Modèles graphiques probabilistes

Francis Bach & Guillaume Obozinski

INRIA - Ecole Normale Supérieure

M2 MVA 2011-2012

Informations générales

- Tous les mercredi 9h-12h salle C103 jusqu'au 23 Novembre.
- **Validation du cours**: DM1 (20 %) + DM2 (20 %) + examen final en temps libre (30%)+ projet (30%).
- **Calendrier**:

Mi-novembre	choisir un projet à faire seul ou en binôme.
Avant 1/12	envoyer un email annonçant le choix du projet.
Avant 17/12	déposer votre examen (au secrétariat / par email).
Avant 17/12	envoyer un draft de projet (1 page)+1er résultats.
Le 5/1	Session poster aux Pavillon des Jardins.
Le 5/1	Rendu rapport de projets ($\approx$ 6 pages).

- **Polycopié** à aller chercher au secrétariat
- **Email** `francis.bach@ens.fr + guillaume.obozinski@ens.fr` écrire aux 2 + mettre "**MVA**" dans le titre de l'email.
- **Notes de cours:** certaines existantes, bonus pour les scribes volontaires (à faire en binôme) pour les notes inexistantes.

Apprentissage automatique

Objet

- Extraire des “relations statistiques” entre
  - un grand nombre de variables d’entrées / prédicteurs / descripteurs
  - une ou plusieurs variables de sorties / décision(s)
- Construire une **connaissance empirique**:

  Transformation d’information empirique en connaissance statistique

Apprentissage automatique

Objet

- Extraire des “relations statistiques” entre
  - un grand nombre de variables d’entrées / prédicteurs / descripteurs
  - une ou plusieurs variables de sorties / décision(s)
- Construire une **connaissance empirique**:

  Transformation d’information empirique en connaissance statistique

Spécificités par rapport aux autres approches en IA

1. Connaissance construite essentiellement à partir des données
2. Capacité de **généralisation**

Spécificité par rapport aux statistiques classiques

But

Modèle prédictif/d'action *vs* modèle explicatif de la réalité.

Spécificité par rapport aux statistiques classiques

But

Modèle prédictif/d'action *vs* modèle explicatif de la réalité.

Difficulté

Nécessité d'intégrer un **un très grand nombre de variables**

Spécificité par rapport aux statistiques classiques

But

Modèle prédictif/d'action *vs* modèle explicatif de la réalité.

Difficulté

Nécessité d'intégrer un **un très grand nombre de variables**

- Vision artificielle: 10^7 dimensions par image
- Imagerie Cérébrale: 10^5 dimensions par volume
- Traitement automatique des langues: $10^4 - 10^{15}$ paramètres
- Génétique: 10^4 gènes, 10^5 SNPs/ microsatellites, 10^9 bases d'ADN

Spécificité par rapport aux statistiques classiques

But

Modèle prédictif/d'action *vs* modèle explicatif de la réalité.

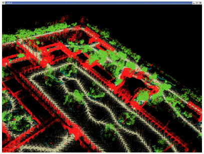
Difficulté

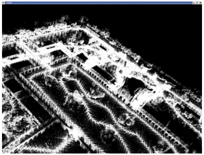
Nécessité d'intégrer un **un très grand nombre de variables**

- Vision artificielle: 10^7 dimensions par image
- Imagerie Cérébrale: 10^5 dimensions par volume
- Traitement automatique des langues: $10^4 - 10^{15}$ paramètres
- Génétique: 10^4 gènes, 10^5 SNPs/ microsatellites, 10^9 bases d'ADN

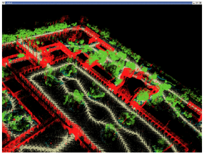
Quel rôle pour la modélisation probabiliste?
Comment la faire?

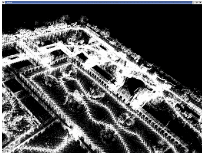
Des problèmes structurés de grande dimension

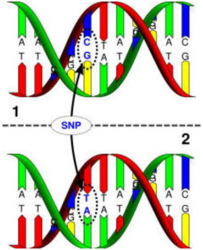




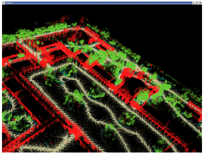
Des problèmes structurés de grande dimension

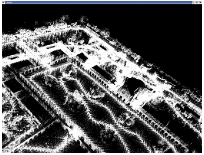


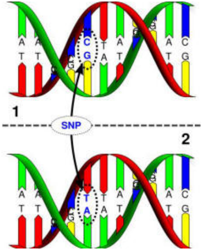


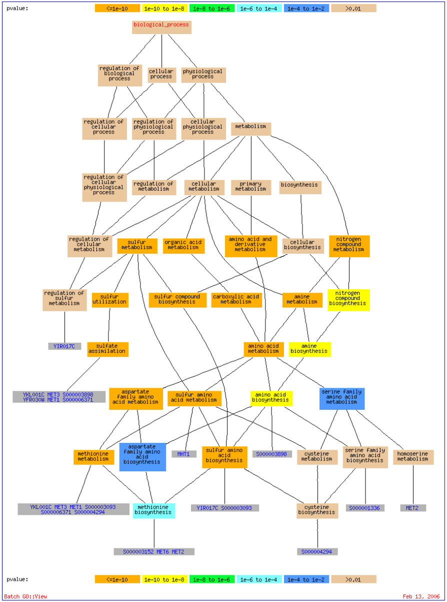


Des problèmes structurés de grande dimension









Modélisation de séquences

Comment modéliser la distribution de séquences d'ADN de longueur k?

Modélisation de séquences

Comment modéliser la distribution de séquences d'ADN de longueur k?

- Modèle naïf $\rightarrow 4^k - 1$ paramètres

Modélisation de séquences

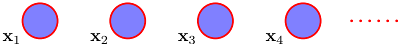
Comment modéliser la distribution de séquences d'ADN de longueur k?

- Modèle naïf $\rightarrow 4^k - 1$ paramètres
- Modèle indépendant $\rightarrow 3k$ paramètres

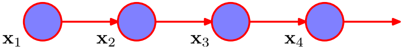
Modélisation de séquences

Comment modéliser la distribution de séquences d'ADN de longueur k?

- Modèle naïf $\rightarrow 4^k - 1$ paramètres
- Modèle indépendant $\rightarrow 3k$ paramètres



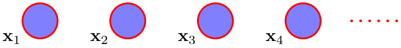
Chaîne de Markov d'ordre 1:



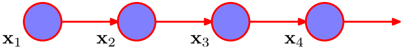
Modélisation de séquences

Comment modéliser la distribution de séquences d'ADN de longueur k?

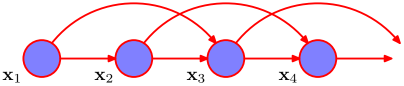
- Modèle naïf $\rightarrow 4^k - 1$ paramètres
- Modèle indépendant $\rightarrow 3k$ paramètres



Chaîne de Markov d'ordre 1:



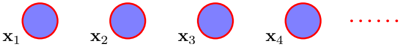
Chaîne de Markov d'ordre 2:



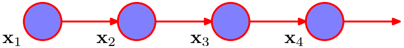
Modélisation de séquences

Comment modéliser la distribution de séquences d'ADN de longueur k?

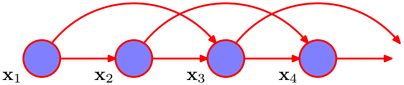
- Modèle naïf $\to 4^k - 1$ paramètres
- Modèle indépendant $\to 3k$ paramètres



Chaîne de Markov d'ordre 1:



Chaîne de Markov d'ordre 2:

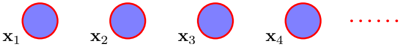


Nombres de paramètres $\mathcal{O}(k)$ pour des chaînes de longueur k.

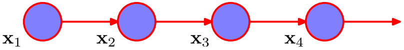
Modélisation de séquences

Comment modéliser la distribution de séquences d'ADN de longueur k?

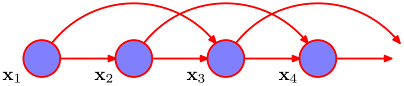
- Modèle naïf $\rightarrow 4^k - 1$ paramètres
- Modèle indépendant $\rightarrow 3k$ paramètres



Chaîne de Markov d'ordre 1:



Chaîne de Markov d'ordre 2:



Nombres de paramètres $\mathcal{O}(k)$ pour des chaînes de longueur k.

Modélisation pour le traitement de la parole

- Parole modélisée par une séquence de phonèmes non-observés

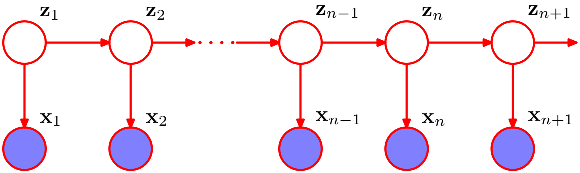
Modélisation pour le traitement de la parole

- Parole modélisée par une séquence de phonèmes non-observés
- Pour chaque phonème un son aléatoire est produit selon une distribution qui caractérise le phonème

Modélisation pour le traitement de la parole

- Parole modélisée par une séquence de phonèmes non-observés
- Pour chaque phonème un son aléatoire est produit selon une distribution qui caractérise le phonème

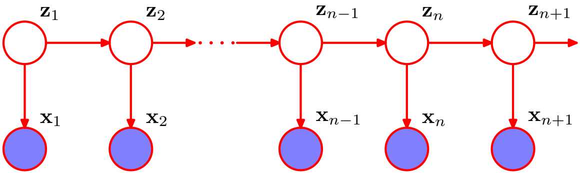
Modèle de Markov caché: HMM (Hidden Markov Model)



Modélisation pour le traitement de la parole

- Parole modélisée par une séquence de phonèmes non-observés
- Pour chaque phonème un son aléatoire est produit selon une distribution qui caractérise le phonème

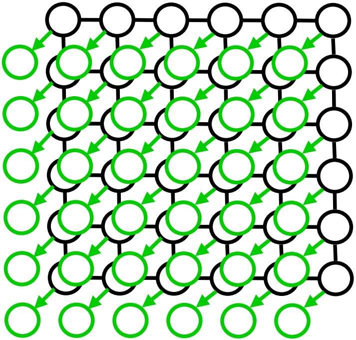
Modèle de Markov caché: HMM (Hidden Markov Model)



$\rightarrow$ modèle dit à variables **latentes**

Modélisation des structure dans les images

Champ de Markov caché

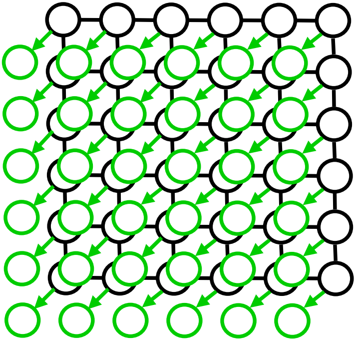


Original image

Segmentation

Modélisation des structure dans les images

Champ de Markov caché



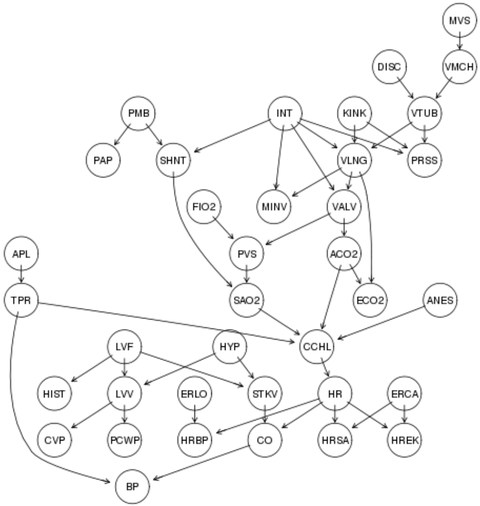
Original image

Segmentation

→ *modèle graphique orienté* vs *non orienté*

Alarme d'anesthésie (Beinlich et al., 1989)

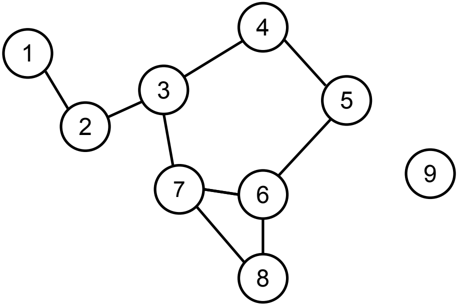
"The ALARM Monitoring system"



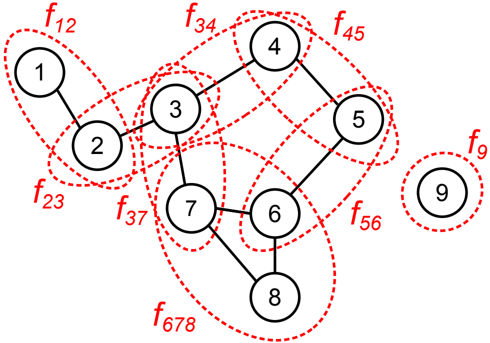
http://www.bnlearn.com/documentation/networks/

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

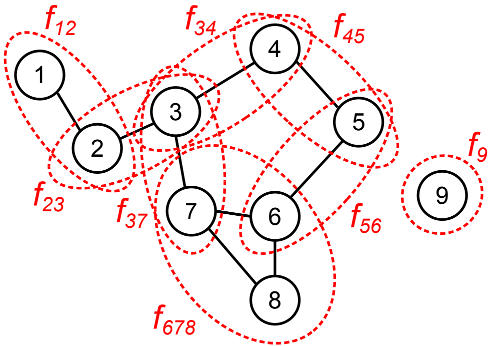
Un modèle probabiliste



Un modèle probabiliste



Un modèle probabiliste



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

Modèles abstraits vs concrets

Modèle abstraits

- Régression linéaire
- Régression logistique
- Modèle de mélange
- Analyse en composantes principales
- Analyse canonique des corrélations
- Analyse en composantes indépendantes
- LDA (ACP multinomiale)
- Classifieur Naive Bayes
- Mélange d'experts

Modèles concrets

- Chaînes de Markov
- HMM
- Modèles arborescents
- HMM double
- Modèles orientés acycliques
- Champs de Markov
- Modèles en étoile
- Modèles de constellation

Modèles abstraits vs concrets

Modèle abstraits

- Régression linéaire
- Régression logistique
- Modèle de mélange
- Analyse en composantes principales
- Analyse canonique des corrélations
- Analyse en composantes indépendantes
- LDA (ACP multinomiale)
- Classifieur Naive Bayes
- Mélange d'experts

Modèles concrets

- Chaînes de Markov
- HMM
- Modèles arborescents
- HMM double
- Modèles orientés acycliques
- Champs de Markov
- Modèles en étoile
- Modèles de constellation

Opérations sur les modèles graphiques

L'Inférence

Calculer une marginale $p(x_i)$ ou $p(x_i | x_1 = 3, x_7 = 0)$

Opérations sur les modèles graphiques

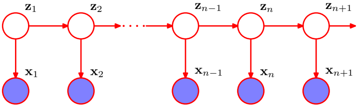
L'Inférence

Calculer une marginale $p(x_i)$ ou $p(x_i|x_1 = 3, x_7 = 0)$

Le Décodage

Quelle est l'instance la plus probable?

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



Opérations sur les modèles graphiques

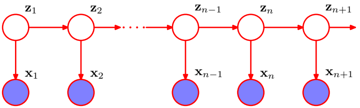
L'Inférence

Calculer une marginale $p(x_i)$ ou $p(x_i|x_1 = 3, x_7 = 0)$

Le Décodage

Quelle est l'instance la plus probable?

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



L'Apprentissage (ou Estimation)

Soit $p(x; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_C \psi(x_C, \theta_C)$, on veut trouver

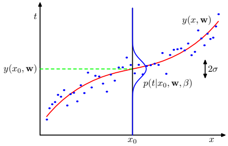
$$\text{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(x^{(i)}; \boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{\theta}} \frac{1}{Z(\boldsymbol{\theta})} \prod_{i=1}^{n} \prod_{C} \psi(x_C^{(i)}, \theta_C)$$

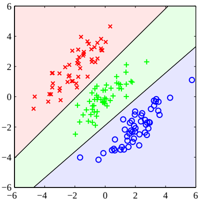
Plan du cours

- **Cours 1**
  Introduction
  Maximum de vraisemblance
  Modèles a un noeud
- **Cours 2**
  Régression lineaire
  Régression logistique
  Classification générative
- **Cours 3**
  K-means
  EM
  Mélanges de Gaussiennes
  Théorie des graphes
- **Cours 4**
  Modèles graphiques orientés
  Modèles graphiques non orientés
- **Cours 5**
  Familles exponentielles
  Théorie de l'information
  Mélanges d'experts
- **Cours 6**
  Variables Gaussiennes
  Analyse factorielle
- **Cours 7**
  Algorithme somme-produit
- **Cours 8**
  Inférence approchée
- **Cours 9**
  Méthodes Bayésiennes
  Sélection de Modèle

Pour commencer: modèles à 1 et 2 noeuds...

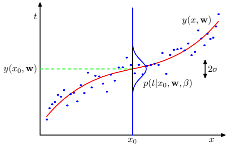
Modèles de régression et classification

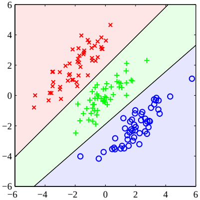




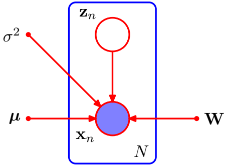
Pour commencer: modèles à 1 et 2 noeuds...

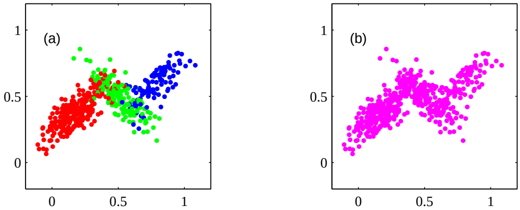
Modèles de régression et classification





Modèles de mixture





Notions transverses

- Modèle *génératif* vs *discriminatif*

Notions transverses

- Modèle *génératif* vs *discriminatif*
- Apprentissage *supervisé* vs *non-supervisé*

Notions transverses

- Modèle *génératif* vs *discriminatif*
- Apprentissage *supervisé* vs *non-supervisé*
- Apprentissage à partir de données *complètement observées* vs *incomplètes*

Notions transverses

- Modèle *génératif* vs *discriminatif*
- Apprentissage *supervisé* vs *non-supervisé*
- Apprentissage à partir de données *complètement observées* vs *incomplètes*
- *Causation* vs *corrélations*:
  Les MG ne sont **pas** des modèles de **causation** $\rightarrow$ des modèles de **corrélation** encodant des ensembles d'hypothèses d'**indépendance conditionelles**.

Outline

Notations, formules,définitions

- Loi jointe de X_A et X_B: $p(x_A, x_B)$
- Loi marginale : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Loi conditionnelle : $p\left(x_A | x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$

Notations, formules,définitions

- Loi jointe de X_A et X_B: $p(x_A, x_B)$
- Loi marginale : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Loi conditionnelle : $p\left(x_A | x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$

Formule de Bayes

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

Notations, formules,définitions

- Loi jointe de X_A et X_B: $p(x_A, x_B)$
- Loi marginale : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Loi conditionnelle : $p\left(x_A | x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$

Formule de Bayes

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

$\rightarrow$ La formule de Bayes n'est pas "bayésienne".

Espérances et Variances

- Espérance de X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$

Espérances et Variances

- Espérance de X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Espérance de $f(X)$, pour f mesurable :

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

Espérances et Variances

- Espérance de X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Espérance de $f(X)$, pour f mesurable :

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

- Variance :

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

Espérances et Variances

- Espérance de X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Espérance de $f(X)$, pour f mesurable :

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

- Variance :

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

- Espérance conditionnelle de X sachant Y :

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

Espérances et Variances

- Espérance de X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Espérance de $f(X)$, pour f mesurable :

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

- Variance :

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

- Espérance conditionnelle de X sachant Y :

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

- Variance conditionnelle :

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

Notions d'indépendance

Indépendance: $X \perp\!\!\!\perp Y$

On dit que X et Y sont indépendantes et on note $X \perp\!\!\!\perp Y$ ssi:

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

Notions d'indépendance

Indépendance: $X \perp\!\!\!\perp Y$

On dit que X et Y sont indépendantes et on note $X \perp\!\!\!\perp Y$ ssi:

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

Indépendance conditionnelle: $X \perp\!\!\!\perp Y \mid Z$

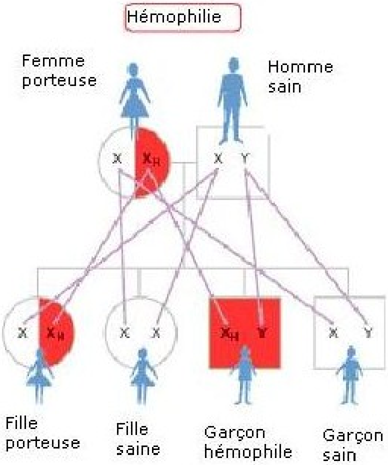
- On dit que X et Y sont indépendantes conditionnellement à Z et
- on note $X \perp\!\!\!\perp Y \mid Z$ ssi:

$\forall x, y, z,$

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

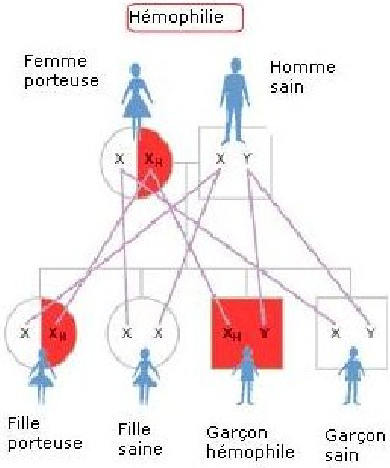
Indépendance Conditionnelle: exemple

Exemple d'une "maladie récessive liée à l'X":
Transmission du gène de l'hémophilie



Indépendence Conditionnelle: exemple

Exemple d'une "maladie récessive liée à l'X":
Transmission du gène de l'hémophilie



Risques de maladie pour des fils d'un père sain:

- dépendant pour deux frères.
- conditionnellement indépendant sachant si la mère est porteuse ou non.

Modèle statistique

Modèle paramétrique – Définition:

Ensemble de distributions paramétré par un vecteur $\theta \in \Theta \subset \mathbb{R}^p$

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

Modèle statistique

Modèle paramétrique – Définition:

Ensemble de distributions paramétré par un vecteur $\theta \in \Theta \subset \mathbb{R}^p$

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

Modèle de Bernoulli: $X \sim \mathrm{Ber}(\theta)$ $\quad \Theta = [0, 1]$

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

Modèle statistique

Modèle paramétrique – Définition:

Ensemble de distributions paramétré par un vecteur $\theta \in \Theta \subset \mathbb{R}^p$

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

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

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

Modèle Binomial : $X \sim \mathrm{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

Modèle statistique

Modèle paramétrique – Définition:

Ensemble de distributions paramétré par un vecteur $\theta \in \Theta \subset \mathbb{R}^p$

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

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

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

Modèle Binomial : $X \sim \mathrm{Bin}(n, \theta) \qquad \Theta = [0, 1]$

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{(1-x)}$$

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

$$p(x|\theta) = \binom{n}{x_1, \ldots, x_k} {\pi_1}^{x_1} \ \ldots \ {\pi_k}^{x_k}$$

Modèle gaussien

Loi gaussienne réelle : $X \sim \mathcal{N}(\mu, \sigma^2)$

X une v.a. réelle, et $\theta = \left(\mu, \sigma^2\right) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

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

Modèle gaussien

Loi gaussienne réelle : $X \sim \mathcal{N}(\mu, \sigma^2)$

X une v.a. réelle, et $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

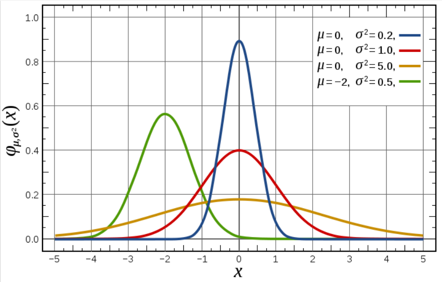
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Loi gaussienne multidimensionelle: $X \sim \mathcal{N}(\mu, \Sigma)$

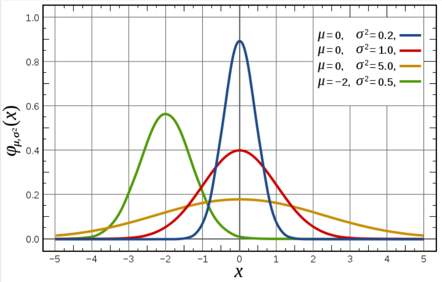
X v.a. à valeur dans $\mathbb{R}^d$. Si $\mathcal{K}_n$ est ensemble des matrices $n \times n$ définies positives, et $\theta = (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \mathcal{K}_n$.

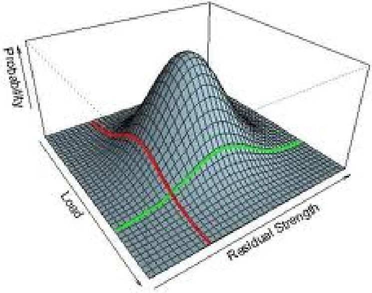
$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Densités gaussiennes



Densités gaussiennes





Principe du Maximum de vraisemblance

- Soit un modèle $\mathcal{P}_\Theta = \left\{ p(x|\theta) \mid \theta \in \Theta \right\}$
- Soit une observation x

Principe du Maximum de vraisemblance

- Soit un modèle $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$
- Soit une observation x

Vraisemblance:

$$\begin{aligned} \mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Principe du Maximum de vraisemblance

- Soit un modèle $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$
- Soit une observation x

Vraisemblance:

$$\begin{aligned} \mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta) \end{aligned}$$

Estimateur du maximum de vraisemblance:

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

Sir Ronald Fisher
(1890-1962)

Principe du Maximum de vraisemblance

- Soit un modèle $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$
- Soit une observation x

Sir Ronald Fisher
(1890-1962)

Vraisemblance:

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

Estimateur du maximum de vraisemblance:

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

Cas de données i.i.d

Pour $(x_i)_{1 \leq i \leq n}$ un *échantillon* de données i.i.d de taille n:

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

Exemples de calculs de l'EMV

- Modèle de Bernoulli
- Modèle Multinomial
- Modèle Gaussien

Estimation bayésienne

On traîte le paramètre θ comme une variable aléatoire.

A priori

On dispose d'un *a priori* $p(\theta)$ sur les paramètres du modèle.

Estimation bayésienne

On traîte le paramètre θ comme une variable aléatoire.

A priori

On dispose d'un *a priori* $p(\theta)$ sur les paramètres du modèle.

A posteriori

Les observations contribuent via la vraisemblance: $p(x|\theta)$.
La probabilité *a posteriori* du modèle est alors

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

$\rightarrow$ L'estimateur bayésien est donc une distribution de probabilité sur les paramètres.

On parle d'*inférence* bayésienne.

Références

- Une bonne partie des illustrations et des notations de cet exposé proviennent du livre de Christopher Bishop:

  Pattern Recognition and Machine Learning, 2006, Springer.

  `http://research.microsoft.com/~cmbishop/PRML/`