Probabilistic graphical models

Francis Bach, INRIA/ENS
Guillaume Obozinski, ENPC

M2 MVA 2013-2014

General information

- Every Wed 9am-12pm amphi Tocqueville until Nov 27.
- Except
  - This Friday Oct 11th 1.30pm-4.30pm amphi Tocqueville
  - Next Wed Oct 16th 9am-12pm amphi Curie
- **Grading** :
  - Homework 1 (20%)
  - Homework 2 (20%)
  - Take Home Exam (a longer Homework) (30%)
  - Project (30%)
- **Programming** :
  - All Hwk + Exam + Project involve programming
  - You may choose the programming language you want
  - We recommend you choose a vector oriented PL such as Python, R Matlab.

General information II

- **Calendar for the project** :

Mid-nov	choose a project to do alone or in pairs.
Before 11/27	send a mail announcing project choice.
Before 12/04	send a project draft (1 page)+ first results.
Before 12/20	Hand in your final exam (to Carine Saint-Prix/by email).
On 12/18	Poster session perhaps Pavillon des Jardins.
Before 01/10	Project reports due ($\approx$ 6 pages).

- **Polycopié** will be available later at the office of Carine Saint-Prix
  - Don't rush there now...
  - If you are not registered in the Master send an email to Carine to say that you would like to attend the course.
- **Email**
  - `francis.bach@ens.fr`
  - `guillaume.obozinski@imagine.enpc.fr`
  - always write to both of us + add "**MVA**" in the email title.
- **Lecture notes** Scribes

Machine learning

Goal

- Extract "statistical relations" between
  - a large number of input variables / features / descriptors
  - one or several decision/output variables
- Construct **empirical knowledge** :

  Turning empirical information into statistical knowledge

Specificities w.r.t. other AI approches

1. Knowledge essentially extracted from des données
2. **Generalization** ability

Specificities w.r.t. classical statistics

Goal

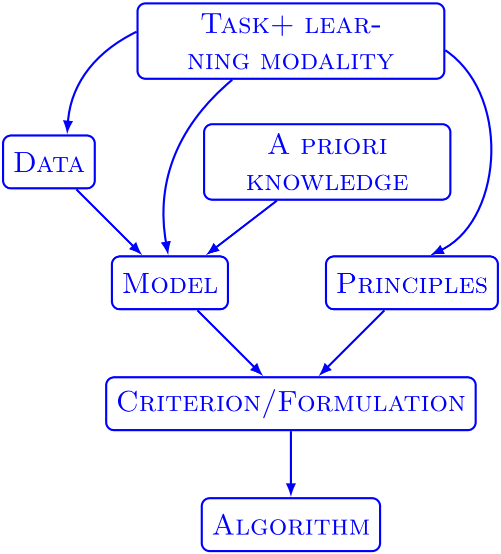
Predictive/Action model *vs* explanatory model of reality

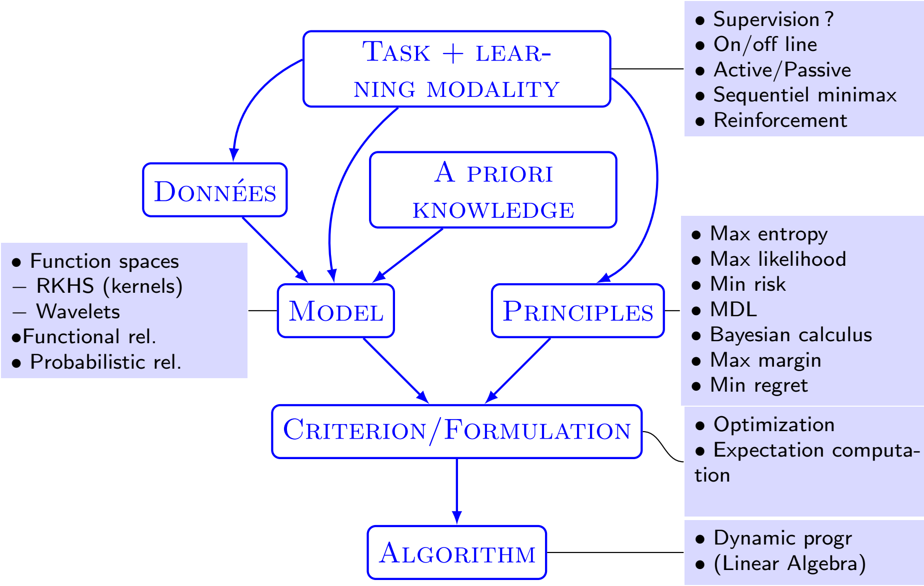
Challenge

Requires to integrate the info from a **very large number of variables**

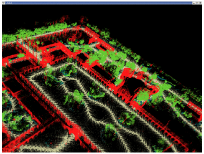
- Computer vision : 10^7 dimensions par image
- Brain imaging : 10^5 dimensions par volume
- Natural Language processing : $10^4 - 10^{15}$ paramètres
- Genetics : 10^4 gènes, 10^5 SNPs/ microsatellites, 10^9 bases d'ADN

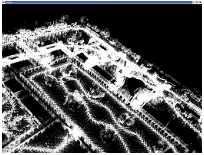
Which role for probabilistic modelling ?
How do proceed ?

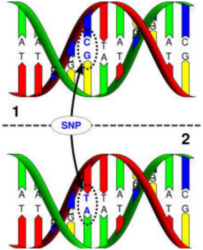

Task+ learning modality
Data
A priori knowledge
Model
Principles
Criterion/Formulation
Algorithm

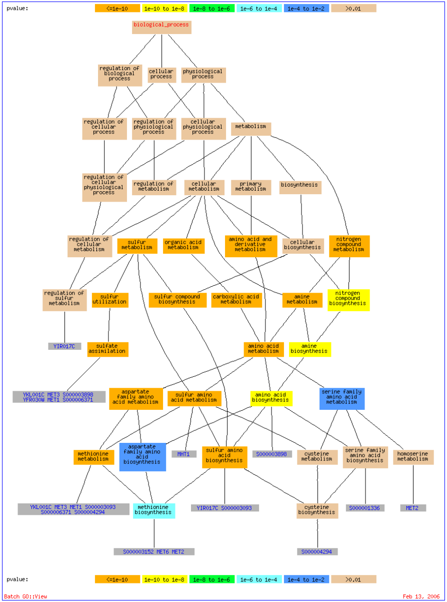

Task + learning modality
• Supervision ?
• On/off line
• Active/Passive
• Sequentiel minimax
• Reinforcement
Données
A priori knowledge
• Function spaces
– RKHS (kernels)
– Wavelets
•Functional rel.
• Probabilistic rel.
Model
Principles
• Max entropy
• Max likelihood
• Min risk
• MDL
• Bayesian calculus
• Max margin
• Min regret
Criterion/Formulation
• Optimization
• Expectation computation
Algorithm
• Dynamic progr
• (Linear Algebra)

Structured problems in HD





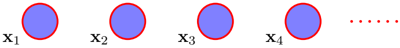




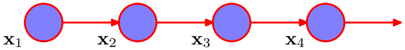
Sequence modelling

How to model the distribution of DNA sequences of length k ?

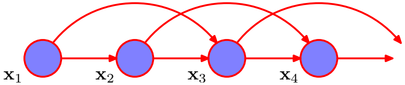
- Naive model $\rightarrow 4^n - 1$ parameters
- Indépendant model $\rightarrow 3n$ parameters



First order Markov chain :



Second order Markov chain :

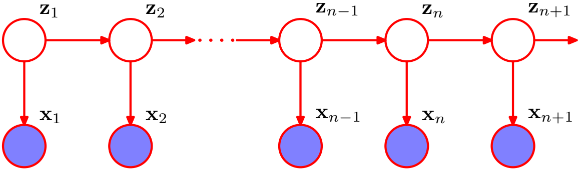


Number of parameters $\mathcal{O}(n)$ for chains of length n.

Models for speech processing

- Speech modelled by a sequence of unobserved phonemes
- For each phoneme a random sound is produced following a distribution which characterizes the phoneme

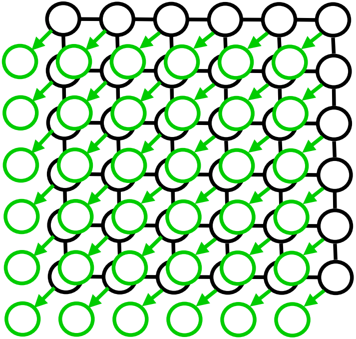
Hidden Markov Model : HMM (Modèle de Markov caché)



$\rightarrow$ **Latent** variable models

Modelling image structures

Markov Random Field
(Champ de Markov caché)



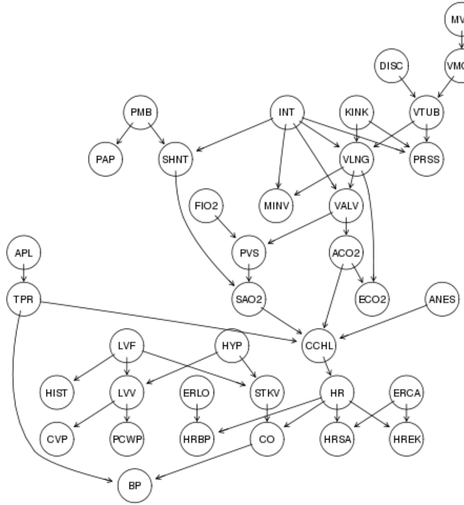
Original image

Segmentation

→ *oriented graphical model* vs *non oriented*

Anaesthesia alarm (Beinlich et al., 1989)

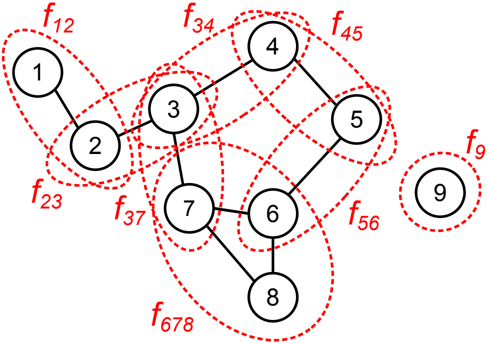
"The ALARM Monitoring system"



http://www.bnlearn.com/documentation/networks/

CVP	central venous pressure
PCWP	pulmonary capillary wedge pressure
HIST	history
TPR	total peripheral resistance
BP	blood pressure
CO	cardiac output
HRBP	heart rate / blood pressure.
HREK	heart rate measured by an EKG monitor
HRSA	heart rate / oxygen saturation.
PAP	pulmonary artery pressure.
SAO2	arterial oxygen saturation.
FIO2	fraction of inspired oxygen.
PRSS	breathing pressure.
ECO2	expelled CO2.
MINV	minimum volume.
MVS	minimum volume set
HYP	hypovolemia
LVF	left ventricular failure
APL	anaphylaxis
ANES	insufficient anesthesia/analgesia.
PMB	pulmonary embolus
INT	intubation
KINK	kinked tube.
DISC	disconnection
LVV	left ventricular end-diastolic volume
STKV	stroke volume
CCHL	catecholamine
ERLO	error low output
HR	heart rate.
ERCA	electrocauter
SHNT	shunt
PVS	pulmonary venous oxygen saturation
ACO2	arterial CO2
VALV	pulmonary alveoli ventilation
VLNG	lung ventilation
VTUB	ventilation tube
VMCH	ventilation machine

Probabilistic model



$$
\begin{aligned}
p(x_1, x_2, \ldots, x_9) \quad = \quad & f_{12}(x_1, x_2)\, f_{23}(x_2, x_3)\, f_{34}(x_3, x_4)\, f_{45}(x_4, x_5) \ldots \\
& f_{56}(x_5, x_6)\, f_{37}(x_3, x_7)\, f_{678}(x_6, x_7, x_8)\, f_9(x_9)
\end{aligned}
$$

Abstact models vs concrete ones

Abstracts models

- Linear regression
- Logistic regression
- Mixture model
- Principal Component Analysis
- Canonical Correlation Analysis
- Independent Component analysis
- LDA (Multinomiale PCA)
- Naive Bayes Classifier
- Mixture of experts

Concrete Models

- Markov chains
- HMM
- Tree-structured models
- Double HMMs
- Oriented acyclic models
- Markov Random Fields
- Star models
- Constellation Model

Operations on graphical models

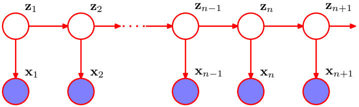
Probabilistic inference

Computing a marginal distr. $p(x_i)$ ou $p(x_i|x_1 = 3, x_7 = 0)$

Decoding (MAP inference)

What is the most likely instance ?

$$\text{argmax}_z p(z|x)$$



Learning (or Estimation)

Soit $p(x;\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_C \psi(x_C, \theta_C)$, we want to find

$$\text{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(x^{(i)};\boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{\theta}} \frac{1}{Z(\boldsymbol{\theta})} \prod_{i=1}^{n} \prod_{C} \psi(x_C^{(i)}, \theta_C)$$

Course outline

- **Course 1**
  Introduction
  Maximum likelihood
  Models with a single node
- **Course 2**
  Linear regression
  Logistic regression
  Generative classification (Fisher discriminant)
- **Cours 3**
  K-means
  EM
  Gaussian mixtures
  Graph Theoretic aspects
- **Cours 4**
  Unoriented graphical models
  Oriented graphical models
- **Cours 5**
  Exponential families
  Information Theory
- **Cours 6**
  Gaussian Variables
  Factorial Analysis
- **Cours 7**
  Sum-product algorithm
- **Cours 8**
  Approximate inférence
- **Cours 9**
  Bayesian methods

General information

- Every Wed 9am-12pm amphi Tocqueville until Nov 27.
- Except
  - This Friday Oct 11th 1.30pm-4.30pm amphi Tocqueville
  - Next Wed Oct 16th 9am-12pm amphi Curie
- **Grading** :
  - Homework 1 (20%)
  - Homework 2 (20%)
  - Take Home Exam (a longer Homework) (30%)
  - Project (30%)
- **Programming** :
  - All Hwk + Exam + Project involve programming
  - You may choose the programming language you want
  - We recommend you choose a vector oriented PL such as Python, R Matlab.

General information II

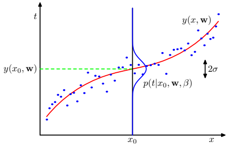
- **Calendar for the project** :

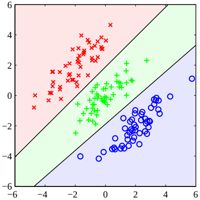
Mid-nov	choose a project to do alone or in pairs.
Before 11/27	send a mail announcing project choice.
Before 12/04	send a project draft (1 page)+ first results.
Before 12/20	Hand in your final exam (to Carine Saint-Prix/by email).
On 12/18	Poster session perhaps Pavillon des Jardins.
Before 01/10	Project reports due ($\approx$ 6 pages).

- **Polycopié** will be available later at the office of Carine Saint-Prix
  - Don't rush there now...
  - If you are not registered in the Master send an email to Carine to say that you would like to attend the course.
- **Email**
  - `francis.bach@ens.fr`
  - `guillaume.obozinski@imagine.enpc.fr`
  - always write to both of us + add "**MVA**" in the email title.
- **Lecture notes** Scribes

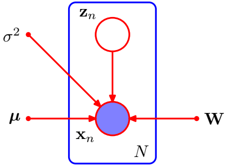
To start : models with 1 and 2 nodes...

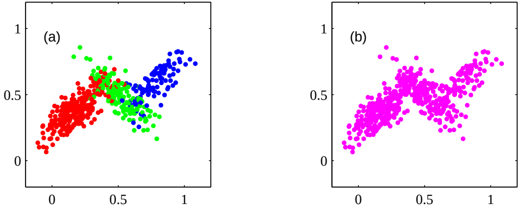
Regression and classification





Mixture models





Transversal concepts

- *Generative* models vs *discriminative*
- *Supervised* vs *unsupervised* learning
- Learning from *completely observed* data vs *incomplete* data
- *Causation* vs *correlations* :
  Graphical models are **not** modelling **causation** $\rightarrow$ modelling **correlation** based on sets of**conditional independences**.

Notations, formulas,definitions

- Joint distribution of X_A et X_B : $p(x_A, x_B)$
- Marginale distribution : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Conditional distribution : $p\left(x_A|x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$

Bayes formula

$$p\left(x_A|x_B\right) = \frac{p\left(x_B|x_A\right) p\left(x_A\right)}{p\left(x_B\right)}$$

$\rightarrow$ Bayes formula is not “bayesian”.

Expectation and Variance

- Expectation of X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Expectation of $f(X)$, for f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

- Conditional Expectation de X given Y :

$$\mathbb{E}[X|Y] = \sum_x x \cdot p(x|y)$$

- Conditional Variance :

$$\text{Var}(X \mid Y) = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2 |Y\right] = \mathbb{E}\left[X^2|Y\right] - \mathbb{E}[X|Y]^2$$

Independence concepts

Independence : $X \perp\!\!\!\perp Y$

We say that X et Y are independents and write $X \perp\!\!\!\perp Y$ ssi :

$$\forall x, y, \qquad P(X = x, Y = y) = P(X = x)\, P(Y = y)$$

Conditional Independence : $X \perp\!\!\!\perp Y \mid Z$

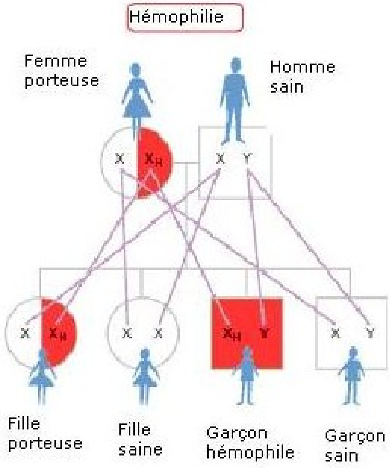
- On says that X and Y are independent conditionally on Z and
- write $X \perp\!\!\!\perp Y \mid Z$ iff :

$\forall x, y, z,$

$$P(X = x, Y = y \mid Z = z) = P(X = x|Z = z)\, P(Y = y|Z = z)$$

Conditional Independence : example

"X-linked recessive disease" :
Transmission of the gene of hemophilia



Risk of illness or sons of a healthy father :

- dependent for two brothers.
- conditionally independant given whether the mother is a carrier or not.

Statistical model

Parametric model – Definition :

Ensemble of probability distributions parameterized by a vector $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \{p(x;\theta) \mid \theta \in \Theta\}$$

Bernoulli model : $X \sim \text{Ber}(\theta) \qquad \Theta = [0,1]$

$$p(x;\theta) = \theta^x (1-\theta)^{(1-x)}$$

Binomial model : $Y \sim \text{Bin}(n,\theta) \qquad \Theta = [0,1]$

$$p(Y;\theta) = \binom{n}{x} \theta^y (1-\theta)^{(n-y)}$$

Multinomial model : $Z \sim \mathcal{M}(n, \pi_1, \pi_2, \ldots, \pi_K) \qquad \Theta = [0,1]^K$

$$p(z;\theta) = \binom{n}{z_1, \ldots, z_k} {\pi_1}^{z_1} \, \ldots \, {\pi_k}^{z_k}$$

Gaussian model

Univariate gaussian : $X \sim \mathcal{N}(\mu, \sigma^2)$

X is real valued r.v., et $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

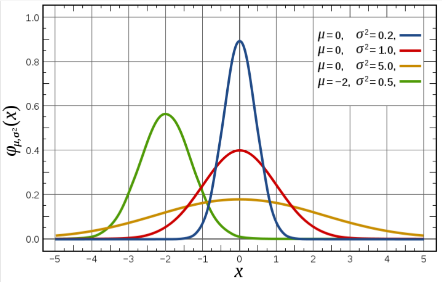
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

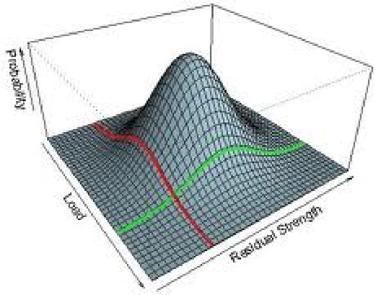
Multivariate gaussian : $X \sim \mathcal{N}(\mu, \Sigma)$

X takes values in $\mathbb{R}^d$. Si $\mathcal{K}_n$ is the set of $n \times n$ positive definite matrices, and $\theta = (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \mathcal{K}_n$.

$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Gaussian densities





Maximum likelihood principle

- Let a model $\mathcal{P}_\Theta = \left\{ p(x;\theta) \mid \theta \in \Theta \right\}$
- Let an observation x

Sir Ronald Fisher
(1890-1962)

Likelihood :

$$\begin{array}{rcl} \mathcal{L} : \Theta & \to & \mathbb{R}_+ \\ \theta & \mapsto & p(x;\theta) \end{array}$$

Maximum likelihood estimator :

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(x;\theta)$$

Case of i.i.d. data

For $(x_i)_{1 \le i \le n}$ a *sample* of i.i.d. data of size n :

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i;\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log\, p(x_i;\theta)$$

Examples of calculations of the MLE

- Bernoulli model
- Multinomial model
- Gaussien model

Bayesian estimation

Parameters θ are modelled as a random variable.

A priori

We have an *a priori* $p(\theta)$ on the model parameters.

A posteriori

The data contribute to the likelihood : $p(x|\theta)$.
The *a posteriori* probability of parameters is then

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} \propto p(x|\theta)\,p(\theta)\,.$$

$\rightarrow$ The Bayesian estimator is thus a probability distibution on the parameters.

One talks about Bayesian inference.

References

- Book of Christopher Bishop :

  Pattern Recognition and Machine Learning, 2006, Springer.
  http://research.microsoft.com/~cmbishop/PRML/

- David Barber's book is available online :

  Bayesian Reasoning and Machine Learning.
  Cambridge University Press, 2012.
  http://www.cs.ucl.ac.uk/staff/d.barber/brml/

- A good book on optimization theory :

  Nonlinear Programming, 1999, Dimitri Bertsekas. Athena Scientific.