$$

that has to be maximized in the domain defined by $\alpha_i \geq 0, \forall i$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Remarks

Simple optimization problem

Quadratic (hence convex) function to optimize and linear constraints

Norm of the solution $\boldsymbol{w}_0$

$$\|\boldsymbol{w}_0\|^2 = \sum_{\text{support vectors}} \alpha_i^0 \left(1 + y_i b\right) = \sum_{\text{support vectors}} \alpha_i^0$$

because of the constraint $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Classifier margin

$$\boxed{\gamma = \frac{1}{\|\boldsymbol{w}_0\|} = \left(\sum \alpha_i^0\right)^{-1/2}}$$

Classification rule for a vector $\boldsymbol{x}$

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{z_i \text{ support vectors}} \alpha_i^0 y_i \boldsymbol{x}_i^T \boldsymbol{x} - b_0\right), \quad b_0 = \frac{1}{2}\left(\boldsymbol{w}_0^T \boldsymbol{z}^+ + \boldsymbol{w}_0^T \boldsymbol{z}^-\right),$$

where $\boldsymbol{z}^+$ (resp. $\boldsymbol{z}^-$) is a support vector belonging to the 1st (resp. 2nd) class.

Extensions

The "soft-margin" SVM classifier

$$\text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{with the constraints } y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq 1 - \xi_i, \forall i$$

The ν-SVM classifier

$$\text{Minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \nu\gamma$$
$$\text{with the constraints } y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq \gamma - \xi_i, \forall i$$

Non-linear preprocessing

Search a non-linear transformation $\phi : \mathbb{R}^p \to \mathbb{R}^q$ ensuring linear separability.

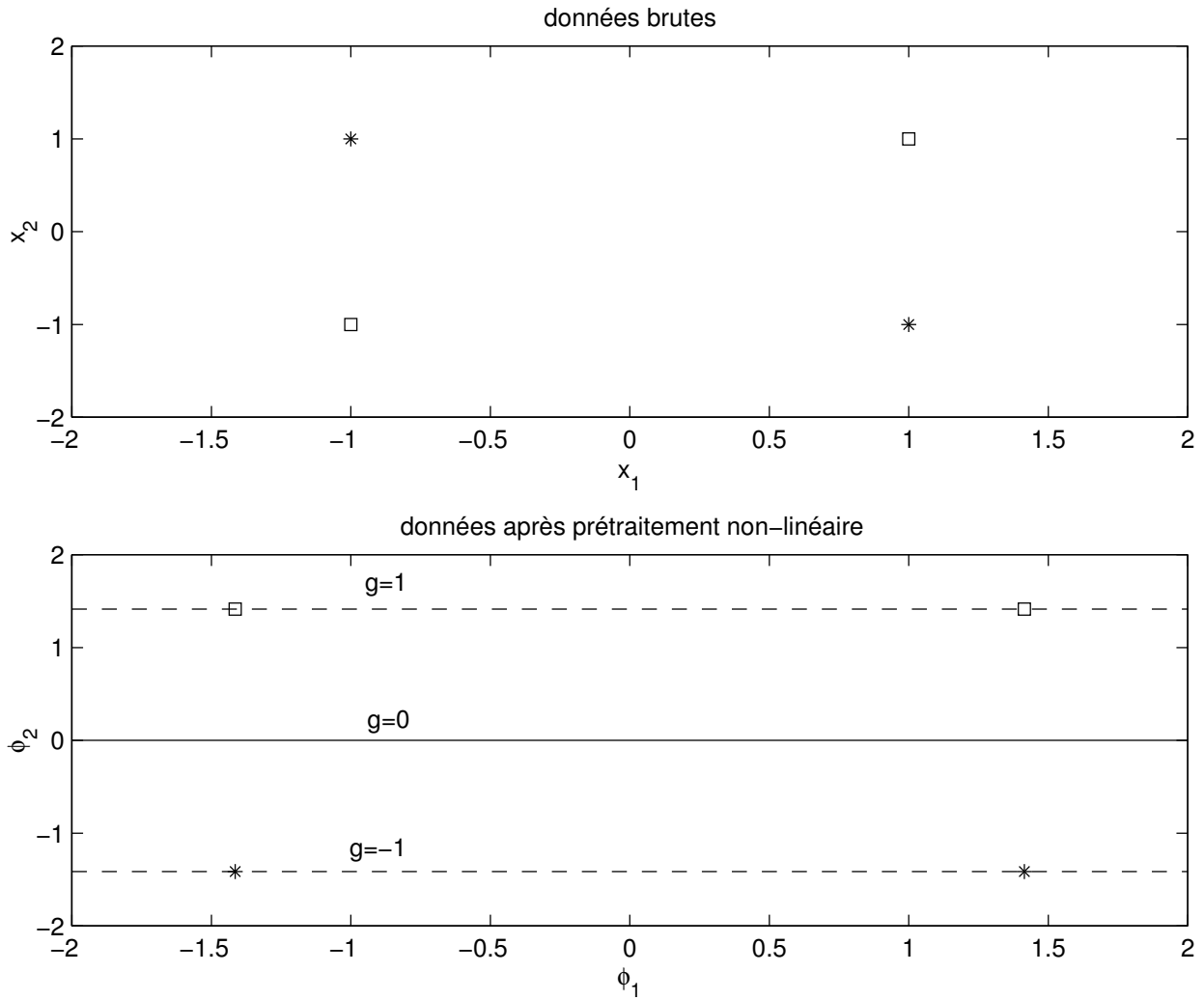
Classical example

The classe classes $\chi_1 = \{(1,1), (-1,-1)\}$ and $\chi_2 = \{(1,-1), (-1,1)\}$ are not linearly separable. Consider the application ϕ defined by

$$\phi : \begin{array}{l} \mathbb{R}^2 \to \mathbb{R}^6 \\ (x_1, x_2)^T \mapsto \left(\sqrt{2}x_1, \sqrt{2}x_1x_2, 1, \sqrt{2}x_2, x_1^2, x_2^2\right)^T \end{array}$$

The data are separable in the plane (ϕ_1, ϕ_2)

Example of separability



Non-linear SVM classifier

Decision rule after preprocessing

$$\begin{aligned} f(\tilde{\boldsymbol{x}}) &= \text{sign}\left(g_{\boldsymbol{w},b}(\tilde{\boldsymbol{x}})\right) = \text{sign}\left(\boldsymbol{w}^T\phi\left(\boldsymbol{x}\right) - b\right) \\ &= \text{sign}\left(\sum_{\text{support vectors}} \alpha_i^0 y_i \phi\left(\boldsymbol{x}_i\right)^T \phi\left(\boldsymbol{x}\right) - b_0\right) \end{aligned}$$

Cost function to optimize

$$\boxed{U\left(\boldsymbol{\alpha}\right) = -\tfrac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Y}\boldsymbol{G}\boldsymbol{Y}^T\alpha + \textstyle\sum_{i=1}^n \alpha_i}$$

where $\boldsymbol{G}$ is the Gram matrix defined by $G_{ij} = \phi\left(\boldsymbol{x}_i\right)^T\phi\left(\boldsymbol{x}_j\right)$. The cost function $U\left(\alpha\right)$ has to be maximized under the constraints

$$\boxed{0 \leq \alpha_i \leq \tfrac{1}{n},\ \forall i} \quad \text{and} \quad \boxed{\textstyle\sum_{i=1}^n \alpha_i \geq \nu}$$

Conclusions: the cost function and the decision rule only depend on the inner products $\phi\left(\boldsymbol{x}_i\right)^T\phi\left(\boldsymbol{x}_j\right)$ and $\phi\left(\boldsymbol{x}_i\right)^T\phi\left(\boldsymbol{x}\right)$.

Kernels

A kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{y})$ allows inner products $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ and $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x})$ to be computed with a reduced computational cost.

Example: $\boldsymbol{x} = (x_1, x_2)^T$ and $\phi(\boldsymbol{x}) = \left(x_1^2, x_2^2, \sqrt{2}x_1 x_2\right)$

$$\begin{aligned} \phi(\boldsymbol{x})^T \phi(\boldsymbol{y}) &= \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1 x_2 \end{pmatrix} . \begin{pmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}y_1 y_2 \end{pmatrix} \\ &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2 \\ &= \left(\boldsymbol{x}^T \boldsymbol{y}\right)^2 \end{aligned}$$

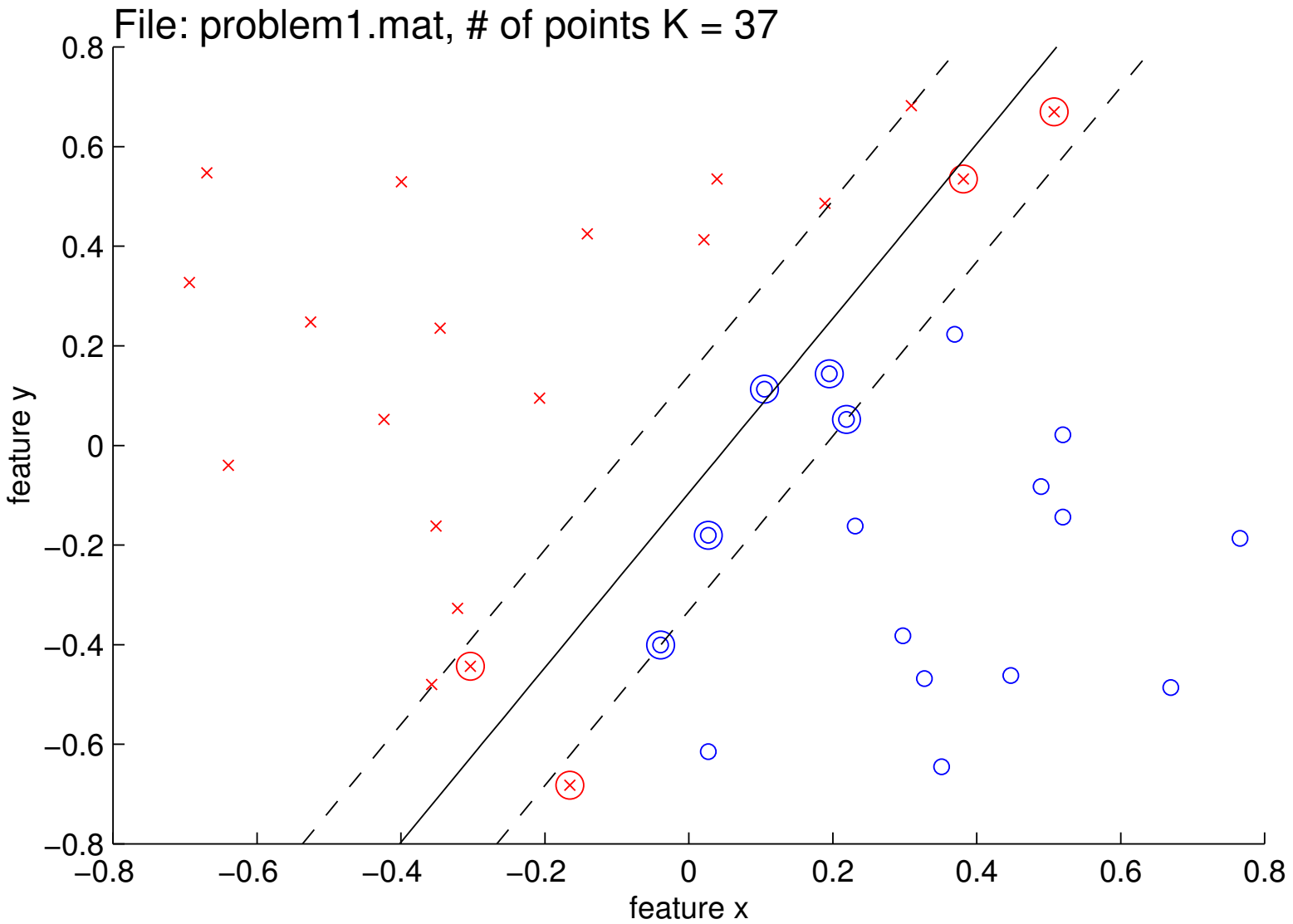
Mercer kernels can be expressed as inner products

$$k(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{y})$$

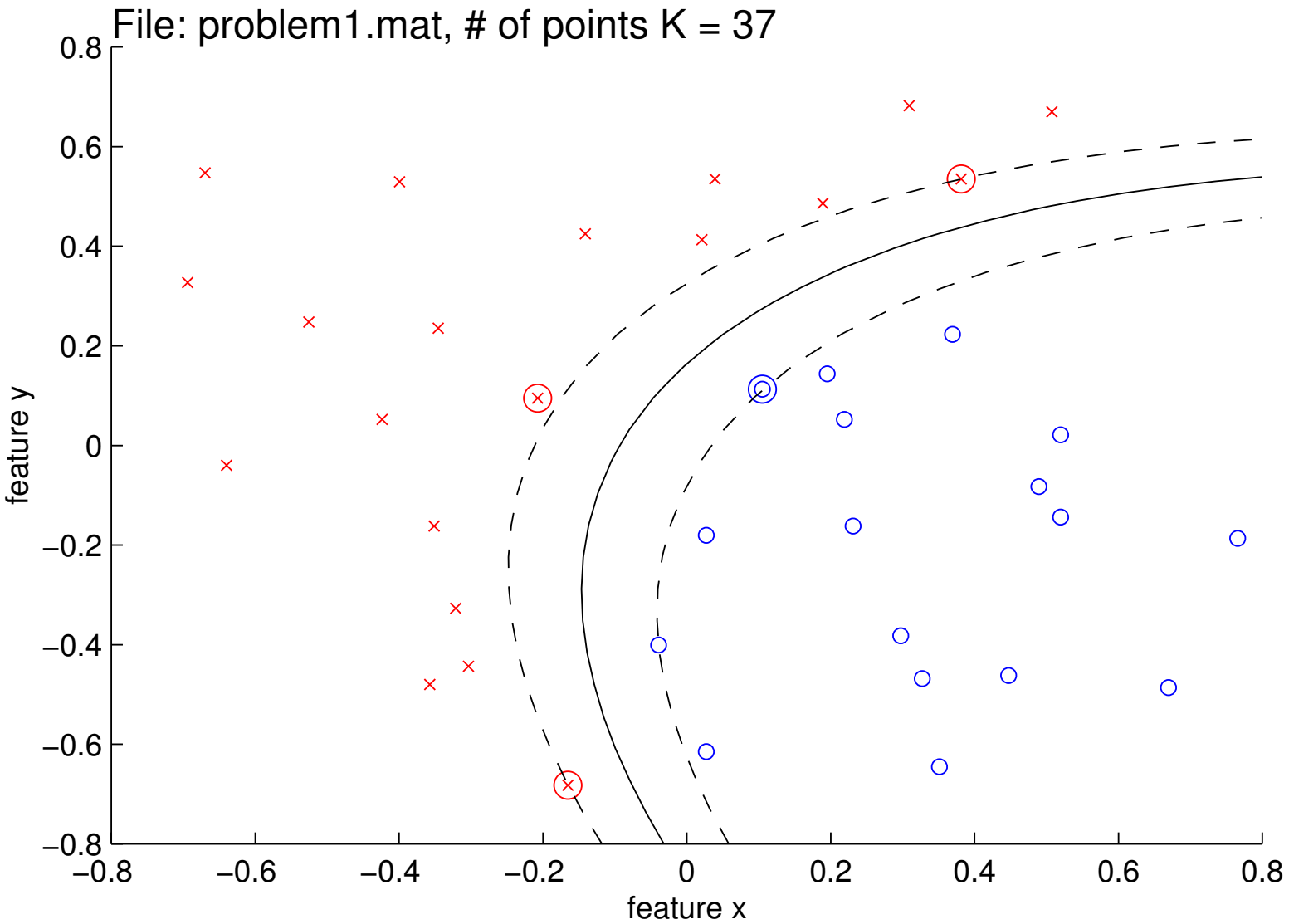
Classical kernels

Kernel	Expression
Polynomial (of degree p)	$k(\boldsymbol{x},\boldsymbol{y}) = (\langle \boldsymbol{x},\boldsymbol{y}\rangle)^q$ $q \in \mathbb{N}^+$
Full polynomial	$k(\boldsymbol{x},\boldsymbol{y}) = (\langle \boldsymbol{x},\boldsymbol{y}\rangle + c)^q$ $c \in \mathbb{R}^+$, $q \in \mathbb{N}^+$
RBF	$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)$ $\sigma \in \mathbb{R}^+$
Mahalanobis	$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left[-(\boldsymbol{x}-\boldsymbol{y})^T \Sigma(\boldsymbol{x}-\boldsymbol{y})\right]$ $\Sigma = \text{diag}\left(\frac{1}{\sigma_1^2},...,\frac{1}{\sigma_p^2}\right)$, $\sigma_i \in \mathbb{R}^+$

Example of linear SVMs

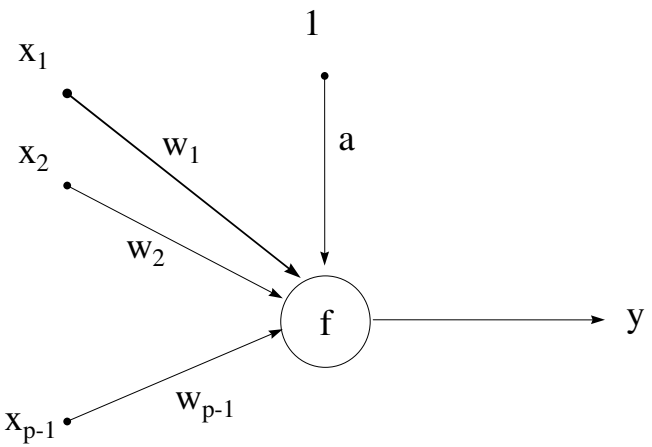


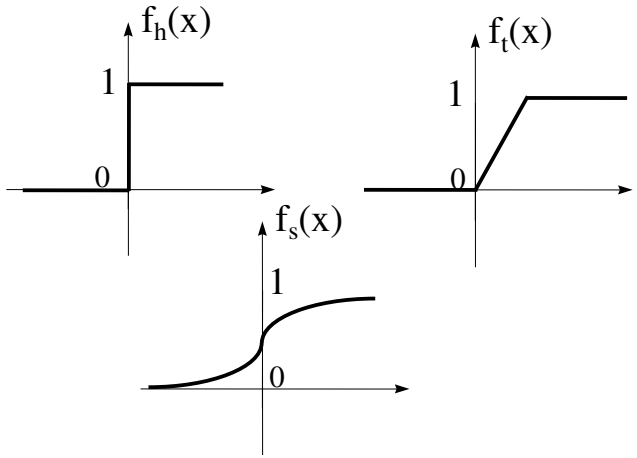
Example of non-linear SVMs



Summary

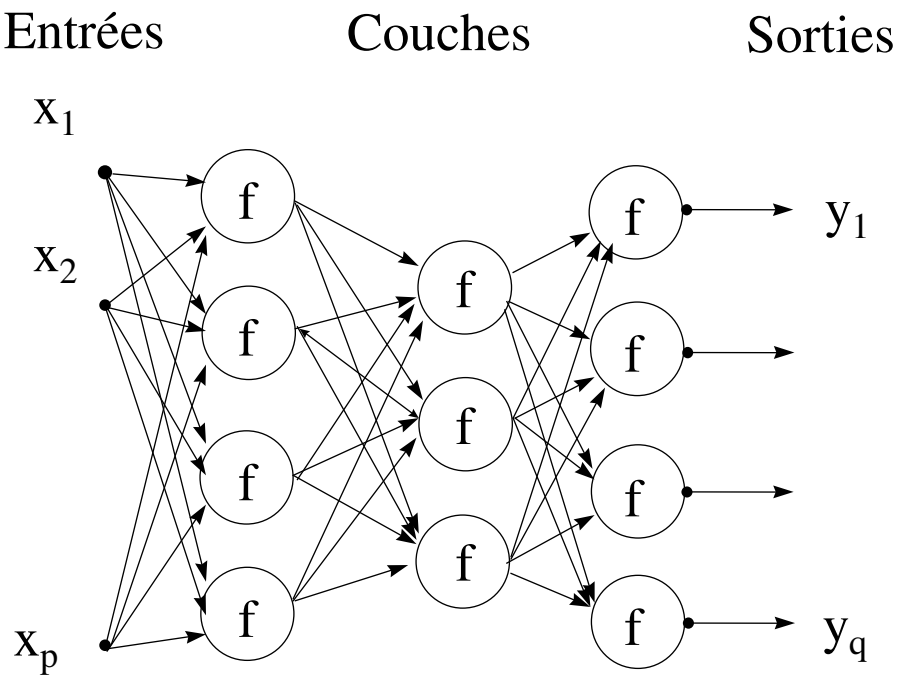
Neural networks





- inputs: $x_1, ..., x_{p-1}$
- weights: $w_1, ..., w_{p-1}$
- bias: a (or offset)
- output : $y = f\left(\sum_{i=1}^{p-1} w_i x_i - a\right)$
- Non-linearity
  - Heaviside: $f_h(x)$
  - Logical threshold: $f_t(x)$
  - Sigmoid: $f_s(x)$
- Equivalent definition: $x_p = -1, w_p = a$ et $y = f\left(\sum_{i=1}^{p} w_i x_i\right)$

Multi-layer perceptron



- Definition of the network structure
  - ☞ Number of outputs?
  - ☞ Number of layers?
  - ☞ Number of nodes per layer?
- Learning rules?

One-layer perceptron with $f_L(x) = x$

- Learning rule
  - ☞ Minimization of $J(w) = \frac{1}{2}E[e^2(n)] \Rightarrow$ Wiener-Hopf filter
  - ☞ Steepest descent algorithm with instantaneous error

$$J_{\text{inst}}(w) = \frac{1}{2}\left[d(n) - f_L\left(\sum_{i=1}^{p} w_i x_i(n)\right)\right]^2$$

- LMS algorithm

$$\widehat{w}_k(n+1) = \widehat{w}_k(n) + \mu e(n) x_k(n)$$

- Properties: convergence of the LMS algorithm to the Wiener-Hopf solution

One-layer perceptron with sigmoidal function

- Learning rule
  - ☞ Minimization of $J(w) = \frac{1}{2}E[e^2(n)] \Rightarrow$ system of non-linear equations (no Wiener-Hopf solution)
  - ☞ Descent algorithm with instantaneous error

$$J_{\text{inst}}(w) = \frac{1}{2}\left[d(n) - f_s\left(\sum_{i=1}^{p} w_i x_i(n)\right)\right]^2$$

- The LMS algorithm

$$\widehat{w}_k(n+1) = \widehat{w}_k(n) + \mu e(n) x_k(n) y(n) [1 - y(n)]$$

- Properties
  - ☞ Convergence to a local minimum of $J(w)$
  - ☞ Same performance obtained with f_h instead of the sigmoid
  - ☞ Generalization (multi-layer network): the back-propagation algorithm

Backpropagation

- Weight Initialization: $w_{ij}^{(l)}(0)$ with footnotesize values
- at iteration n
  - ☞ Compute the desired output

$$d(n) = 1 \text{ if } x(n) \in \omega_1, \quad d(n) = 0 \text{ if } x(n) \in \omega_2$$

  - ☞ Compute the network output

$$y_j^{(l)}(n) = f_s\left(\sum_{i=1}^{N_{l-1}} w_{ij}^{(l)}(n)\, y_i^{(l-1)}(n)\right)$$

where $y_j^{(l)}$ is the jth output of the lth layer (with $y^{(L)}(n) = y(n)$ et $y_j^{(0)}(n) = x_j(n)$), N_l is the number of nodes of the lth layer and L is the number of layers.

  - ☞ Weight update

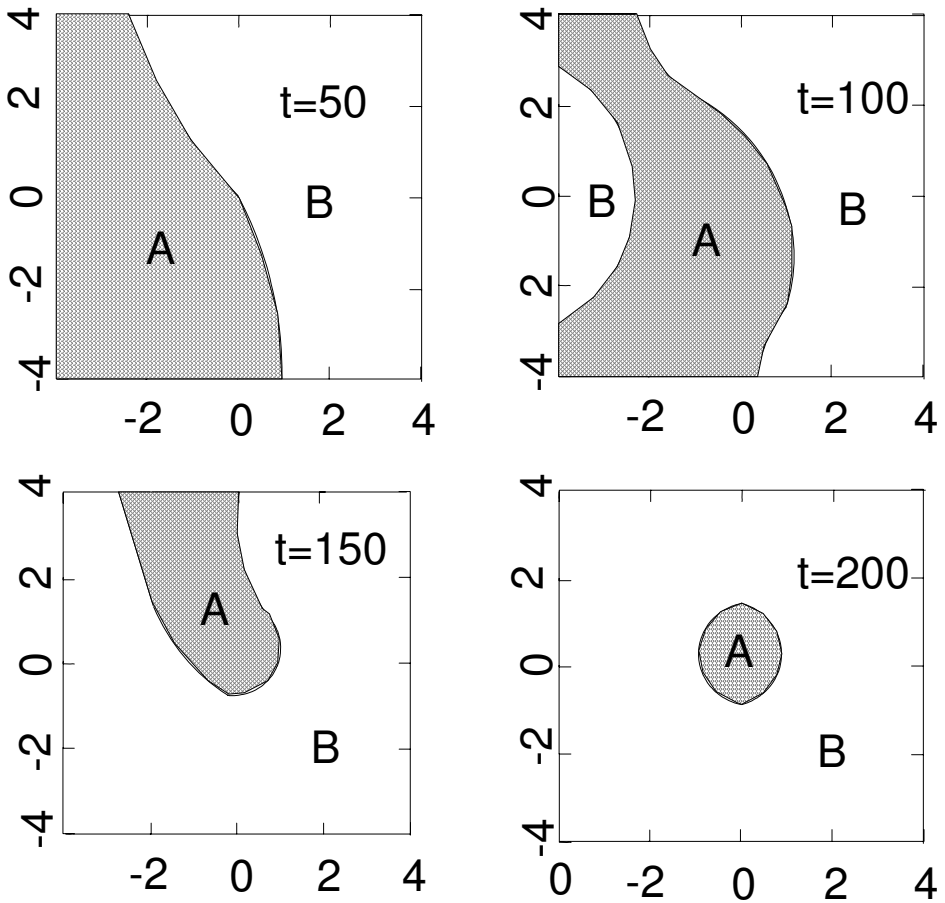
$$\widehat{w}_{ij}^{(l)}(n+1) = \widehat{w}_{ij}^{(l)}(n) + \mu\delta_j^{(l)}(n)\, y_i^{(l-1)}(n)$$

with $\delta^{(L)}(n) = e(n)y(n)[1 - y(n)]$ and

$$\delta_j^{(l)}(n) = y_j^{(l)}(n)\left[1 - y_j^{(l)}(n)\right]\sum_k \delta_k^{(l+1)}(n)\, w_{jk}^{(l+1)}(n)$$

where the summation covers all the errors of the $(l+1)$th layer and $e(n) = d(n) - y(n)$ is the error at the output of the network.

Class 1: $D(0,1)$ Class 2: $D(0,5)\backslash D(0,1)$



Remarks

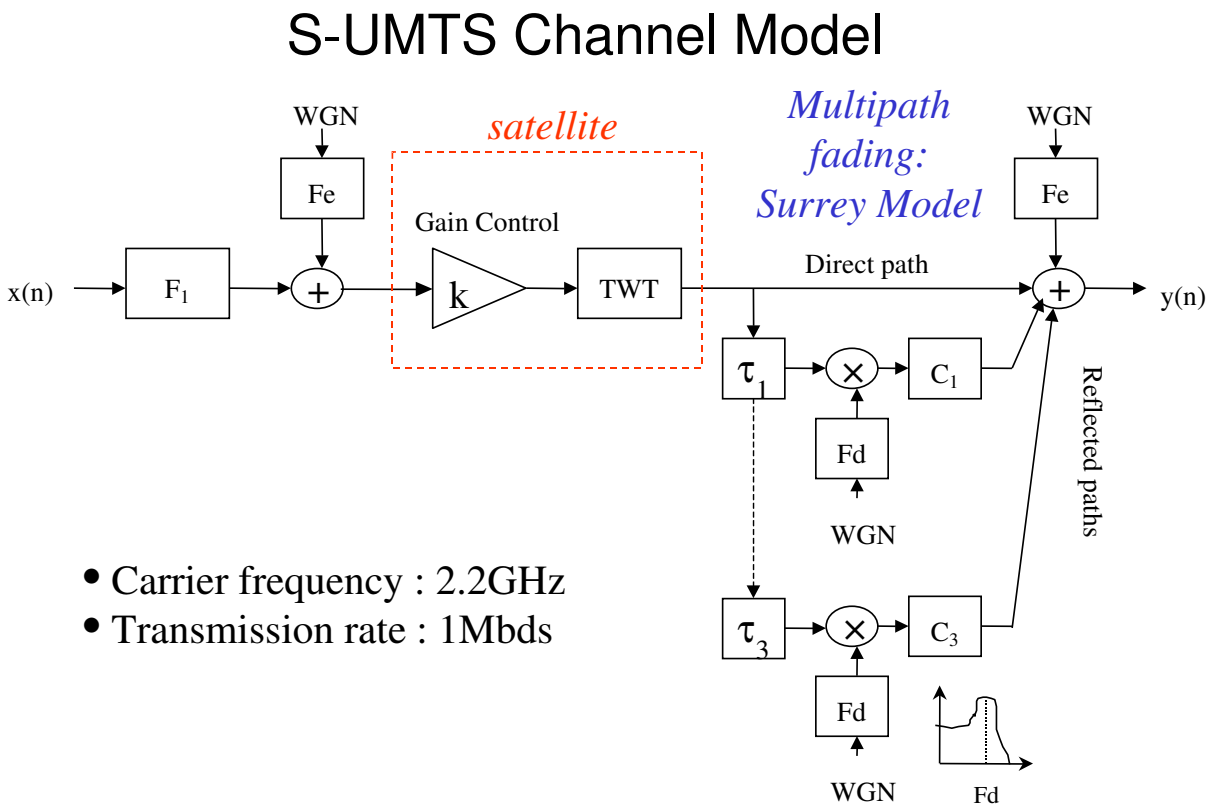
- Epoch
- How should we choose the parameter μ?
  - small μ: slow convergence, small residual error
  - large μ: fast convergence, large residual error
- Classification rule with constant momentum noté $\beta \in [0, 1[$

$$\widehat{w}_{ij}(n+1) = \widehat{w}_{ij}(n) + \Delta\widehat{w}_{ij}(n)$$

$$\Delta\widehat{w}_{ij}(n) = \beta\Delta\widehat{w}_{ij}(n-1) + \mu\delta_j(n)\, y_i(n)$$

  For $\beta = 0$: classical backpropagation rule
  For $\beta > 0$: $\Delta\widehat{w}_{ij}(n) = \mu\sum_{k=0}^{n}\beta^{n-k}\delta_j(n)\, y_i(n)$
  Convergence is ensured if $0 \leq |\beta| \leq 1$.
- The convergence is accelerated when consecutive errors have the same sign whereas it is reduced when two consecutive errors have different signs.
- Stopping rules are problem dependent. For example, in some equalization problems, we use a sequence of known samples (pilot symbols) and the learning algorithm is stopped when $\|$input - output$\| < \varepsilon$
- Universal approximation theorem of Cybenko.

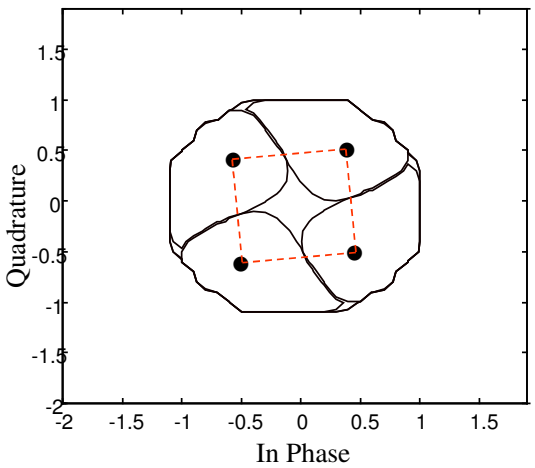
Application to the equalization of non-linear channels



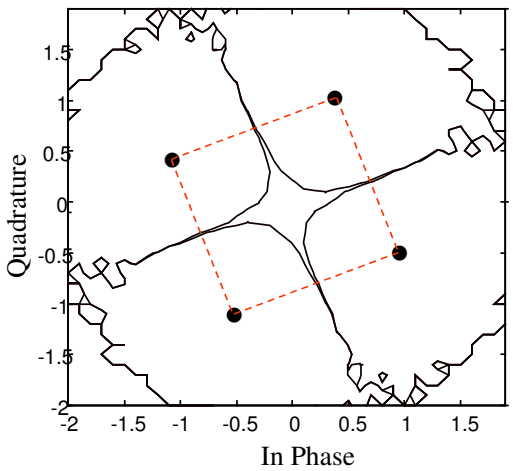
Application to the equalization of non-linear channels

Consequences of the non-linearity

Distortion of a 4-QAM signal



backoff 0dB
no downlink noise

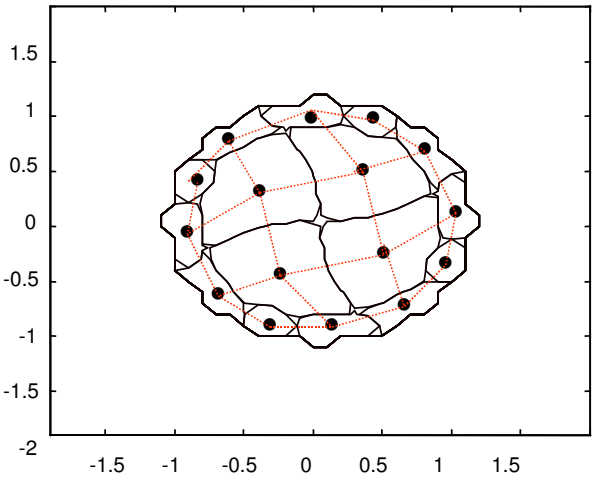


backoff 0dB
15dB downlink SNR

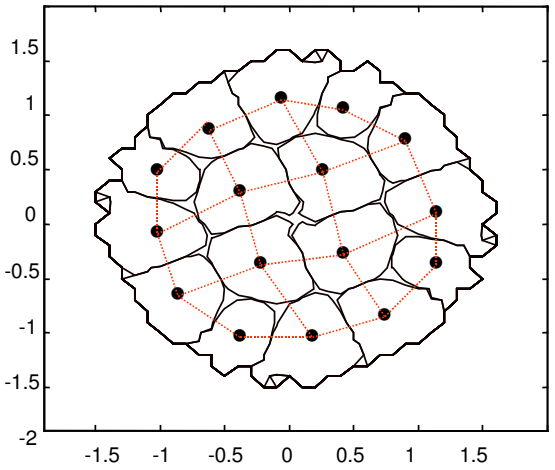
Application to the equalization of non-linear channels

Consequences of the non-linearity

Distortion of a 16-QAM signal

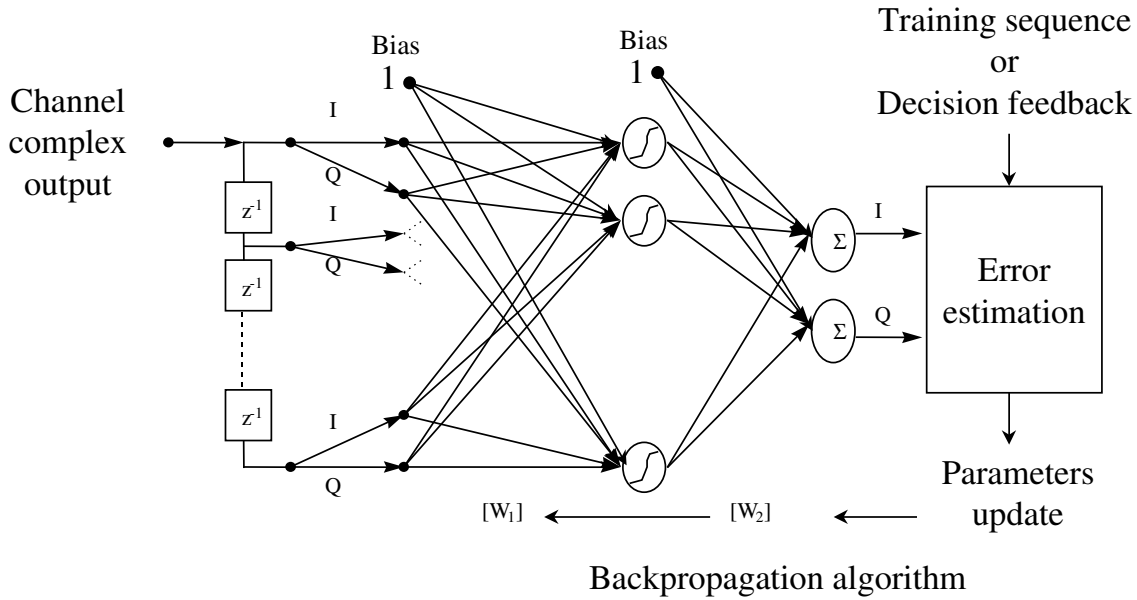


backoff -1dB
no downlink noise



backoff -1dB
15dB downlink noise

Application to the equalization of non-linear channels

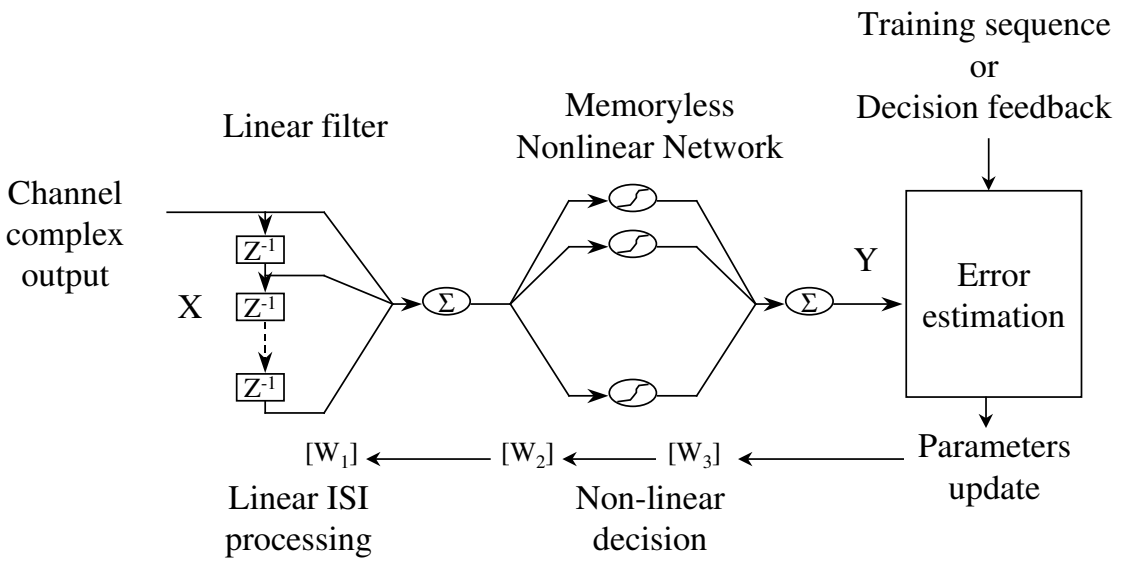


- **Advantages :** universal approximation theorem, parallelism.
- **Drawback :** slow convergence.

Application to the equalization of non-linear channels

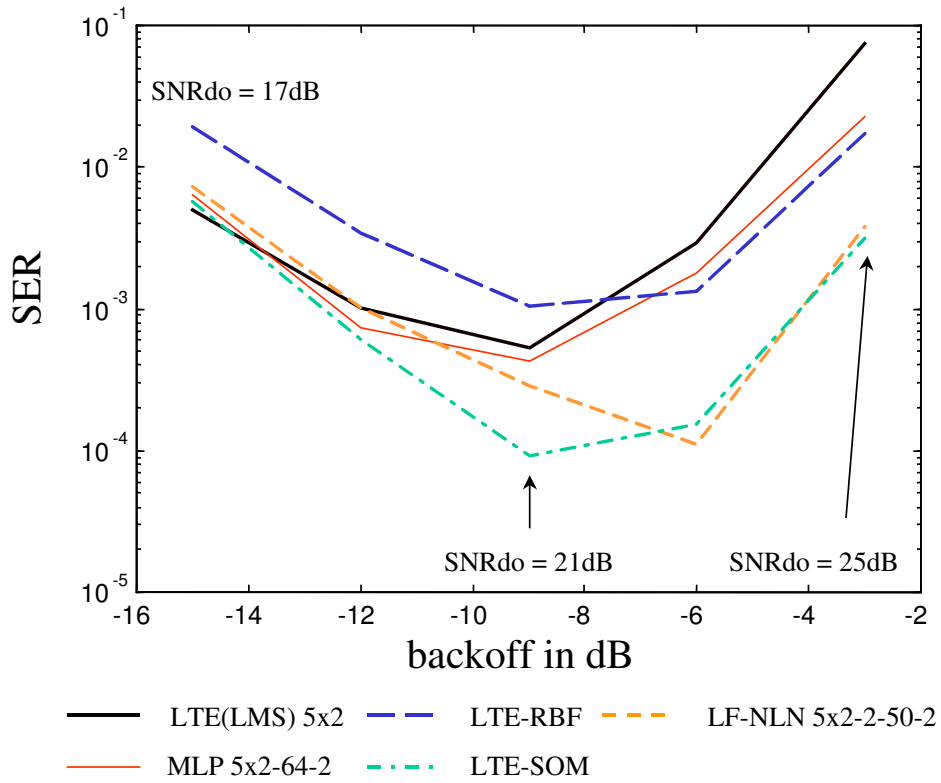
Neural Network Equalizers :

Linear Filter - NonLinear Network (LF-NLN)



- **Advantage :** More simple than the MLP.
- **Drawback :** Non optimal.

Application to the equalization of non-linear channels



Summary

Example of decision tree

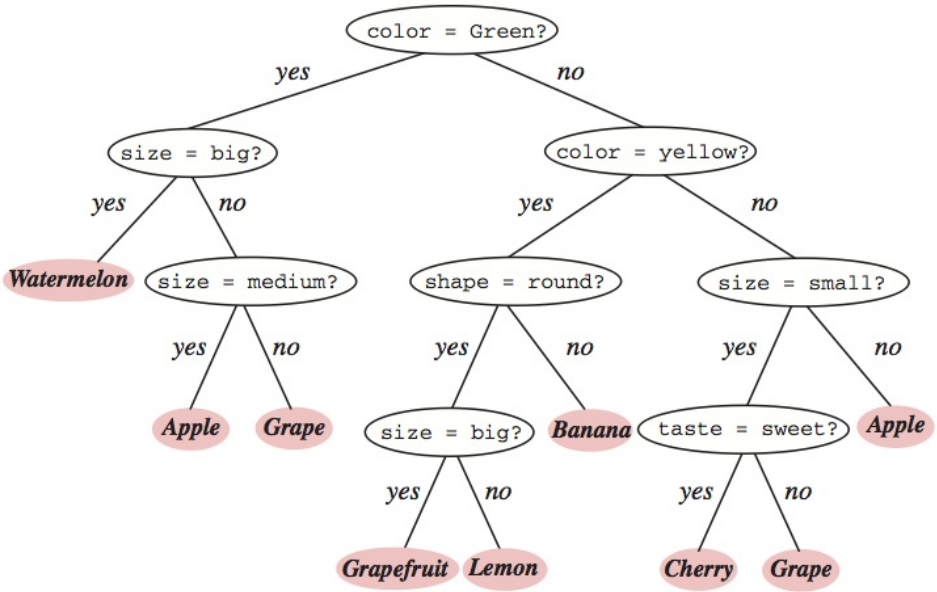


FIGURE 8.2. A tree with arbitrary branching factor at different nodes can always be represented by a functionally equivalent binary tree—that is, one having branching factor $B = 2$ throughout, as shown here. By convention the "yes" branch is on the left, the "no" branch on the right. This binary tree contains the same information and implements the same classification as that in Fig. 8.1. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Construction of the tree
- Classification rule

Splitting Rules

Inhomogeneity or impurity of the data

- Entropy (Algorithm C4.5)

$$i_n = -\sum_j \frac{n_j}{n} \log_2 \left(\frac{n_j}{n} \right)$$

- Gini Index (CART)

$$i_n = \sum_j \frac{n_j}{n} \left(1 - \frac{n_j}{n} \right)$$

Drop of Impurity

$$\Delta i_n = i_n - P_L i_L - P_R i_R$$

where $P_L = \frac{n_L}{n}, P_R = \frac{n_R}{n}$ are the proportions of the sets D_L, D_R.

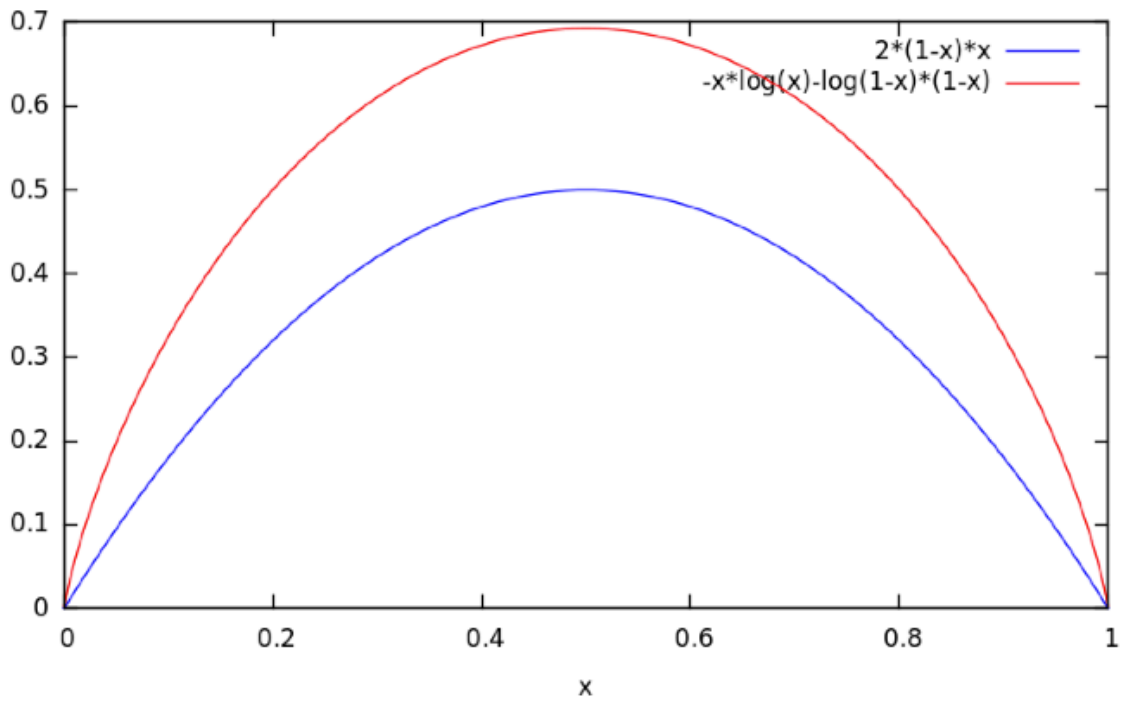
Choose the split associated with the maximum drop of impurity!

Gini Index or Entropy?

Example of 2 classes ($x = n_1/n$)

$$\text{Gini}(S) = 2x(1-x)$$

$$\text{Ent}(S) = -x\log x - (1-x)\log(1-x)$$



Implementation by default

Algorithm CART (classregtree.m)

- All variables are considered for each split
- All splits are considered
- Stopping rule: pure node or number of elements less than $n_{\min}$ (specified by the user)
- Splitting rule: Gini index

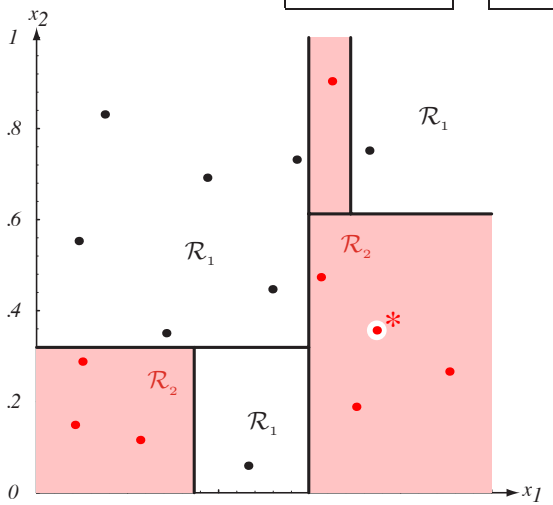
CART algorithm (example #1 for vectors in $\mathbb{R}^2$)

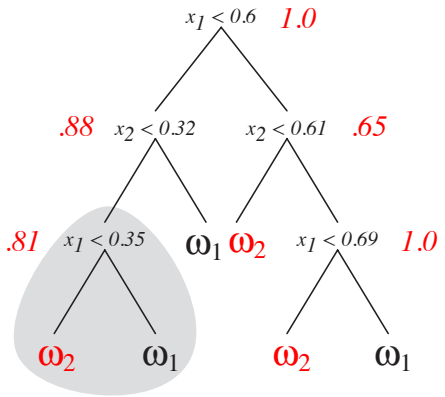
Example 1: A simple tree classifier

Consider the following $n = 16$ points in two dimensions for training a binary CART tree ($B = 2$) using the entropy impurity (Eq. 1).

ω_1 (black)	
x_1	x_2
.15	.83
.09	.55
.29	.35
.38	.70
.52	.48
.57	.73
.73	.75
.47	.06

ω_2 (red)	
x_1	x_2
.10	.29
.08	.15
.23	.16
.70	.19
.70	.19
.62	.47
.91	.27
.65	.90
.75	.36* (.32†)



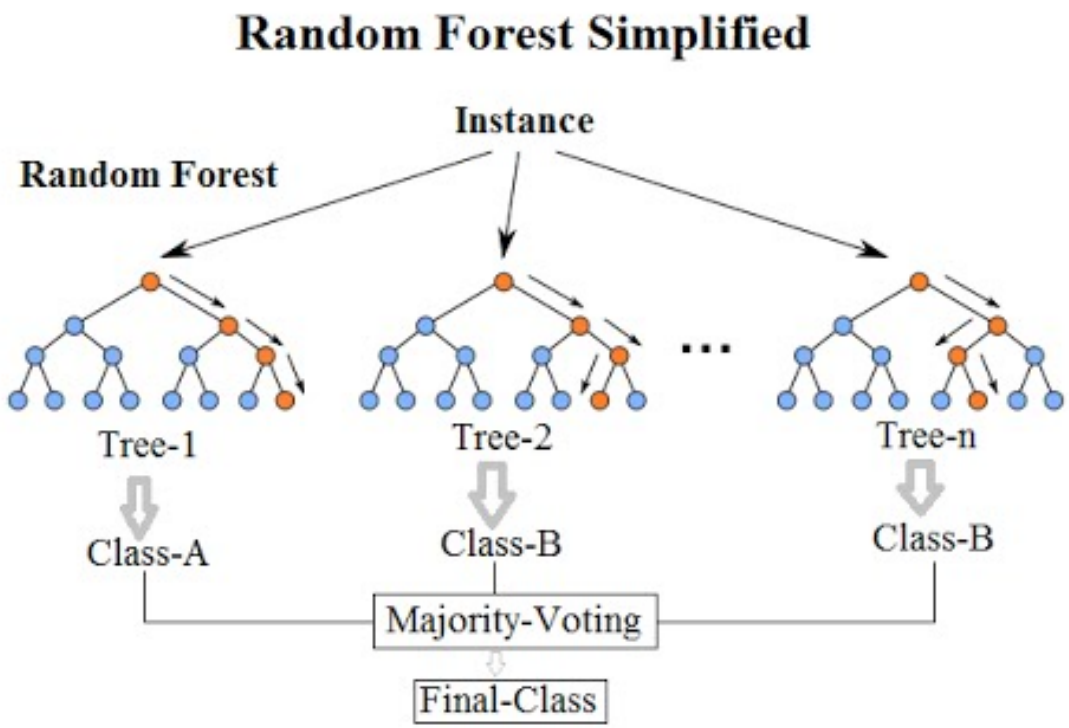


CART algorithm (example #2 for qualitative data)

	Weight	Size	Age	Result
$\boldsymbol{x}_1$	Light	Small	Young	Pass
$\boldsymbol{x}_2$	Light	Small	Young	Pass
$\boldsymbol{x}_3$	Light	Tall	Young	Pass
$\boldsymbol{x}_4$	Light	Tall	Old	Fail
$\boldsymbol{x}_5$	Light	Tall	Old	Pass
$\boldsymbol{x}_6$	Light	Tall	Old	Fail
$\boldsymbol{x}_7$	Heavy	Small	Old	Fail
$\boldsymbol{x}_8$	Heavy	Small	Young	Fail
$\boldsymbol{x}_9$	Light	Small	Old	Pass
$\boldsymbol{x}_{10}$	Heavy	Tall	Old	Pass

- First branch (Gini Index)
  - Weight $\Rightarrow 2(1/7 + 1/15) \sim 0.42$
  - Size $\Rightarrow 12/25 \sim 0.48$
  - Age $\Rightarrow 2(3/40 + 3/20) \sim 0.45$

Random Forests

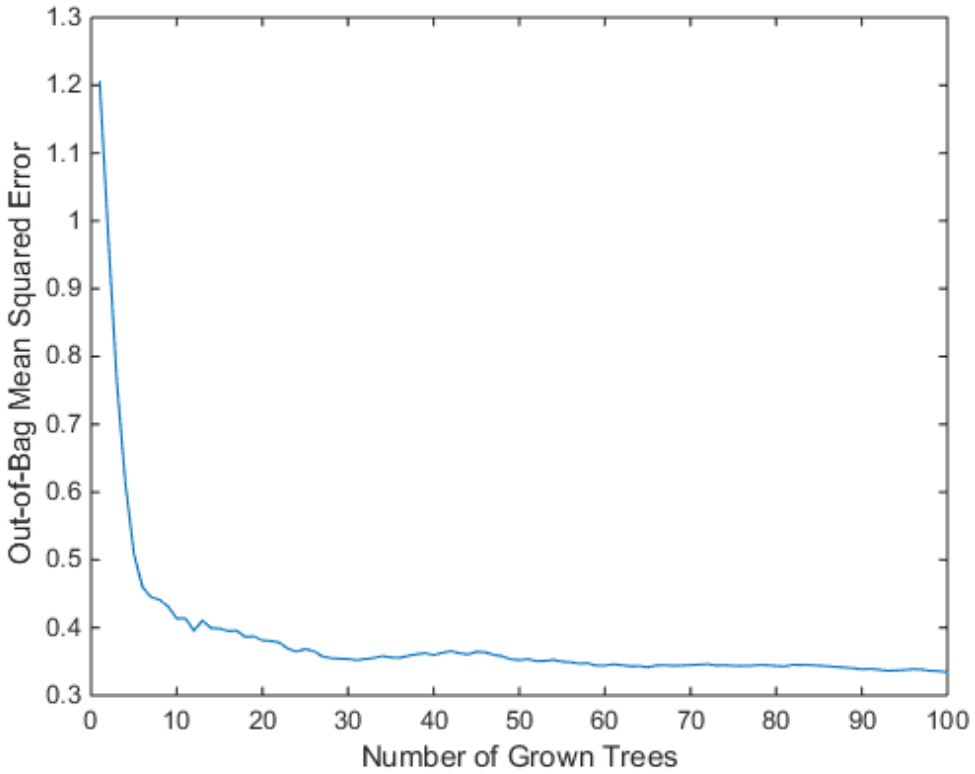


TreeBagger.m (options by default)

- Resampling all the data in the training set by bootstrap (and not a subset)
- Number of variables to select at random for each decision split: $\sqrt{n_{\text{var}}}$

Out-of-bag error

Matlab example



Comparison between CART and Random Forests

Book by G. James, D. Witten, T. Hastie and R. Tibshirani

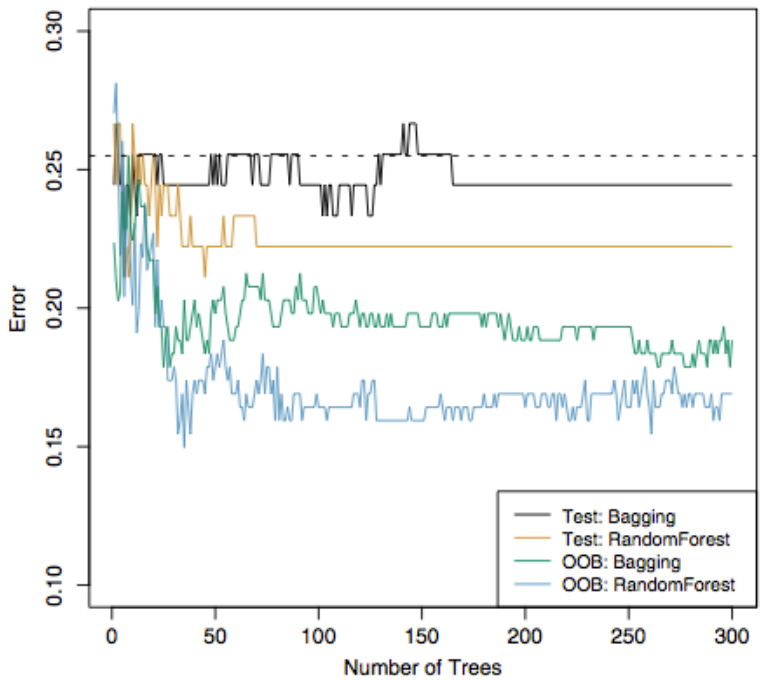


FIGURE 8.8. *Bagging and random forest results for the* `Heart` *data. The test error (black and orange) is shown as a function of* B*, the number of bootstrapped training sets used. Random forests were applied with* $m = \sqrt{p}$*. The dashed line indicates the test error resulting from a single classification tree. The green and blue traces show the OOB error, which in this case is considerably lower.*

Importance of the different variables

Mean decrease in Gini index

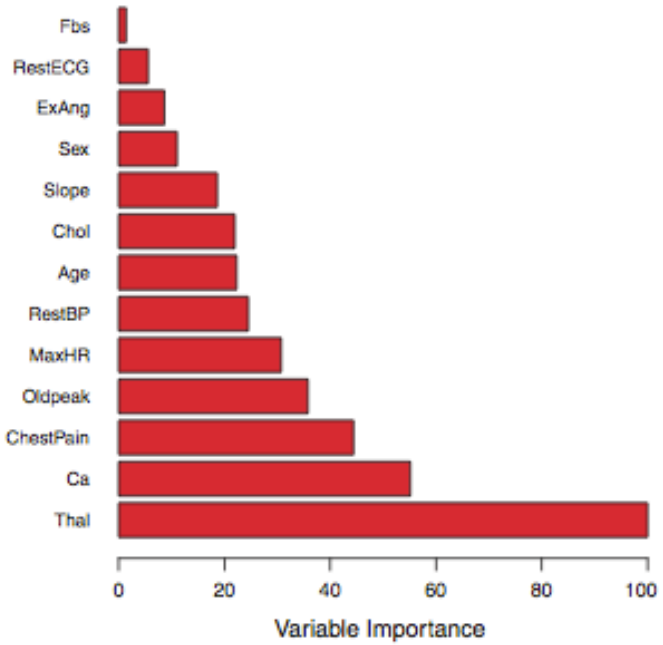


FIGURE 8.9. *A variable importance plot for the* `Heart` *data. Variable importance is computed using the mean decrease in Gini index, and expressed relative to the maximum.*

References

CART and Random Forests

- L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, Chapman & Hall, New-York, 1993.
- L. Breiman, Random Forests, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
- R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd edition, Wiley, 2000.
- G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, New-York, 2013.

Thanks for your attention