

IFT3395

Fondements de l'apprentissage machine

**Brève introduction aux
modèles graphiques probabilistes**

**Première partie: modèles dirigés
(réseaux Bayesiens)**

Professeur: Pascal Vincent

Une partie du matériel pour ce cours est inspiré/emprunté/traduit d'une présentation d'Aaron Courville ainsi que de l'excellente introduction de Kevin Murphy:

A Brief Introduction to Graphical Models and Bayesian Networks

<http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html>

que je vous invite à consulter pour plus de précisions sur le sujet.

Rappel: opérations avec les distributions

Avec une distribution, on peut vouloir:

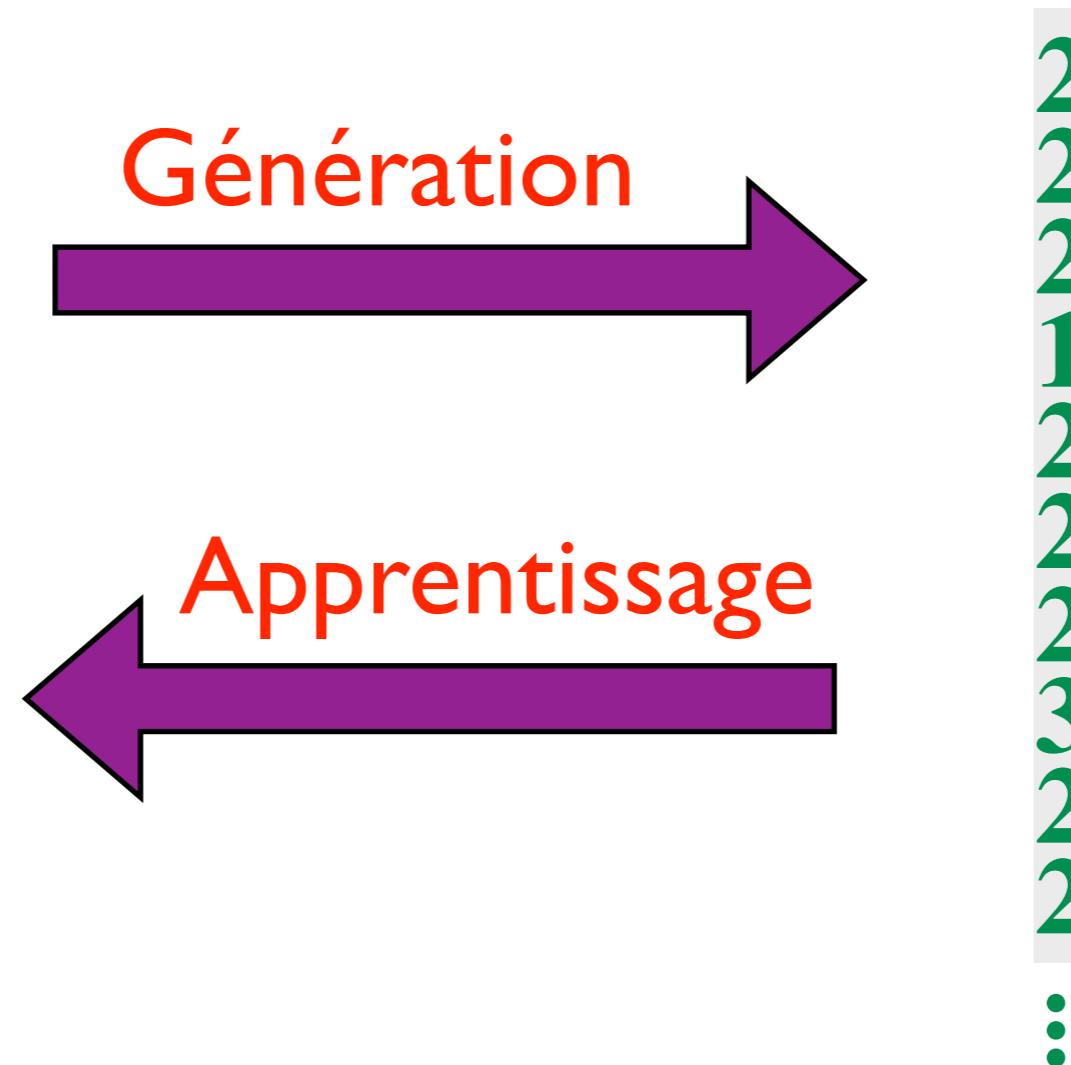
- **Générer des données**, c.a.d. tirer des échantillons selon la distribution.
- **Calculer la (log) probabilité d'une configuration** (sachant la valeur de certaines des variables et ayant maginalisé celles dont on ne connaît pas la valeur).
- **Inférence**: inférer la valeur la plus probable ou la valeur espérée d'un sous ensemble de variables sachant la valeur des autres.
- **Apprentissage** des paramètres de la distribution **à partir d'un ensemble de données** (de sorte à maximiser la probabilité que les données soient générées par cette distribution avec ces paramètres: principe du *maximum de vraisemblance*).

Ex. variable scalaire discrète X

Table de probabilité:

x	$P(X=x)$
1	0.10
2	0.80
3	0.10

Ensemble de données:



Modèles graphiques probabilistes

- Utiles pour des distributions multivariées (dimension $d > 1$)
 $P(X_1, X_2, \dots, X_d)$
- Ils définissent une *structure* de la relation entre ces variables aléatoires:
 - ➡ Les relations d'**indépendance conditionnelle** entre les variables sont représentées par un **graphe**. Chaque variable est un noeud.
- Deux grandes familles de modèles:
 - ➡ Graphes orientés (arcs): **directed graphical models**
= Réseaux Bayesiens (*Bayes Nets*)
 - ➡ Graphes non orientés (arêtes): **undirected graphical models**
= Champs aléatoires de Markov (*Markov Random Fields*)

Rappel de probabilités: Indépendance marginale

Définition: X est *marginalement* indépendant de Y si pour tout (i, j)

$$\begin{aligned} P(X = x_i, Y = y_j) &= P(X = x_i)P(Y = y_j) \\ P(X, Y) &= P(X)P(Y) \end{aligned}$$

Ou de manière équivalente:

$$P(X \mid Y) = P(X) \quad P(Y \mid X) = P(Y)$$

Y ne nous apprend rien sur X,
et X ne nous apprend rien sur Y

Rappel de probabilités: Indépendance conditionnelle

Définition: X est *conditionnellement* indépendant de Y sachant Z si la loi de probabilité de X est indépendante de la valeur de Y lorsqu'on connaît la valeur de Z . Pour tout (i, j, k) :

$$\begin{aligned} P(X = x_i, Y = y_j \mid Z = z_k) &= P(X = x_i \mid Z = z_k)P(Y = y_j \mid Z = z_k) \\ P(X, Y \mid Z) &= P(X \mid Z)P(Y \mid Z) \end{aligned}$$

Ce qui équivaut à :

$$P(X \mid Y, Z) = P(X \mid Z) \quad P(Y \mid X, Z) = P(Y \mid Z)$$

Quand on connaît la valeur de Z ,
 Y ne nous apprend rien de plus sur X ,
et X ne nous apprend rien de plus sur Y

Modèles graphiques dirigés

- Ensemble de noeuds avec des flèches (arcs) entre certains des noeuds
 - Les noeuds représentent des variables aléatoires
 - Les flèches indiquent une factorisation des probabilités conditionnelles
- Considérons une distribution jointe arbitraire:

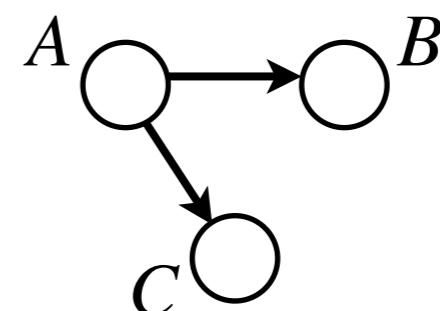
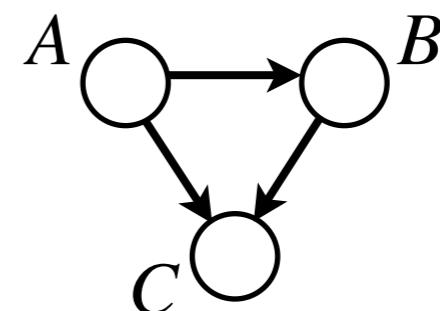
$$P(A, B, C)$$

- On peut *toujours* la factoriser ainsi:

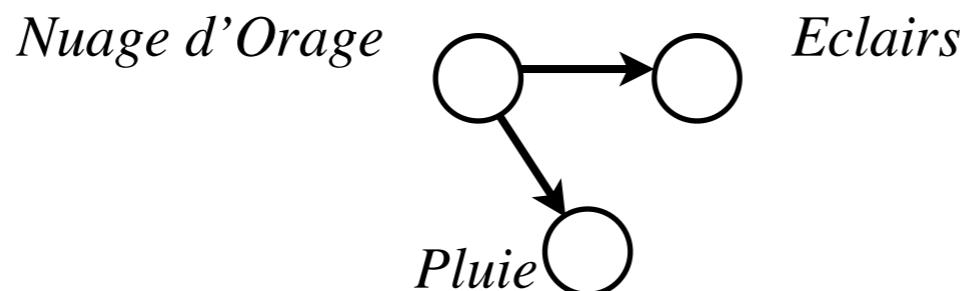
$$\begin{aligned} P(A, B, C) &= P(A)P(B, C | A) \\ &= P(A)P(B | A)P(C | A, B) \end{aligned}$$

- Si en plus C est conditionnellement indépendant de B :

$$P(A, B, C) = P(A)P(B | A)P(C | A)$$



- Ex: $P(\text{Pluie}, \text{Eclairs} | \text{Nuage d'orage}) = P(\text{Pluie} | \text{Nuage d'orage}) P(\text{Eclairs} | \text{Nuage d'orage})$

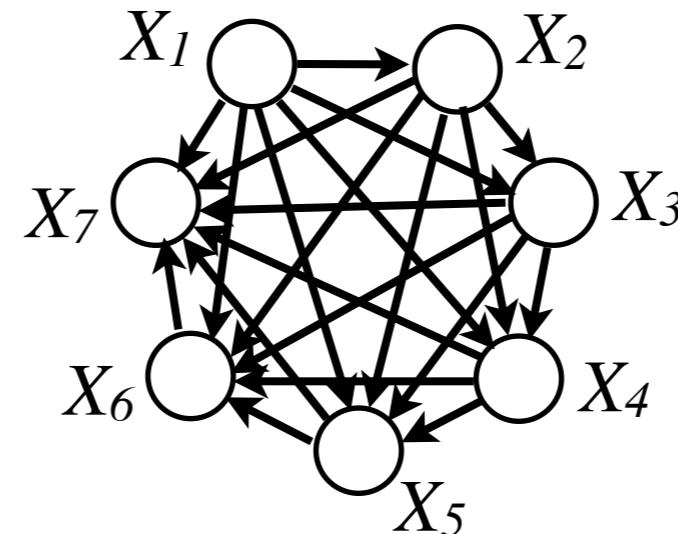


Cas général

- Une distribution jointe arbitraire $P(X_1, X_2, \dots, X_d)$ peut **toujours** se factoriser ainsi:

$$P(X_1, X_2, \dots, X_d) = P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots P(X_d | X_1, \dots, X_{d-1})$$

- Ceci peut être représenté par un graphe *complètement connecté* où chaque noeud reçoit un arc depuis tous les noeuds qui le précèdent (dans un ordre de numérotation).



- Remarque: d'autres ordres de numérotation donneraient d'autres factorisations valides. L'ordre de numérotation est arbitraire!
- Il s'agit d'un graphe orienté acyclique (DAG).

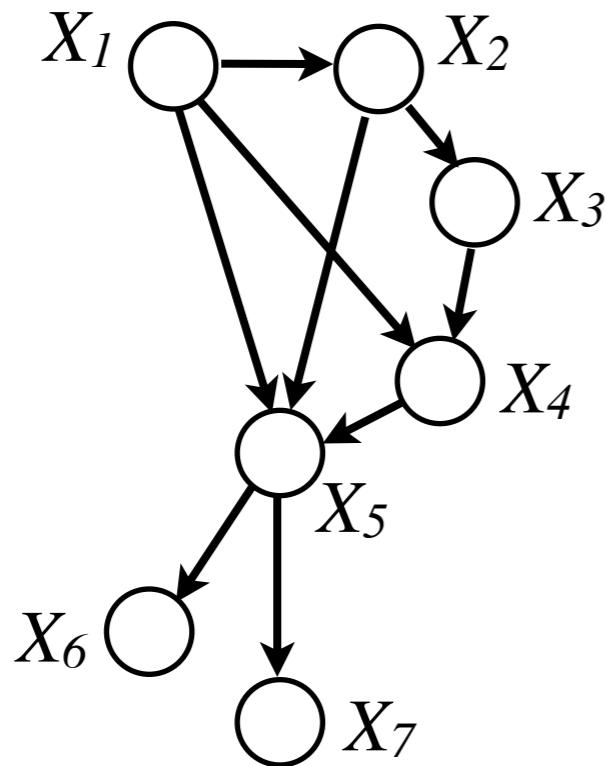
Que représente un graphe ? (acyclique, complètement connecté ou non)

- Il représente une factorisation particulière de la probabilité jointe:

$$P(X_1, \dots, X_d) = \prod_{i=1}^d P(X_i | \text{Pa}_i)$$

où Pa_i est l'ensemble des **parents** immédiats du noeud X_i .

- Ex:



Terminologie:

- Parents** d'un noeud i = noeuds desquels part une flèche vers i .
- Ancêtres** d'un noeud i = parents, parents des parents, etc...
- Enfants** d'un noeud i = noeuds sur lesquels pointe un flèche partant de i
- Descendants** = enfants, enfants des enfants, etc...

$$\begin{aligned} P(X_1, \dots, X_7) &= P(X_1)P(X_2|X_1)P(X_3|X_2) \\ &\quad P(X_4|X_3, X_1)P(X_5|X_1, X_2, X_4) \\ &\quad P(X_6|X_5)P(X_7|X_5) \end{aligned}$$

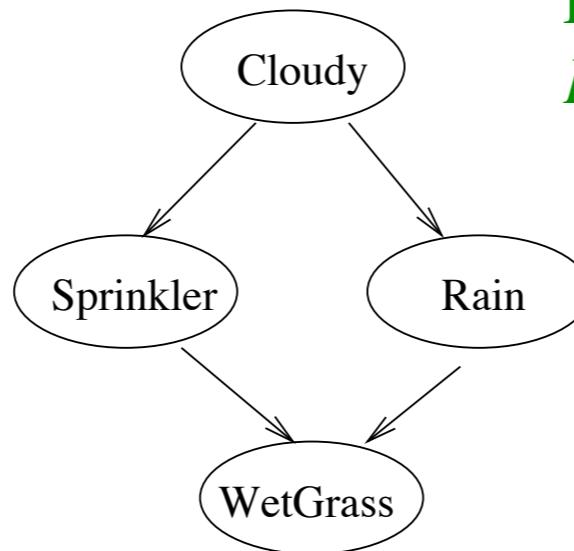
- Sachant ses parents immédiats, X_i est indépendant de tous ses non-descendants.
(représente les indépendance conditionnelle)

Spécification complète de la distribution

- Le graphe ne dit pas tout (il indique seulement les relations d'indépendance conditionnelles).
- Il reste à spécifier chacune des lois $P(X_i | \text{Pa}_i)$ (les facteurs de la jointe)
- Ex: tables de probabilité pour des variables catégoriques

T=True=1
F=False=0

	$P(C=F)$	$P(C=T)$
	0.5	0.5



Le graphe indique la factorisation:

$$P(W,S,R,C) = P(W|R,S) P(S|C) P(R|C) P(C)$$

C	$P(S=F)$	$P(S=T)$
F	0.5	0.5
T	0.9	0.1

C	$P(R=F)$	$P(R=T)$
F	0.8	0.2
T	0.2	0.8

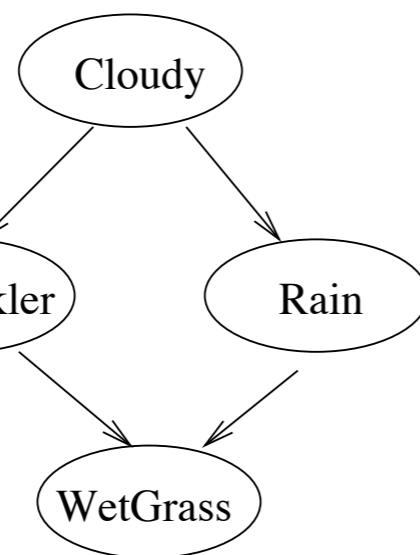
S	R	$P(W=F)$	$P(W=T)$
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

L'histoire générative

- Le graphe donne aussi une **procédure pour «générer»** des données de la distribution jointe $P(X_1, X_2, \dots, X_d)$ si on sait générer depuis les lois $P(X_i | \text{Pa}_i)$
- Il faut suivre un ordonnancement partiel du graphe

$$\begin{array}{c} P(C=F) \quad P(C=T) \\ \hline \end{array}$$

0.5 0.5



Le graphe indique la factorisation:

$$P(W,S,R,C) = P(W|R,S) P(S|C) P(R|C) P(C)$$

C	$P(S=F) \quad P(S=T)$	
F	0.5	0.5
T	0.9	0.1

C	$P(R=F) \quad P(R=T)$	
F	0.8	0.2
T	0.2	0.8

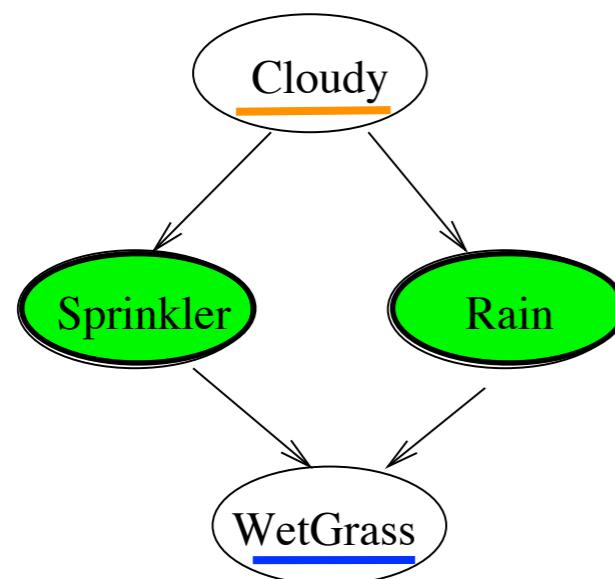
S	R	$P(W=F) \quad P(W=T)$	
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Pour générer un échantillon $x=(w,s,r,c)$:

- $c \sim P(C)$
- $s \sim P(S|C=c)$
- $r \sim P(R|C=c)$
- $w \sim P(W|R=r, S=s)$

Le problème de l'inférence

- L'inférence consiste à prédire un sous-ensemble des variables sachant les valeurs d'un autre sous-ensemble des variables.
- c.a.d. connaissant les paramètres du modèle, calculer $P(\text{noeuds à prédire} \mid \text{noeuds observés})$
- Les variables conditionnantes sont dites **observées** ou **visibles** (on connaît leur valeur) habituellement représentées par des **noeuds pleins**.
- Les variables qui ne sont ni observées, ni celles à prédire sont dites **variables latentes** ou **cachées**.
- Ex: $P(W=1 \mid S=1, R=0)$



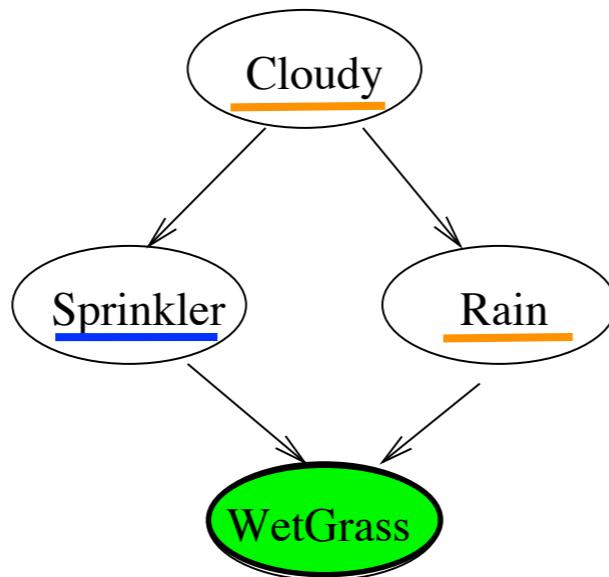
T=True=1
F=False=0

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

- **Un cas facile** car c'est une des distributions conditionnelles qui nous est donnée!

Le problème de l'inférence

- Connaissant les paramètres du modèle,
calculer $P(\text{variables à prédire} \mid \text{variables observées})$
- Ex: $P(S=1 \mid W=1)$



$$P(S = 1 \mid W = 1) = \frac{P(S = 1, W = 1)}{P(W = 1)} = \frac{\sum_{c,r} P(C = c, S = 1, R = r, W = 1)}{P(W = 1)} = \frac{0.2781}{0.6471} = 0.430$$

Le problème de l'inférence

Soient

- X l'ensemble des variables observées, et x leurs valeurs
- Y la ou les variables à prédire, et \mathcal{Y} l'ensemble des combinaisons de valeurs qu'elles peuvent prendre.
- Z la ou les variables latentes, et \mathcal{Z} l'ensemble des combinaisons de valeurs qu'elles peuvent prendre.

- On cherche $P(Y=y|X=x)$

• Solution générale:

$$\begin{aligned} P(Y = y|X = x) &= \frac{P(Y = y, X = x)}{P(X = x)} \\ &= \frac{\sum_{z \in \mathcal{Z}} P(Y = y, X = x, Z = z)}{\sum_{y' \in \mathcal{Y}} \sum_{z' \in \mathcal{Z}} P(Y = y', X = x, Z = z')} \end{aligned}$$

On sait calculer la jointe

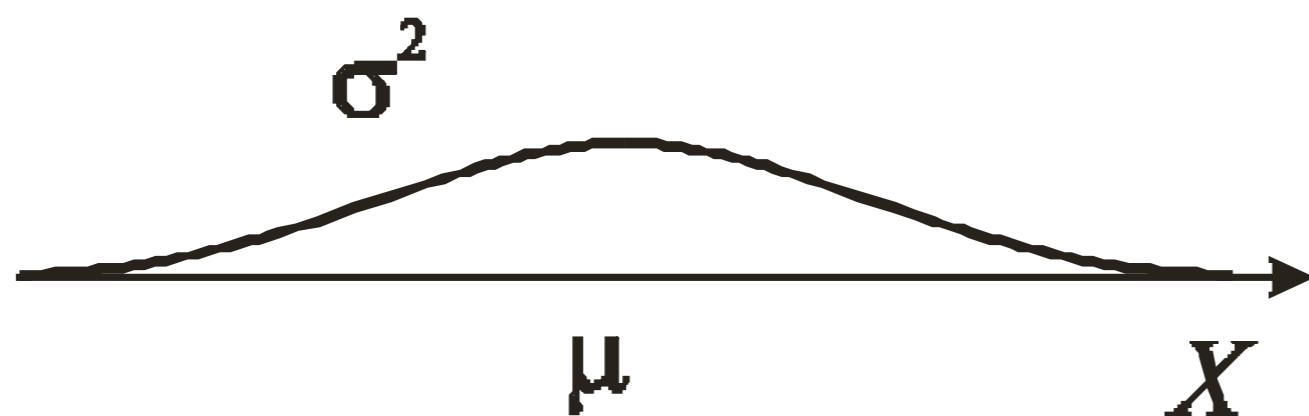
Ex: si $Z \in \{0, 1\}^{d_Z}$ alors 2^{d_Z} configurations.

- La structure du graphe et la nature des variables permet parfois de faire le calcul exact efficacement.
- Sinon techniques d'inférence approximative:
méthodes Variationnelles ou techniques d'échantillonnage Monte-Carlo
- Pour des variables continues: remplacer les sommes de probas par des intégrales de densités (p.d.f.)

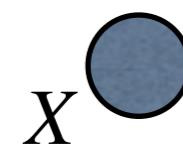
Exemple: Gaussienne univariée

- On peut considérer une variable aléatoire Gaussienne comme un (simple) modèle graphique.

$$p(X = x) = \mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

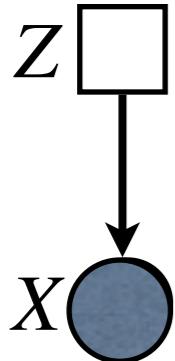


Modèle graphique:

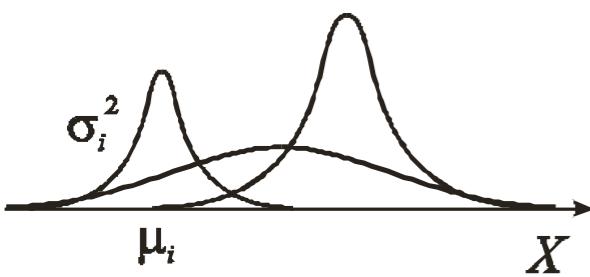


Exemple II: Mélange de Gaussiennes

- Modèle graphique:



carré: variable catégorique
rond: variable continue



- Distributions conditionnelles:

$$P(Z = i) = w_i$$

$$p(X = x | Z = i) = \mathcal{N}(x | \mu_i, \sigma_i^2) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right)$$

$i \in \{1, \dots, K\}$ est l'indice de la composante Gaussienne du mélange.
 w_i est la probabilité marginale que X soit générée par la i^{eme} composante (la i^{eme} Gaussienne).

- Distribution jointe: (je me permets ici un abus de notation)

$$p(Z = i, X = x) = P(Z = i)p(X = x | Z = i) = w_i \mathcal{N}(x | \mu_i, \sigma_i^2)$$

- Distribution marginale:

$$p(X = x) = \sum_{i=1}^K p(X = x | Z = i)P(Z = i) = \sum_{i=1}^K w_i \mathcal{N}(x | \mu_i, \sigma_i^2)$$

- Inférence: (règle de Bayes)

$$P(Z = i | X = x) = \frac{p(X = x | Z = i)P(Z = i)}{\sum_{i'} p(X = x | Z = i')P(Z = i')}$$

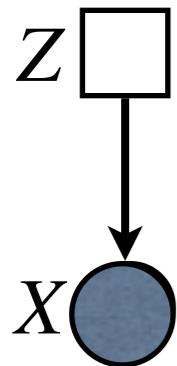
$$= \frac{w_i \mathcal{N}(x | \mu_i, \sigma_i^2)}{\sum_{i'} w_{i'} \mathcal{N}(x | \mu'_{i'}, \sigma'^2_{i'})}$$

- Histoire générative

- ▶ on choisit une des K Gaussiennes i selon $P(Z)$
- ▶ on génère x selon cette Gaussienne i .

Exemple II: Mélange de Gaussiennes

- Modèle graphique:



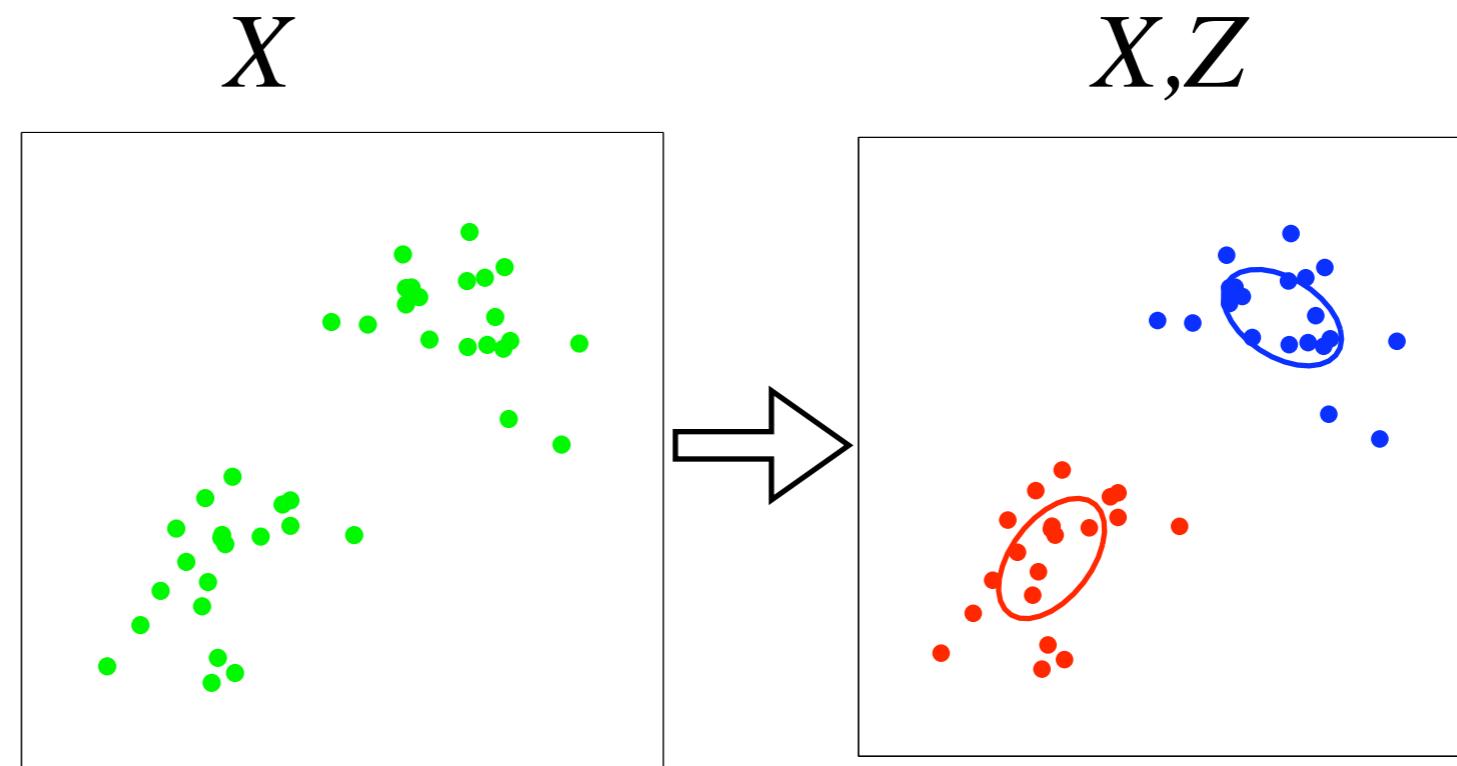
Si $X \in \mathbb{R}^d$:

Mélange de K Gaussiennes **multivariées**

Paramètres θ : moyenne μ_i et matrice de covariance Σ_i de chaque Gaussienne

carré: variable catégorique

rond: variable continue

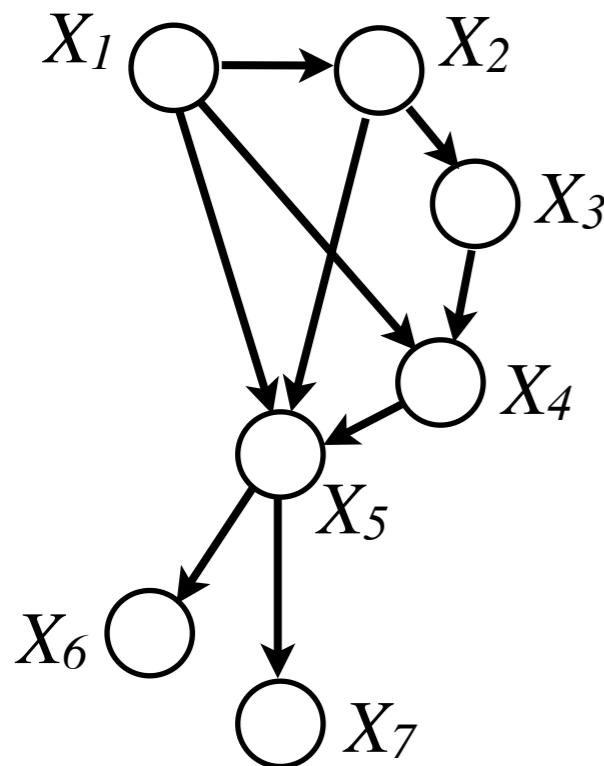


Apprentissage des paramètres

cas entièrement observé

- Si on a un modèle et des données D_n où toutes les variables du modèle sont observées
- L'apprentissage des paramètres par maximum de vraisemblance est facile
- On peut apprendre les paramètres de chacun des facteurs (les probabilités conditionnelles du graphe) indépendamment (si leurs paramètres ne sont pas liés).

$$D_n = \{x^{(1)}, \dots, x^{(n)}\}, \quad x^{(t)} = (x_1^{(t)}, \dots, x_d^{(t)}).$$



$$\begin{aligned} P(X_1, \dots, X_7) &= P(X_1)P(X_2|X_1)P(X_3|X_2) \\ &\quad P(X_4|X_3, X_1)P(X_5|X_1, X_2, X_4) \\ &\quad P(X_6|X_5)P(X_7|X_5) \end{aligned}$$

Ex: on peut apprendre les paramètres de $P(X_4|X_3, X_1)$ ainsi:

$$\theta_{4|3,1}^* = \arg \max_{\theta_{4|3,1}} \sum_{t=1}^n \log P(X_4 = x_4^{(t)} | X_3 = x_3^{(t)}, X_1 = x_1^{(t)})$$

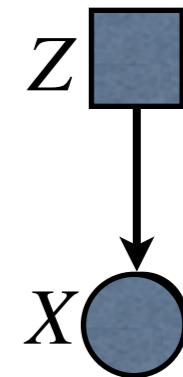
Apprentissage des paramètres

cas entièrement observé

- Ex: pour mélange de Gaussiennes

- **Modèle graphique:**

$$D_n = \{(x^{(1)}, z^{(1)}), \dots, (x^{(n)}, z^{(n)})\}, \quad x^{(t)} \in \mathbb{R}^d, \quad z^{(t)} \in \{1, \dots, K\}.$$



Maximum de vraisemblance:

$$\begin{aligned}
 & \arg \max_{\theta} \sum_{t=1}^n \log p(X = x^{(t)}, Z = z^{(t)}; \theta) \\
 &= \arg \max_{\theta} \sum_{t=1}^n \log p(Z = z^{(t)}) + \log p(X = x^{(t)} | Z = z^{(t)}) \\
 &= \arg \max_{\{w, \mu, \Sigma\}} \sum_{t=1}^n \log w_{z^{(t)}} + \log \mathcal{N}(x^{(t)} | \mu_{z^{(t)}}, \Sigma_{z^{(t)}}) \\
 \text{avec la contrainte } & w_k \geq 0 \text{ et } \sum_{k=1}^K w_k = 1
 \end{aligned}$$

carré: variable catégorique
rond: variable continue

Apprentissage des paramètres

cas entièrement observé

- Ex: pour mélange de Gaussiennes

$$D_n = \{(x^{(1)}, z^{(1)}), \dots, (x^{(n)}, z^{(n)})\}, \quad x^{(t)} \in \mathbb{R}^d, \quad z^{(t)} \in \{1, \dots, K\}.$$

$$\arg \max_{\{w, \mu, \Sigma\}} \sum_{t=1}^n \log w_{z^{(t)}} + \log \mathcal{N}(x^{(t)} | \mu_{z^{(t)}}, \Sigma_{z^{(t)}}) \quad \text{avec la contrainte} \quad w_k \geq 0 \text{ et } \sum_{k=1}^K w_k = 1$$

Solution:

$$\mu_k^*, \Sigma_k^* = \arg \max_{\mu_k, \Sigma_k} \sum_{t|z^{(t)}=k} \log \mathcal{N}(x^{(t)} | \mu_k, \Sigma_k)$$

μ_k^*, Σ_k^* = Solution du maximum de vraisemblance pour la Gaussienne k sur les exemples pour lesquels $t=k$.

$$n_k = \sum_{t=1}^n \mathbb{I}_{\{z^{(t)}=k\}}$$

$$w_k^* = \frac{n_k}{n}$$

$$\mu_k^* = \frac{1}{n_k} \sum_{t|z^{(t)}=k} x^{(t)}$$

$$\Sigma_k^* = \frac{1}{n_k} \sum_{t|z^{(t)}=k} (x^{(t)} - \mu_k^*)(x^{(t)} - \mu_k^*)^T$$

A quel algo cela correspond-t-il???

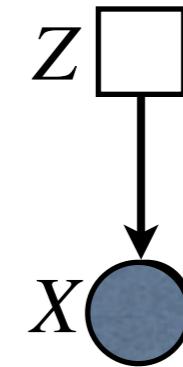
➡ Classifieur de
Bayes Gaussien!

Apprentissage des paramètres avec des variables latentes (non observées)

- Généralement plus difficile:
il faut marginaliser les variables latentes
(ou inférer leur distribution)
- À partir d'un ensemble d'entraînement
 $D_n = \{x^{(1)}, \dots, x^{(n)}\}$, $x^{(t)} \in \mathbb{R}^d$.
- Principe du maximum de vraisemblance

$$\arg \max_{\theta} \sum_{t=1}^n \log p(X = x^{(t)}; \theta)$$

- Modèle graphique:



Ex: mélange de Gaussiennes (on peut exprimer explicitement la marginale $p(X)$):

$$= \arg \max_{\theta} \sum_{t=1}^n \log \sum_{z=1}^K w_z \mathcal{N}(x^{(t)} | \mu_z, \Sigma_z)$$

- On peut faire de la descente de gradient directement sur cet objectif
- Ou on peut utiliser l'algorithme E.M. (Expectation Maximization)

Expectation Maximization

[Dempster, Laird and Rubin, 1977]

- Soit Z la ou les variables latentes
- Le principe d'EM est simple:

Si on connaissait les Z , l'apprentissage (maximisation de vraisemblance) serait facile.

On va donc **itérer**:

- Inférer la distribution des Z pour chaque exemple X étant donné les paramètres *actuels* du modèle
- Faire comme si les Z étaient observés, suivant vraiment cette distribution,
et trouver les **paramètres qui maximisent** alors la **vraisemblance**.

- L'algorithme E.M. met à jour les paramètres selon la formule:

$$\theta^{\text{nouveau}} = \arg \max_{\theta} \sum_{t=1}^n \mathbb{E}_{p(Z|X=x^{(t)}; \theta^{\text{actuels}})} [\log p(X = x^{(t)}, Z; \theta)]$$

EXPECTATION

Espérance de la log vraisemblance des $x^{(t)}, z^{(t)}$
selon l'estimé actuel des $z^{(t)} | x^{(t)}$.

MAXIMISATION

Expectation Maximization

- Ex: Mélange de Gaussienne

- Inférer la distribution des Z pour chaque exemple X étant donné les paramètres actuels du modèle

$$\begin{aligned}
 P(Z|X = x) &= (P(Z = 1|X = x), \dots, P(Z = K|X = x)) \\
 &= \left(\frac{P(X = x|Z = 1)P(Z = 1)}{P(X = x)}, \dots, \frac{P(X = x|Z = K)P(Z = K)}{P(X = x)} \right) \\
 &= (w_1\mathcal{N}(x|\mu_1, \Sigma_1), \dots, w_K\mathcal{N}(x|\mu_K, \Sigma_K)) \frac{1}{\sum_{j=1}^K w_j\mathcal{N}(x|\mu_j, \Sigma_j)}
 \end{aligned}$$

- On calcule les «responsabilités» $\mathbf{R}_{ti} = P(Z = i|X = x^{(t)})$ avec les valeurs actuelles des paramètres.
Cas limite instructif: si on fixe $w_i = 1/K$ et $\Sigma_i = \epsilon I$ avec $\epsilon \rightarrow 0$,
alors $\mathbf{R}_t = P(Z|X = x^{(t)}) \rightarrow \text{onehot}(\arg \min(\|x^{(t)} - \mu_1\|^2, \dots, \|x^{(t)} - \mu_K\|^2))$
- Faire comme si les (X, Z) étaient entièrement observés, suivant vraiment cette distribution:
comme si, correspondant à chaque exemple $x^{(t)}$, on observait une proportion \mathbf{R}_{ti} d'exemples $(X=x^{(t)}, Z=i)$.
- Trouver les paramètres qui maximisent alors la vraisemblance.

$$\begin{aligned}
 \theta^{\text{nouveau}} &= \arg \max_{\theta} \sum_{t=1}^n \mathbb{E}_{P(Z|X=x^{(t)}; \theta^{\text{actuels}})} [\log p(X = x^{(t)}, Z; \theta)] \\
 \{\mu, \Sigma\}^{\text{nouveau}} &= \arg \max_{\theta} \sum_{t=1}^n \sum_{i=1}^K \underbrace{P(Z = i|X = x^{(t)}; \theta^{\text{actuels}})}_{\mathbf{R}_{ti}} \underbrace{\log p(X = x^{(t)}, Z = i; \theta)}_{\log w_i + \log \mathcal{N}(x|\mu_i, \Sigma_i)}
 \end{aligned}$$

Expectation Maximization

$$\{\mu, \Sigma\}^{\text{nouveau}} = \arg \max_{\theta} \sum_{t=1}^n \sum_{i=1}^K \mathbf{R}_{ti} (\log w_i + \log \mathcal{N}(x | \mu_i, \Sigma_i))$$

C'est une version pondérée du cas complètement observé vu précédemment.

SOLUTION:

$$n_i = \sum_{t=1}^n \mathbf{R}_{ti}$$

$$w_i^* = \frac{n_i}{\sum_{k=1}^K n_k}$$

$$\boxed{\mu_i^* = \frac{1}{n_i} \sum_{t=1}^n \mathbf{R}_{ti} x^{(t)}}$$

$$\Sigma_i^* = \frac{1}{n_i} \sum_{t=1}^n \mathbf{R}_{ti} (x^{(t)} - \mu_i^*) (x^{(t)} - \mu_i^*)^T$$

- Cas limite instructif: si on fixe $w_i=1/K$ et $\Sigma_i = \varepsilon I$ avec $\varepsilon \rightarrow 0$

alors $\boxed{\mathbf{R}_t = P(Z|X = x^{(t)}) \rightarrow \text{onehot}(\arg \min(\|x^{(t)} - \mu_1\|^2, \dots, \|x^{(t)} - \mu_K\|^2))}$

A quel algo cela correspond-t-il???

Apprentissage des modèles graphiques dirigés: résumé

- On a vu qu'on pouvait **apprendre les paramètres** d'un modèle
- Facile si toutes les variables sont observées
- Si il y a des **variables latentes** il faut pouvoir **efficacement**
 - soit les marginaliser (les ignorer en sommant sur toutes les combinaisons possibles de leur valeur)
 - soit inférer leur distribution, et calculer l'espérance sous cette distribution (algo E.M.)
 - Sinon possibilité de recourir à des méthodes approximatives (*variationnelles, Monte-Carlo, ...*).
- Il existe aussi des techniques qui tentent d'**apprendre la structure du graphe** (les arcs reliant les noeuds) à partir de données -> Domaine de recherche actif.

De nombreux algorithmes d'apprentissage peuvent être compris comme des modèles graphiques

