IFT6390
Fondements de l'apprentissage machine

A short introduction to Probabilistic Graphical Models

Directed PGM / Bayes nets

Professor: Ioannis Mitliagkas
Slides: Pascal Vincent

(largely inspired by a presentation from Aaron Courville and the very good introduction from Kevin Murphy:
A Brief Introduction to Graphical Models and Bayesian Networks
http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html
that I encourage you to read).

Reminder: operations on distributions

What do we want to do with a distribution:

- Generate data, i.e. draw samples from the distribution.
- Compute the (log) probability of a configuration (taking into account the values of known variables, and marginalizing over the unknown variables).
- Inference: *infer (i.e. predict)* the most likely or expected value of some of the variables knowing the values for other variables.
- Learn the parameters of the distribution from a data set drawn from the distribution (so as to maximize the probability of having generated the data from the distribution with these parameters: *Maximum Likelihood principle*).

Ex. discrete variable X

Probability table:

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

Generate

Learn

Dataset:

2
2
2
1
2
2
2
3
2
2
2
⋮

Probabilistic Graphical Models

- Useful for multivariate distributions (dimension $d>1$)
  $P(X_1 , X_2 , \dots , X_d)$

- Used to define *structure* on the relation between the random variables $X_1 , X_2 , \dots , X_d$:

➡ Conditional independence relations between the variables are represented by a **graph**.
Each variable is a node.

- Two main families of models:

➡ Directed graphical models
= Bayes Net

➡ Undirected graphical models
= Markov Random Fields (MRF)

Probability refresher: Marginal independence

Définition: X is marginally independent of Y if for all (i, j)

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$
$$P(X, Y) = P(X)P(Y)$$

Or equivalently:

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

Y gives no information about X,
and X gives no information about Y

Probability refresher: Conditional independence

Definition: X is *conditionally independent of* Y *given* Z if the distribution of X is independent of the value of Y when we know the value of Z.

Formally, for all *(i,j,k)*:

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

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

Or equivalently:

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

When we know the value of Z,
Y gives no information about X,
and X gives no information about Y

Probability refresher: Conditional independence

Definition: X is *conditionally independent of* Y *given* Z if the distribution of X is independent of the value of Y when we know the value of Z.

Formally, for all *(i,j,k)*:

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

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

Or equivalently:

Examples:

- X,Y are the outcomes of two dice rolls and Z=X+Y
(or, less extreme, Z is the parity of X+Y)

- Height and vocabulary of a person are not independent, but they become independent if you know the person's age.

Directed graphical models

- Directed graph: set of nodes with arrows/edges connecting some of the nodes
  - Nodes represent random variable
  - Arrows denotes a factorization of the conditional probabilities
- Consider some arbitrary joint distribution:

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

- We can always factorize it as:

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

- If in addition C is conditionally independent of B given A:

$$P(A, B, C) = P(A)P(B \mid A)P(C \mid A)$$

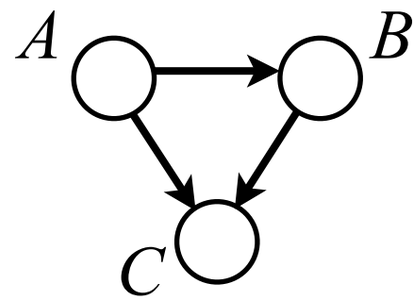
Directed graphical models

- Directed graph: set of nodes with arrows/edges connecting some of the nodes
  - Nodes represent random variable
  - Arrows denotes a factorization of the conditional probabilities
- Consider some arbitrary joint distribution:

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

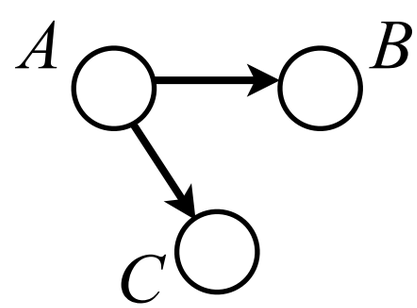
- We can always factorize it as:

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

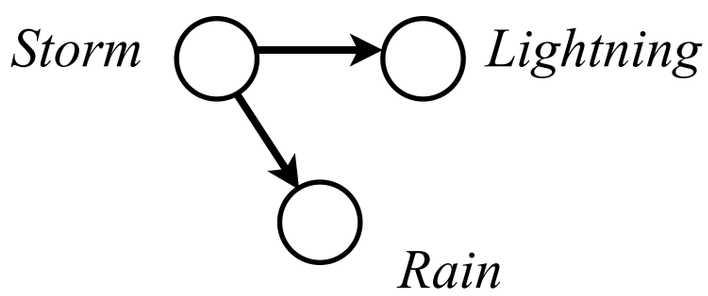


- If in addition C is conditionally independent of B given A:

$$P(A, B, C) = P(A)P(B \mid A)P(C \mid A)$$



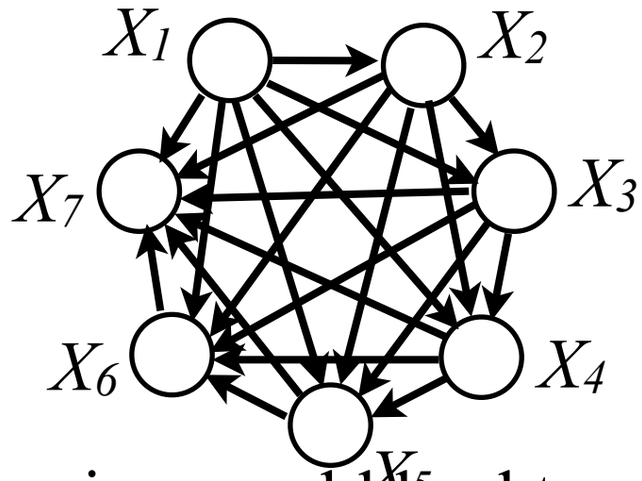
- **Ex:** *P(Rain, Lightning | Storm) = P(Rain | Storm) P(Lightning|Storm)*



General case

- Any joint distribution $P(X_1, X_2, \ldots, X_d)$
  can **always** be factorized as:
  $P(X_1, X_2, \ldots, X_d) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2) \ldots P(X_d \mid X_1, \ldots, X_{d-1})$

- This can be represented by a *complete graph* where each node is connected with arrows pointing to all nodes with a greater index (given some numbering of the nodes).



- Remark: any other choice of numbering would lead to a valid factorization of the joint distribution. The ordering is arbitrary!

- In Bayes nets, graphs are always directed acyclic graph (DAG).

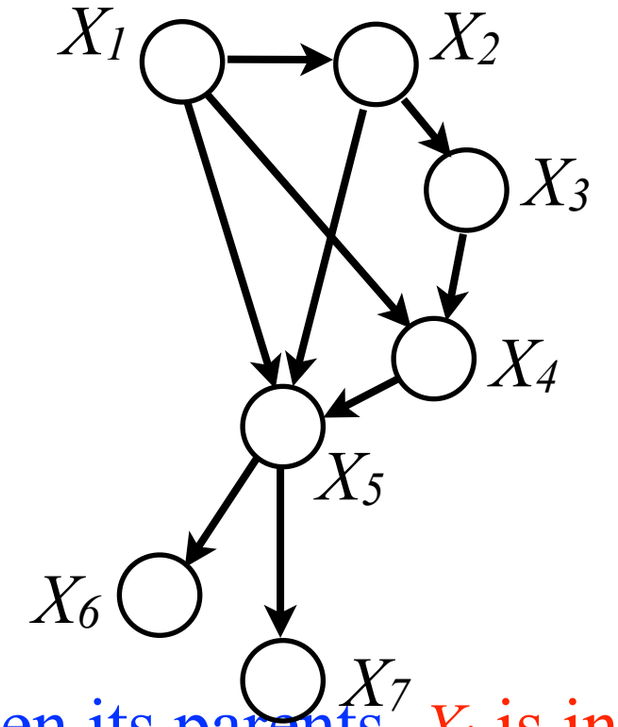
What does a Bayes net (DAG) represent ?

- It represents a particular factorization of the joint distribution:

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

where Pa_i is the set of parents of node X_i.

- Ex:



Terminology:

- Parents of a node i = nodes with an outgoing arrow pointing to i.
- Ancestors of a node i = parents, grandparents, etc...
- Children of a node i = nodes with an incoming arrow coming from i.
- Descendants = children, great children, etc...

$$\begin{aligned} P(X_1, \ldots, X_7) \quad = \quad & P(X_1)P(X_2|X_1)P(X_3|X_2) \\ & P(X_4|X3, X_1)P(X_5|X_1, X_2, X_4) \\ & P(X_6|X_5)P(X_7|X_5) \end{aligned}$$

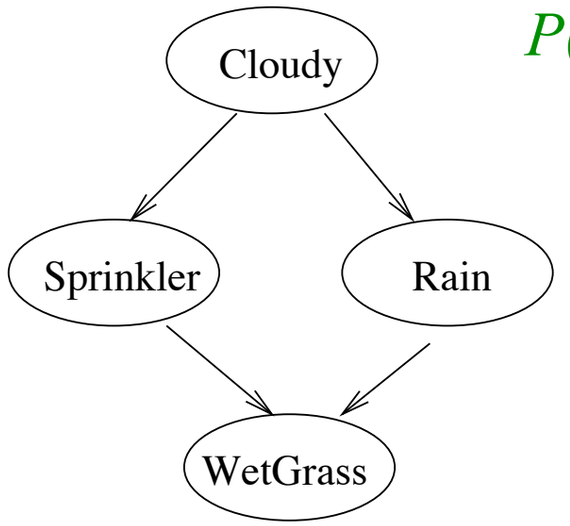
- Given its parents, X_i is independent of all its non-descendants. (represent the conditional independence relations)

Specifying the complete distribution

- The graph does not contain all the information (only the conditional independence relations).
- Still need to specify each of the distributions $P(X_i \mid \mathrm{Pa}_i)$ (the factors of the joint distribution)
- Ex: Probability tables for categorical variables.

T=True=1
F=False=0

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



This graph specify the factorization:
$P(W,S,R,C) = P(W|R,S)\ P(S|C)\ P(R|C)\ P(C)$

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

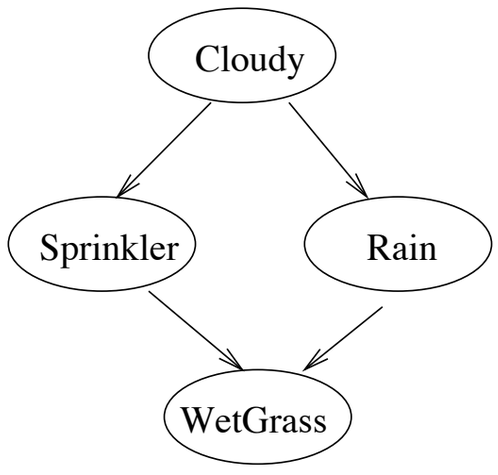
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Ex. emprunté à Kevin Murphy

Generating data

- The graph gives a procedure to sample data from the joint distribution $P(X_1, X_2, \ldots, X_d)$ if we know how to draw samples from each factor $P(X_i \mid \mathrm{Pa}_i)$
- Follow a bottom-up partial ordering: sample a value for each root (nodes without parents), then sample a value for all children of root nodes, etc.

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



This graph specify the factorization:
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

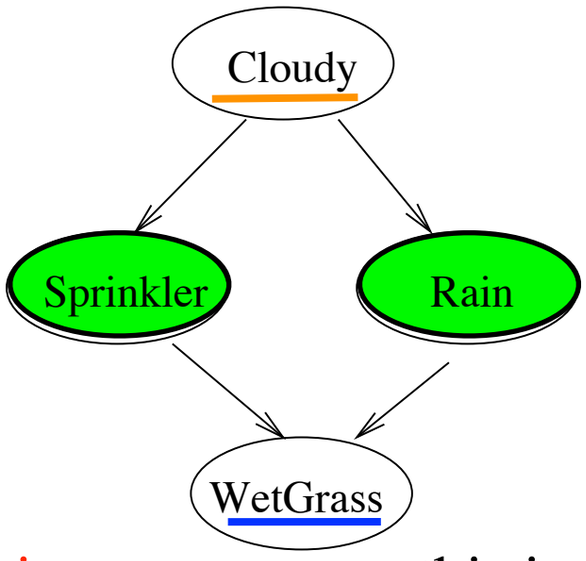
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

To draw a sample x=(w,s,r,c):

- *c~P(C)*
- *s~P(S|C=c)*
- *r~P(R|C=c)*
- *w~P(W|R=r,S=s)*

Inference problem

- The goal of *inference* is to predict the values of some subset of nodes, knowing the values of nodes in a another subset of nodes
- That is, knowing the model parameters (e.g. prob tables), compute *P*(nodes to predict | observed nodes)
- We call the condition variables observed or ”visibles” (we know their values), often represented by filled nodes in the graph.
- Variables that are not observed and that we don't need to predict or called latent/hidden variables.
- Ex: *P(W=1|S=1,R=0)*



T=True=1
F=False=0

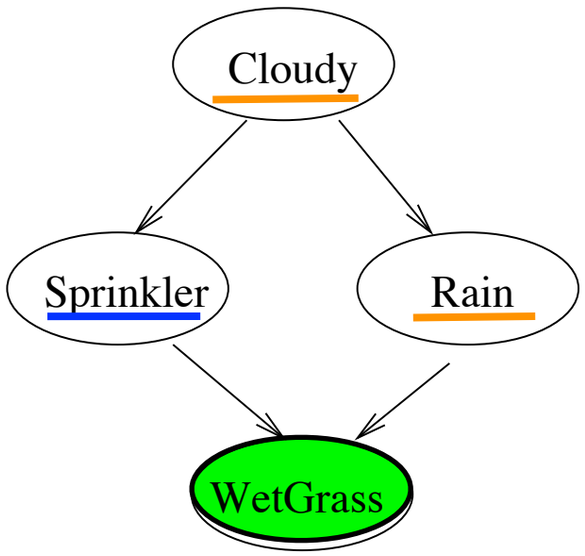
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

- This is an easy case: this is one of the conditionnal distribution we know directly from the table!

Inference problem

- Knowing the model parameters (e.g. prob. tables), compute *P*(nodes to predict | observed nodes)

- Ex: *P(S=1|W=1)*



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

Inference problem

Let
- X be the set of observed variables, and x their values.
- Y be the variable(s) we want to predict, and $\mathcal{Y}$ the set of values that Y can take.
- Z be the latent variable(s), and $\mathcal{Z}$ the set of values that Z can take.

- We want to compute *P(Y=y|X=x)*
- General solution:

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

We know how to compute the joint distribution!

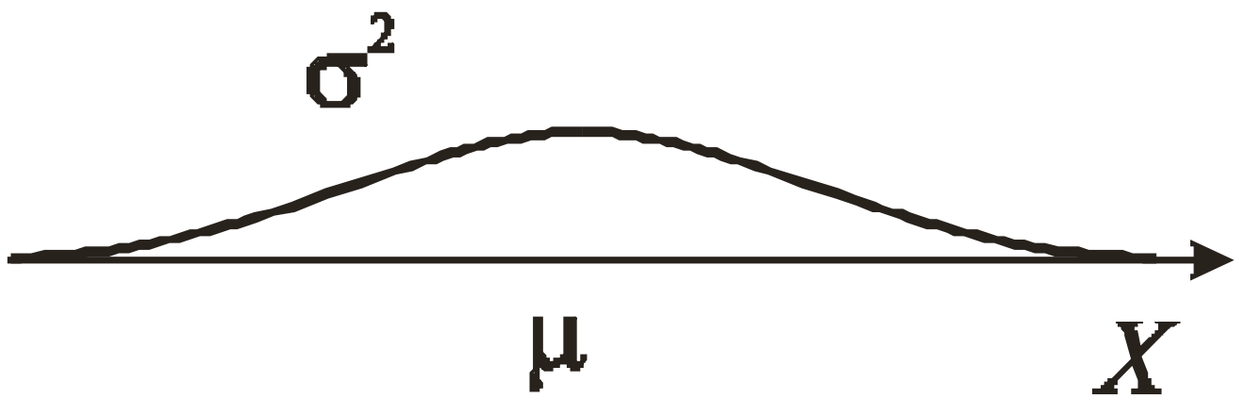
$\hookrightarrow$ If Z is a set of d binary variables, i.e. $Z \in \{0,1\}^d$, then there are 2^d configurations!

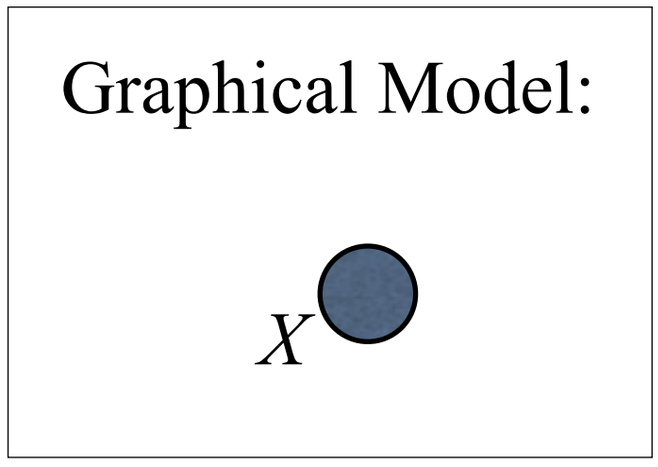
- The graph structure can sometimes help us to efficiently compute this exactly.
- Otherwise we can use approximate inference techniques:
  e.g. variational methods or MCMC sampling
- Note: for continuous variables, we replace sums of probabilities by integral of densities (p.d.f.)

Example: univariate Gaussian

- We can consider a Gaussian random variable as a simple (trivial) graphical model!

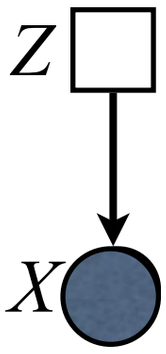
$$p(X = x) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$





Example II: Gaussian mixture

- **Graphical model:**



square: categorical variable
circle: continuous variable

- **Conditional distribution:**

parameters of the model

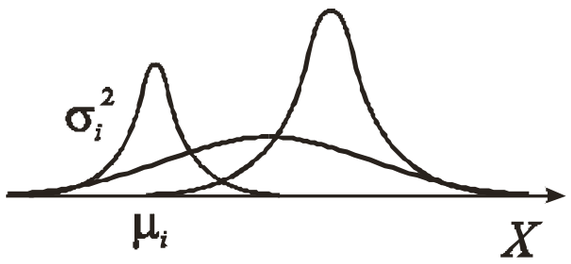
$$P(Z=i) = w_i$$

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

$i \in \{1, \ldots, K\}$ is the index for the component of the Gaussian mixture.
w_i is the probability to generate X from the *i*-th component (the *i*-th Gaussian).

- **Joint distribution:**

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



- **Marginal distribution:**

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

- **Generate data**
  - randomly choose one of the *K* Gausians, *i*, using *P(Z)*
  - Draw *x* from the *i*-th Gaussian.
  - ...

- **Inference:** (Bayes rule)

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

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

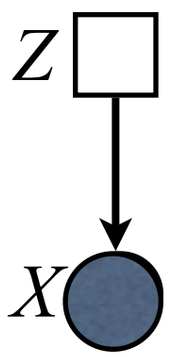
Example II: Gaussian mixture

If $X \in \mathbb{R}^d$: Mixture of K multivariate Gaussians
Parameters θ:

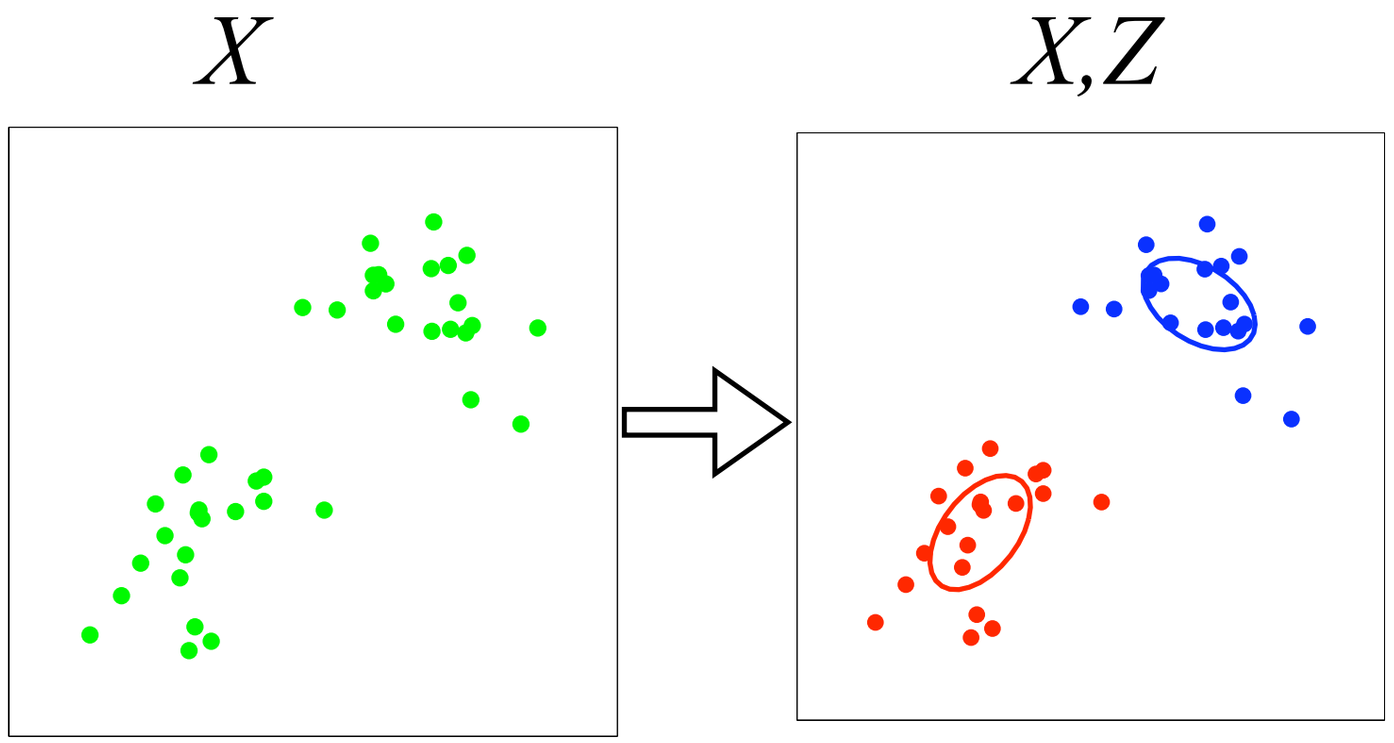
$$\begin{cases} w_i & \text{probability of drawing each component (positives and sum to 1)} \\ \mu_i \in \mathbb{R}^d & \text{mean of the } i\text{-th Gaussian} \\ \Sigma_i \in \mathbb{R}^{d \times d} & \text{covariance matrix of the } i\text{-th Gaussian (p.s.d. matrix)} \end{cases}$$

for each i$\in \{1, \ldots, K\}$

- **Graphical model:**



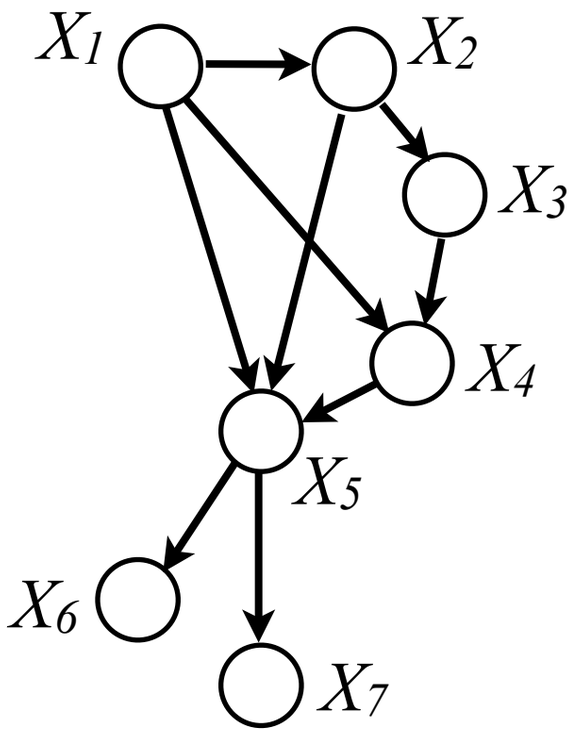
square: categorical variable
circle: continuous variable



Learning the parameters
when all variables are observed

- If we know the model (graph structure) and we have data D_n where all the variables are observed
- Then, learning the model parameters using maximum likelihood is easy!
- We can learn the parameters of each factor (the conditional probabilities in the graph) independently (assuming their parameters are independent...).

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



$$P(X_1, \ldots, X_7) = P(X_1)P(X_2|X_1)P(X_3|X_2) P(X_4|X3, X_1)P(X_5|X_1, X_2, X_4) P(X_6|X_5)P(X_7|X_5)$$

Ex: We can learn the parameters for $P(X_4|X_3,X_1)$ by solving

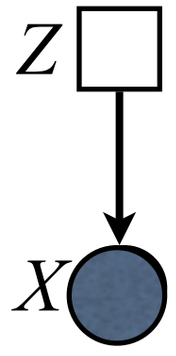
$$\theta^*_{4|3,1} = \arg\max_{\theta_{4|3,1}} \sum_{t=1}^{n} \log P(X_4 = x_4^{(t)}|X_3 = x_3^{(t)}, X_1 = x_1^{(t)})$$

Learning the parameters
when all variables are observed

- Ex: Gaussian mixture model

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

- **Graphical model:**



square: categorical variable
circle: continuous variable

Maximum likelihood:

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

$$= \arg\max_\theta \sum_{t=1}^{n} \log p(Z = z^{(t)}) + \log p(X = x^{(t)} | Z = z^{(t)})$$

$$= \arg\max_{\{w,\mu,\Sigma\}} \sum_{t=1}^{n} \log w_{z^{(t)}} + \log \mathcal{N}(x^{(t)} | \mu_{z^{(t)}}, \Sigma_{z^{(t)}})$$

$$\text{such that } w_k \geq 0 \text{ et } \sum_{k=1}^{K} w_k = 1$$

Learning the parameters
when all variables are observed

- Ex: Gaussian mixture model

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

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

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 of the maximum likelihood using only examples generated by the *k*-th Gaussian (i.e. examples such that z^(t) =*k)*.

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

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

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

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

Wait... we know this algorithm!

⇨ Bayes classifier with Gaussians

Learning the parameters with latent variables (non observed)

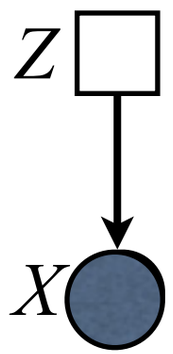
- Usually considerably more difficult:
  we need to marginalize the latent variables
  (or learn/infer their distribution)
- From the training data

$$D_n = \{x^{(1)}, \ldots, x^{(n)}\}, x^{(t)} \in \mathbb{R}^d.$$

- Maximum likelihood principle:

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

- **Graphical model**



Ex: Gaussian mixture (we have an explicit epxresion for *p(X)*, but no analytical solution):

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

- We can directly use gradient descent/ascent on this objective
(but how do we enforce the constraints on the w_i?)
- Or we can use the E.M. (Expectation Maximization) algorithm

Expectation Maximization

[Dempster, Laird and Rubin, 1977]

- Let Z be the latent variable(s)
- The idea behind EM is simple:
  If we know the value for the variables Z, learning would be easy (maximum likelihood).
  So we iterate:
  - Infer the distribution of Z for each example X given our current estimate of the model's parameters
  - Suppose the value of Z are observed (using the distribution we guessed in the last step)
    and find the parameters that maximize the likelihood.
- The E.M. algorithm update the parameters using the formula:

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

Expectation of the log likelihood of $x^{(t)}, z^{(t)}$
According to the current estimate of $z^{(t)} \mid x^{(t)}$.

MAXIMIZATION

Expectation Maximization

- Ex: Gaussian mixture
  - Infer the distribution of Z for each example X given our current estimtate of the model's parameters

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

  - Compute the «responsibilities» $\mathbf{R}_{ti} = P(Z=i|X=x^{(t)})$ using the current estimate of the parameters.

Extreme case for intuition: if we set $w_i = 1/K$ and $\Sigma_i = \varepsilon\, I$ with $\varepsilon \to 0$,
then $\mathbf{R}_t = P(Z|X=x^{(t)}) \to \text{onehot}(\arg\min(\|x^{(t)}-\mu_1\|^2,\ldots,\|x^{(t)}-\mu_K\|^2))$

- We proceed as if *(X, Z)* were fully observed, following this distribution:
  i.e. as if, for each example $x^{(t)}$, we observed a proportion $\mathbf{R}_{ti}$ of examples of the form $(X=x^{(t)}, Z=i)$.
- Find the parameters maximizing the likelihood.

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

$$
\{\mu,\Sigma\}^{\text{new}} = \arg\max_{\theta} \sum_{t=1}^{n}\sum_{i=1}^{K} \underbrace{P(Z=i|X=x^{(t)};\theta\,\text{old}\ )}_{\mathbf{R}_{ti}} \underbrace{\log p(X=x^{(t)},Z=i;\theta)}_{\log w_i + \log\mathcal{N}(x|\mu_i,\Sigma_i)}
$$

Expectation Maximization

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

This is simply a weighted version of the fully observed case we saw earlier.

SOLUTION:

$$\begin{aligned}
n_i &= \sum_{t=1}^{n} \mathbf{R}_{ti} \\
w_i^* &= \frac{n_i}{\sum_{k=1}^{K} n_k} \\
\mu_i^* &= \frac{1}{n_i} \sum_{t=1}^{n} \mathbf{R}_{ti} x^{(t)} \\
\Sigma_i^* &= \frac{1}{n_i} \sum_{t=1}^{n} \mathbf{R}_{ti} (x^{(t)} - \mu_k^*)(x^{(t)} - \mu_k^*)^T
\end{aligned}$$

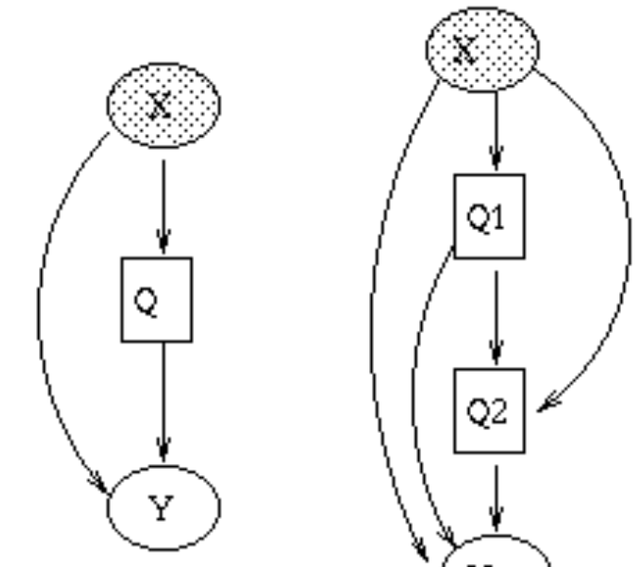
- **Extreme case for intuition: if we set w_i=1/K et $\Sigma_i = \varepsilon I$ avec $\varepsilon \to 0$ then** $\mathbf{R}_t = P(Z|X = x^{(t)}) \to \text{onehot}(\arg\min(\|x^{(t)} - \mu_1\|^2, \ldots, \|x^{(t)} - \mu_K\|^2))$

What is this algorithm???

Learning directed graphical models
Summary

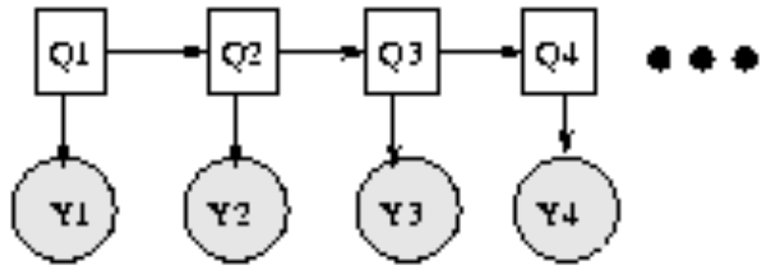
- We saw how we can learn the model parameters
- Easy if all variables are observed
- When there are latent variables we need to efficiently
  - marginalize them (i.e. ignore them by summing over all possible values they can take)
  - or infer their distribution and compute the expectation w.r.t. this distribution (E.M.)
  - or we can use sampling techniques *(variational methods, Monte-Carlo, ...)*.
- There are techniques that try to learn the graph structure (the edges between the nodes) from data -> active research area.

A lot of learning algorithms can be seen as probabilistic graphical models

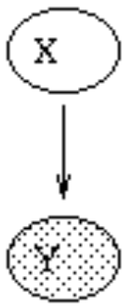


Mixture of Experts

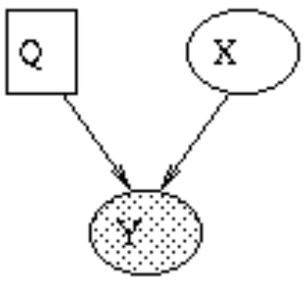
Hierarchical Mixture of Experts



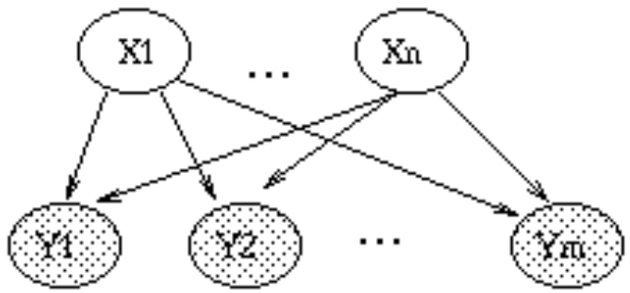
Hidden Markov Model



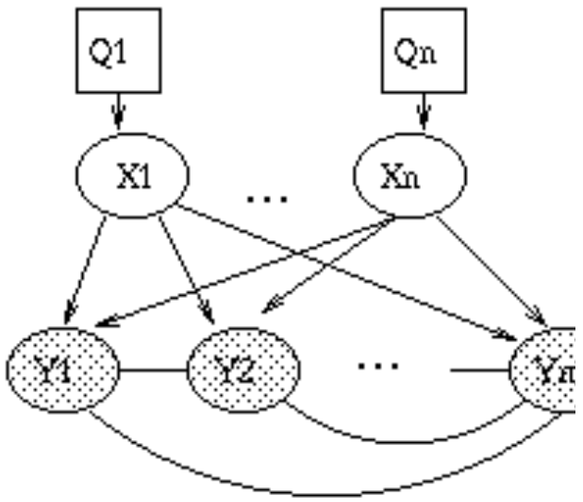
Factor Analysis/PCA



Mixture of FAs



Factor analysis



Independent Factor Analysis