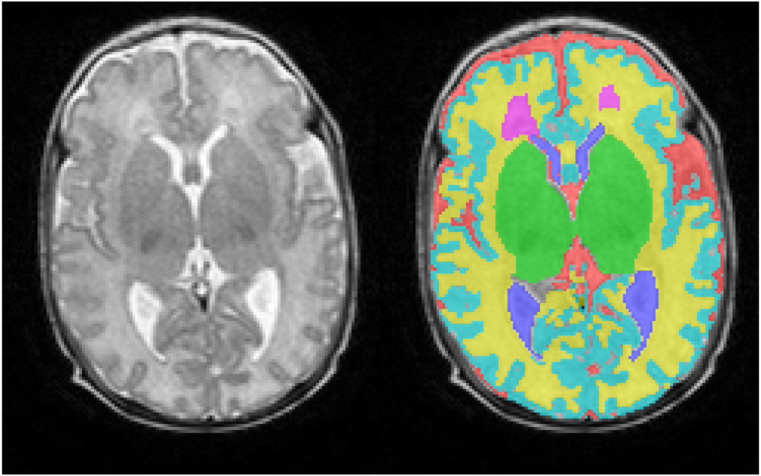
Classification and clustering: a few approaches

Isabelle Bloch

LTCI, Télécom Paris

isabelle.bloch@telecom-paris.fr

Objective: example



Content

A few approaches for classification and clustering with applications to images:

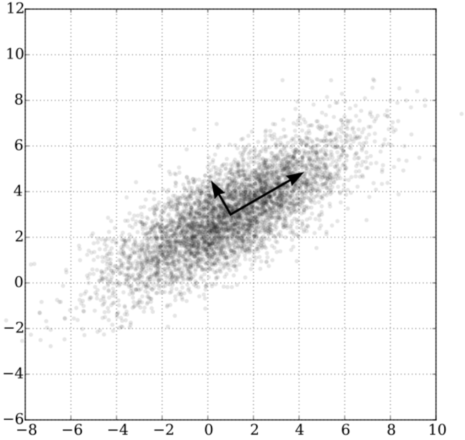
- Principal component analysis.
- Automatic clustering: k-means.
- Bayesian classification.
- Hierarchical clustering.
- SVM.
- Artificial neural networks.

Aim: data (images) $\rightarrow$ features $\rightarrow$ classes or clusters.

Feature space = characteristics of data on which classification will rely.
Decision space = classes, clusters.

Principal component analysis (PCA)

Find uncorrelated variables in the feature space



Reminder on covariance:
$A = (a_1 ... a_n), B = (b_1 ... b_n)$ with 0 mean ("centered").
Variance of A and B:

$$\sigma_A^2 = \frac{1}{n-1} \sum_{i=1}^{n} < a_i a_i > \quad \sigma_B^2 = \frac{1}{n-1} \sum_{i=1}^{n} < b_i b_i >$$

Covariance between A and B:

$$\sigma_{AB}^2 = \frac{1}{n-1} \sum_{i=1}^{n} < a_i b_i > = \frac{1}{n-1} AB^t$$

Properties of covariance:

- $\sigma_{AB}^2 = 0$ iff A and B are entirely decorrelated.
- $\sigma_{AB}^2 = \sigma_A^2$ iff $A = B$.

m samples of dimension $n \rightarrow$ matrix X of size $m \times n$.

- row of $X =$ all measurements of a particular type
- column of $X =$ measurements of a particular sample

Covariance $= S_X = \frac{1}{n-1} XX^t$ (square matrix $m \times m$).

- diagonal terms $=$ variance of particular measurement types
- off-diagonal terms $=$ covariance between measurement types

Diagonalization

$$Y = PX \text{ such that } S_Y = \frac{1}{n-1} YY^t \text{ is diagonal}$$

with P orthonormal $(p_i p_j = \delta_{ij})$,
Rows p_i of $P =$ principal components $=$ eigenvectors of XX^t.
ith diagonal value of $S_Y =$ variance of X along p_i.

PCA algorithm:

1. centering data (subtracting the mean for each measurement type);
2. computing the eigenvectors of XX^t.

Underlying assumptions:

- Linearity (extension to apply non-linearity before PCA: kernel PCA).
- Mean and variance sufficient $\Rightarrow$ Gaussian distributions.
- Principal components with larger associated variances represent interesting dimensions, while those with lower variances represent noise $\Rightarrow$ dimension reduction.
- Principal components are orthogonal.

Example: face recognition

e.g. http://www.cs.toronto.edu/~guerzhoy/320/lec/pca.pdf
http://dhoiem.cs.illinois.edu/courses/vision_spring10/lectures/Lecture15-FaceRecognition.ppt

Automatic clustering: k-means

- k classes $C_1 \ldots C_k$.
- Class center = prototype (in the feature space) = $m_1 \ldots m_k$.
- Distance d in the feature space: for sample x, $d(x, m_i)$ (Euclidean, Mahalanobis...)

Minimization of an objective function J:

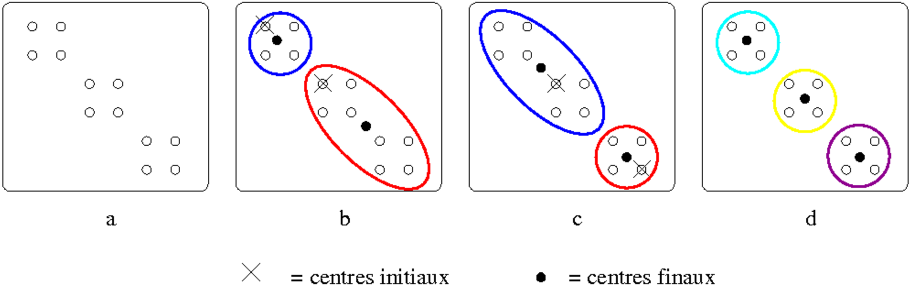
$$J = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, m_i)^2$$

Solution:

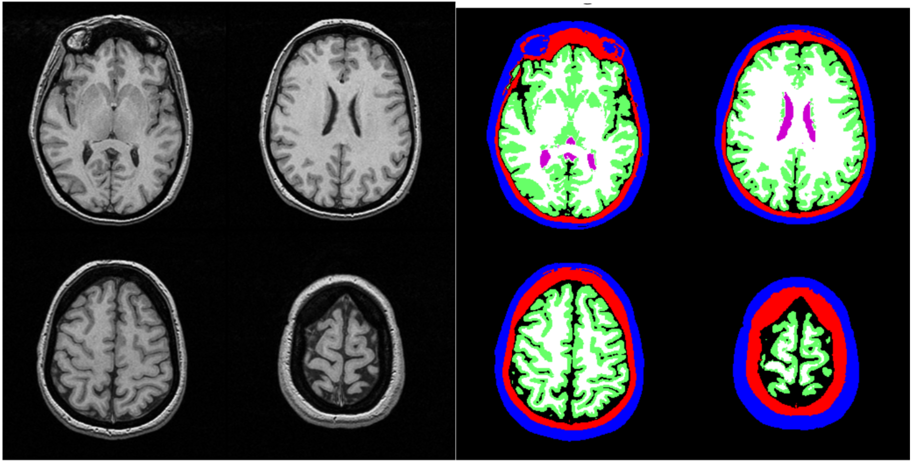
1. m_i fixed $\Rightarrow x \in C_i$ iff $\forall j, d(x, m_i) \leq d(x, m_j)$.
2. class assignment fixed $\Rightarrow m_i = \frac{\sum_{x \in C_i} x}{|C_i|}$.

Algorithm:

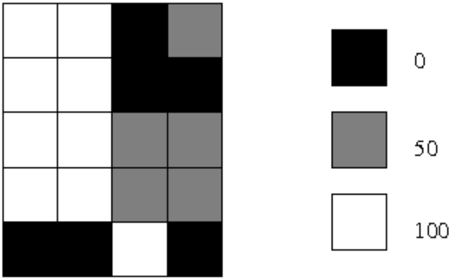
- Choose k and set initial class centers.
- Iterate steps 1 and 2.
- Until convergence... towards a local minimum of J!


a
b
c
d
= centres initiaux
= centres finaux

Example: MRI brain images



Exercise: which classification result for $k = 2$?



ISODATA

Extends k-means algorithm with splitting and grouping steps:

- adaptation of class number,
- more parameters,
- splitting along the direction of maximal variance,
- grouping if class centers are closed to each other.

Bayesian clustering

Probabilistic modeling of an image

- $\Omega = \{\omega_1, \omega_2, ...\}$ = set of classes
- s = site (pixel) and $S = \{s\}$ = set of sites
- Y_s = random variable associated with each site
- y_s gray level (realization of Y_s) – or more generally a feature vector

MAP decision:

s assigned to ω_i such that:

$$P(\omega_i | y_s) \text{ maximal}$$

Bayes rule:

$$P(\omega_i | y_s) = \frac{p(y_s | \omega_i) P(\omega_i)}{p(y_s)}$$

Estimating $p(y_s|\omega_i)$ (likelihood):

- learning from annotated data (supervised)
- histogram $\Rightarrow$ distribution of gray levels (unsupervised)
  - parametric methods
  - non parametric methods

Estimating $P(\omega_i)$ (prior):

- prior knowledge on class occurrences
- learning on a significant sample set
- otherwise equiprobable classes

$$P(\omega_i) = \frac{1}{Card(\Omega)}$$

$$\Rightarrow p(y_s|\omega_i) \text{ maximal}$$

= classification in the sense of maximum likelihood

Usual hypotheses:

- independence of features conditionnally to the classes
- Gaussian distributions
- Markov hypothesis on the class field $\Rightarrow$ included in the prior

Markovian regularization

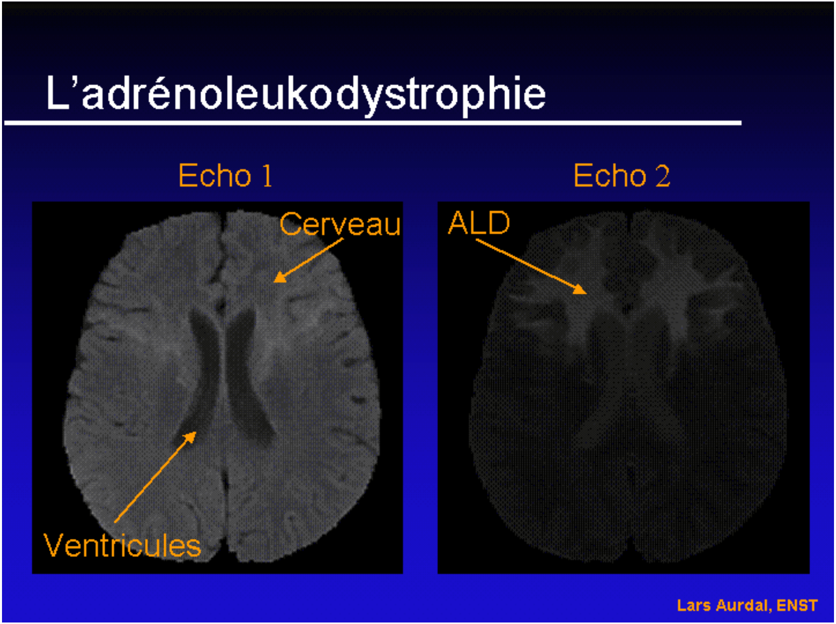
- Minimization of an energy defined globally on the image:

$$U(x) = U(x, y) + U_{prior}(x)$$

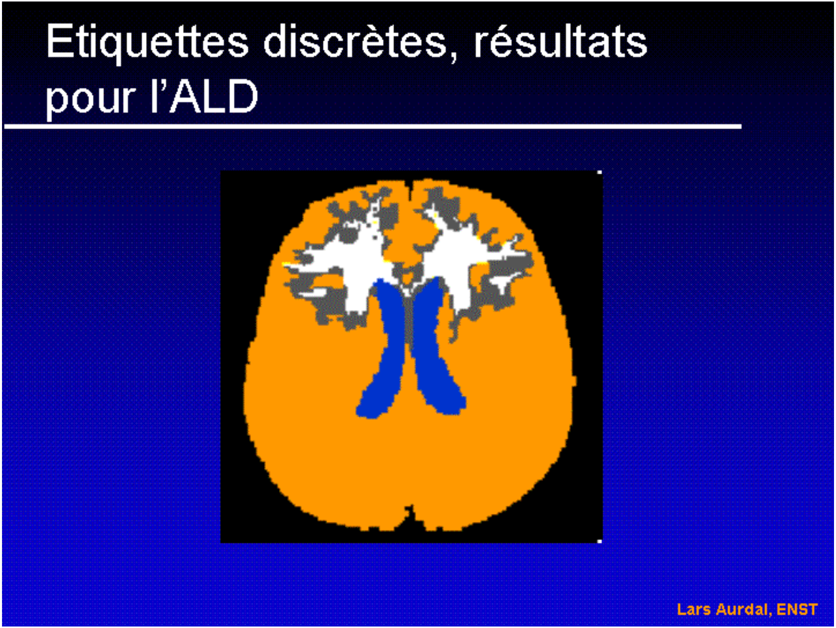
with

$$U(x) = \sum_{s} \frac{(\mu_{x_s} - y_s)^2}{\sigma^2} - \beta \sum_{s,t} \delta(x_s, x_t)$$

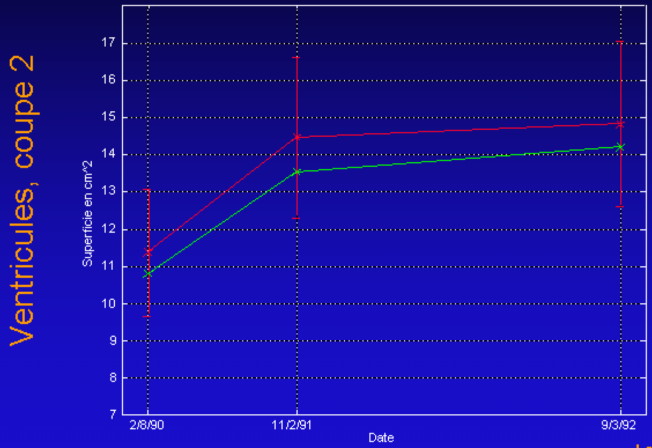
  - $U(x, y)$ = data fidelity
  - $U_{prior}(x)$ = contextual term
- No analytical solution
- Iterative and stochastic optimization (simulated annealing) or iterative conditional modes if the initialization is good enough



L'adrénoleukodystrophie
Echo 1
Echo 2
Cerveau
ALD
Ventricules
Lars Aurdal, ENST


Etiquettes discrètes, résultats pour l'ALD
Lars Aurdal, ENST

Etiquettes discrètes, suivi d'un cas d'ALD

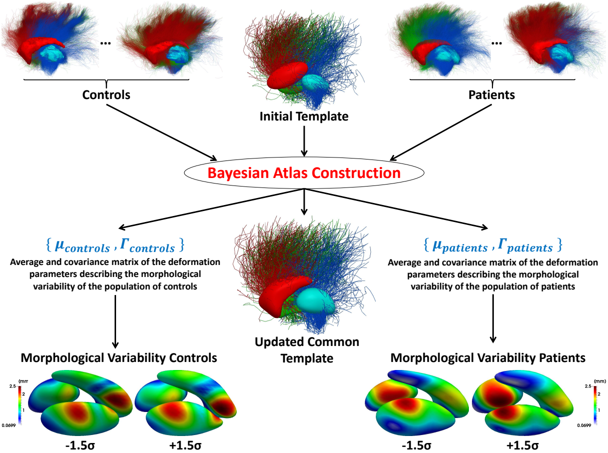


Lars Aurdal, ENST

Etiquettes discrètes, suivi d'un cas d'ALD

Lars Aurdal, ENST

Pietro Gori et al.



Exercise:

Landcover classification in a satellite image:

- wheat (class C_1) which covers 60 % of the region of interest,
- forest (class C_2) which covert 10% of the region of interest,
- remaining: man-made objects, etc.

A large zone Z has been detected, which is homogeneous, with a mean gray level $m = 80$.

After a learning phase, conditional probabilities $P(n|C_1)$ et $P(n|C_2)$ (for a gray level n) are estimated as Gaussian distributions of parameters:

- for wheat: $m_1 = 100$, $\sigma_1 = 20$,
- for forest: $m_2 = 85$, $\sigma_2 = 5$.

1. To which class should Z be assigned according to the maximum likelihood criterion?
2. To which class should Z be assigned according to the maximum a posteriori criterion?
3. Actually, the estimation of forest coverage is not known precisely and can vary between 8% à 20% of the region of interest. Would this change the decision?

Hierarchical clustering

Set of nested clusterings (ex: taxonomy).

http://www.statisticshowto.com

Equivalence between indexed hierarchy in the sample space E and ultra-metric on E.

Ultra-metric: $\delta : E \times E \rightarrow \mathbb{R}^+$ such that

$$\forall (x_1, x_2) \in E^2,\ \delta(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2,$$

$$\forall (x_1, x_2) \in E^2,\ \delta(x_1, x_2) = \delta(x_2, x_1),$$

$$\forall (x_1, x_2, x_3) \in E^3,\ \delta(x_1, x_2) \leq \max(\delta(x_1, x_3), \delta(x_3, x_2)).$$

$\Rightarrow$ any triangle is isocele.

Chain (path) distance:

- path from x_1 to x_2: sequence $c = \{c_1 = x_1, c_2, ...c_n = x_2\}$
- step $P(c) = \max_{i=1}^{n-1} d(c_i, c_{i+1})$
- chain distance: $\delta(x_1, x_2) = \inf_{c \in CH(x_1,x_2)} P(c)$
  ($CH(x_1, x_2)$ = set of all paths from x_1 to x_2)

$\Rightarrow \delta$ is an ultra-metric

Ordering property: $\forall \Delta$ ultra-metric, $\Delta \leq d \; \Rightarrow \; \Delta \leq \delta$,
for the ordering $d_1 \leq d_2 \; \Leftrightarrow \; \forall (x, y) \in E^2, \; d_1(x, y) \leq d_2(x, y)$.

Equivalence relation from an ultra-metric δ:

$$\forall \delta_0 \in \mathbb{R}^+,\ xR_{\delta_0}y \ \Leftrightarrow\ \delta(x,y) \leq \delta_0$$

Partition $P(\delta_0) = E/R_{\delta_0}$ = set of classes

- Within a class: path from any sample to any other sample by steps of length $\leq \delta_0$.
- From a sample of a class to a sample of another class: at least one step of length $> \delta_0$.

Hierarchy on E: $\mathcal{H} \subseteq \mathcal{P}(E)$ such that:

- $E \in \mathcal{H}$,
- $\forall x \in E, \ \{x\} \in \mathcal{H}$,
- $\forall (h, h') \in \mathcal{H}^2, \begin{cases} h \cap h' = \emptyset, \\ \text{or } h \subseteq h', \\ \text{or } h' \subseteq h. \end{cases}$.

Example: $\mathcal{H} = \bigcup_{\delta_0 \in \mathbb{R}^+} P(\delta_0)$

Indexed hierarchy: $(\mathcal{H}, f)$ such that $\mathcal{H}$ is a hierarchy and f is a mapping $\mathcal{H} \to \mathbb{R}^+$ such that:

- $\forall x \in E, \ f(\{x\}) = 0$,
- $\forall (h, h') \in \mathcal{H}^2, h \subset h', \ h \neq h' \ \Rightarrow \ f(h) < f(h')$.

Fundamental result

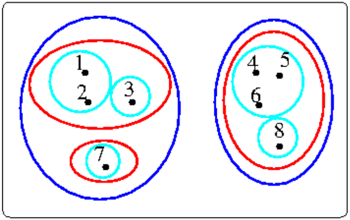
- Let $\mathcal{H} = \bigcup_{\delta_0 \in \mathbb{R}^+} P(\delta_0)$. $(\mathcal{H}, f)$ is an indexed hierarchy for:

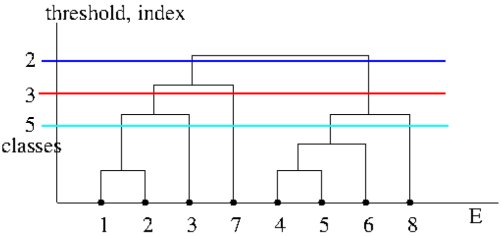
$$\forall h \in \mathcal{H}, \; f(h) = \min\{\delta_0 \mid h \in P(\delta_0)\}.$$

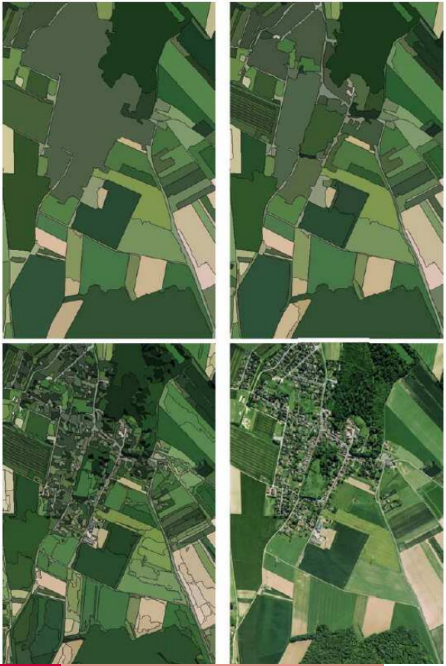
- Conversely, let $(\mathcal{H}, f)$ be an indexed hierarchy. Then δ defined as:

$$\delta(x, y) = \min_{h \in \mathcal{H}} \{f(h) \mid (x, y) \in h^2\}.$$

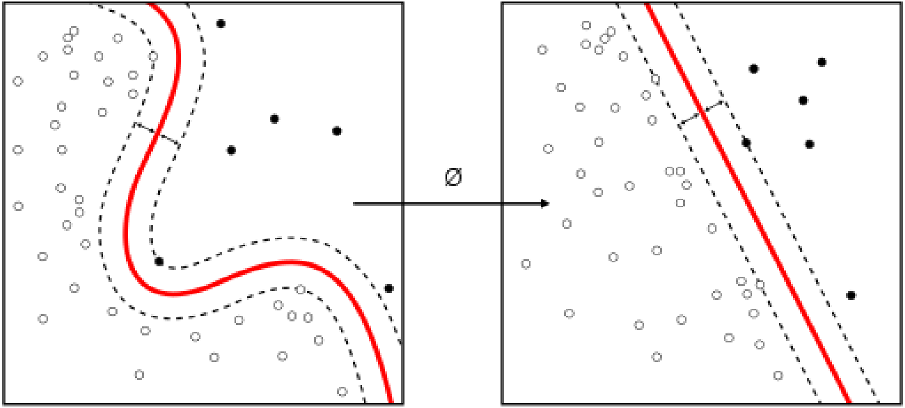
is an ultra-metric.



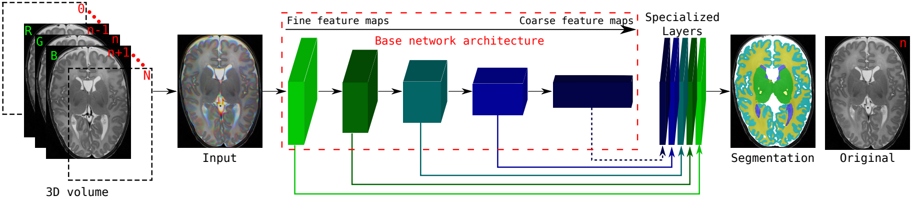


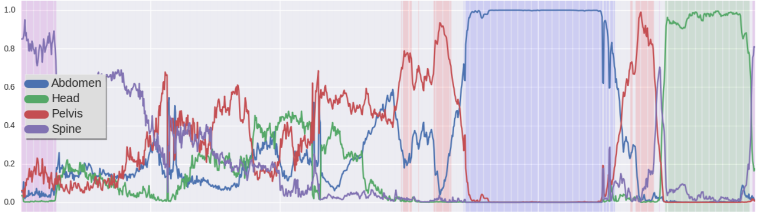


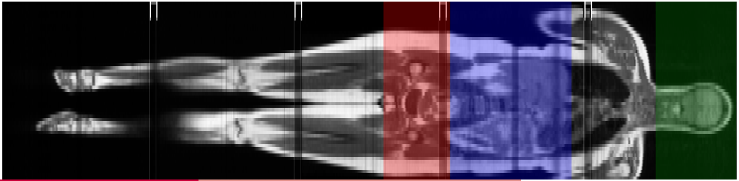
SVM



Artificial neural networks







To know more:
C. Bishop: Pattern Recognition and Machine Learning, Springer, 2006.
(among others!)