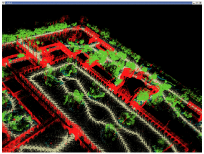
Introduction to Graphical models

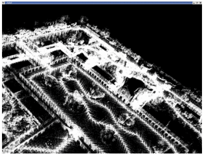
Guillaume Obozinski

Ecole des Ponts - ParisTech

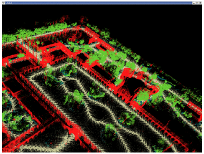
INIT/AERFAI Summer school on Machine Learning
Benicàssim, June 26th 2017

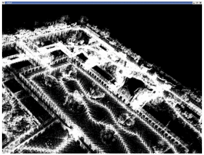
Structured problems in HD

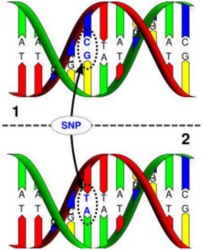




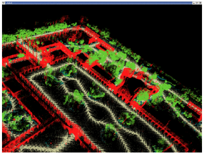
Structured problems in HD

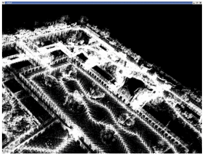


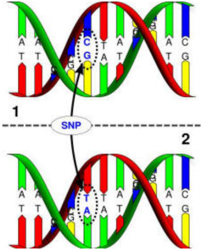


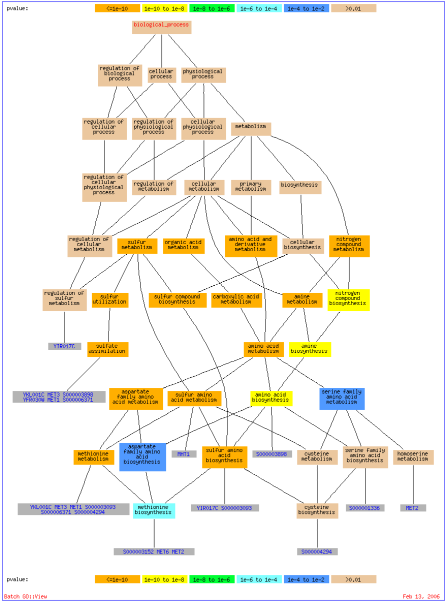


Structured problems in HD









Sequence modelling

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

Sequence modelling

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

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

Sequence modelling

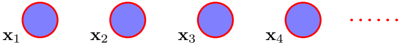
How to model the distribution of DNA sequences of length k?

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

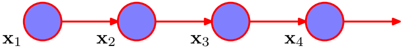
Sequence modelling

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

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



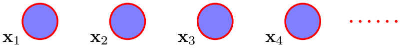
First order Markov chain:



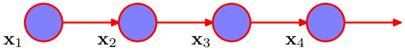
Sequence modelling

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

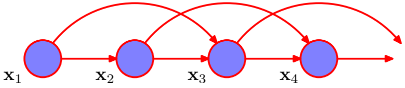
- Naive model $\rightarrow 4^n - 1$ parameters
- Indépendant model $\rightarrow 3n$ parameters



First order Markov chain:



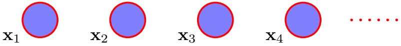
Second order Markov chain:



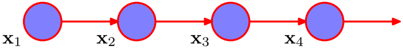
Sequence modelling

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

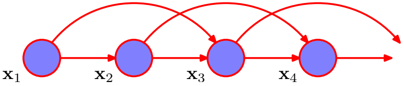
- Naive model $\rightarrow 4^n - 1$ parameters
- Indépendant model $\rightarrow 3n$ parameters



First order Markov chain:



Second order Markov chain:

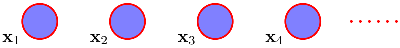


Number of parameters $\mathcal{O}(n)$ for chains of length n.

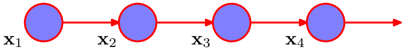
Sequence modelling

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

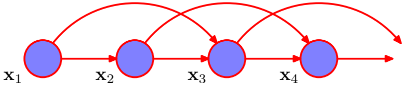
- Naive model $\rightarrow 4^n - 1$ parameters
- Indépendant model $\rightarrow 3n$ parameters



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

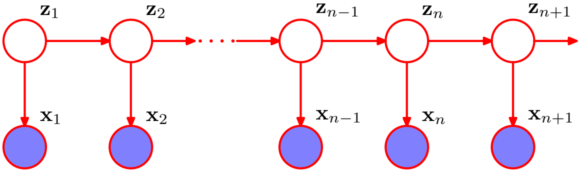
Models for speech processing

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

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

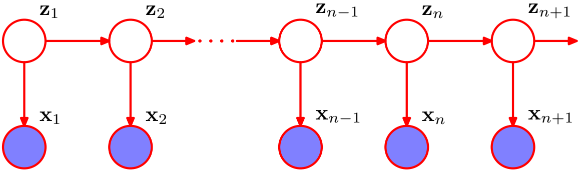
Hidden Markov Model: HMM



Models for speech processing

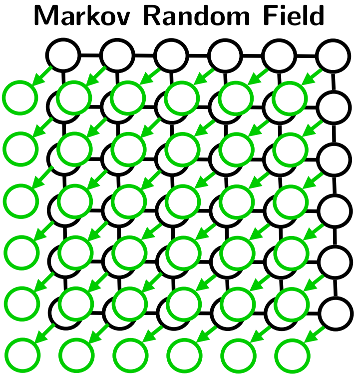
- Speech modelled by a sequence of unobserved phonemes
- For each phoneme a random sound is produced following a distribution which characterizes the phoneme

Hidden Markov Model: HMM



$\rightarrow$ **Latent** variable models

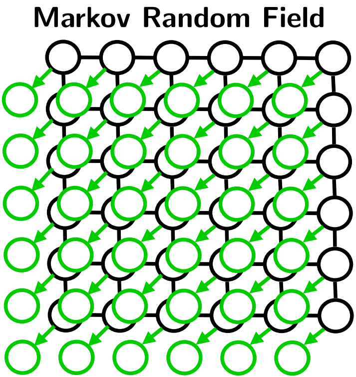
Modelling image structures



Original image

Segmentation

Modelling image structures



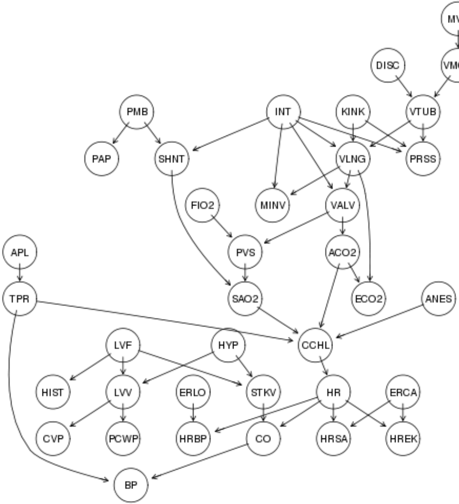
Original image

Segmentation

→ *oriented graphical model* vs *non oriented*

Anaesthesia alarm (Beinlich et al., 1989)

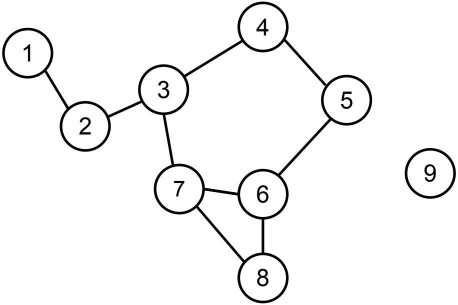
"The ALARM Monitoring system"



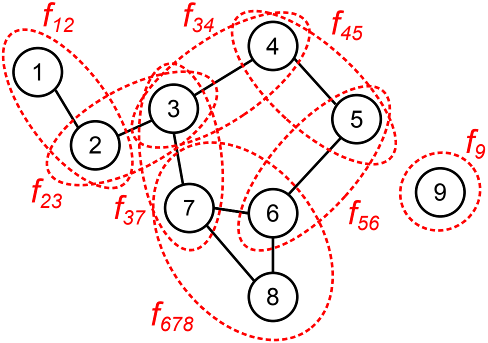
http://www.bnlearn.com/documentation/networks/

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

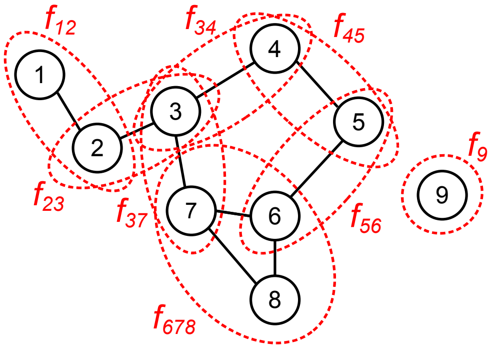
Probabilistic model



Probabilistic model



Probabilistic model



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

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

Poll...

... about some relevant concepts.

Poll...

... about some relevant concepts.

- Markov Chain

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)
- Bayesian inference

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)
- Bayesian inference
- Kullback-Leibler divergence

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)
- Bayesian inference
- Kullback-Leibler divergence
- Expectation maximisation algorithm

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)
- Bayesian inference
- Kullback-Leibler divergence
- Expectation maximisation algorithm
- Sum-product algorithm

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)
- Bayesian inference
- Kullback-Leibler divergence
- Expectation maximisation algorithm
- Sum-product algorithm
- MCMC sampling

Poll...

... about some relevant concepts.

- Markov Chain
- Density of the multivariate Gaussian distribution
- Maximum likelihood estimator
- Logistic regression
- Entropy
- Exponential families of distributions ($\neq$ the exponential distribution)
- Bayesian inference
- Kullback-Leibler divergence
- Expectation maximisation algorithm
- Sum-product algorithm
- MCMC sampling

Outline

1. Preliminary concepts

2. Directed graphical models

3. Markov random field

4. Operations on graphical models

Probability distributions

Joint probability distribution of r.v. $(X_1, \ldots, X_p)$: $p(x_1, x_2, x_3, \ldots, x_n)$.
We assume either that

- X_i takes values in $\{1, \ldots, K\}$ and

$$p(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

- or that $(X_1, \ldots, X_n)$ admits a density in $\mathbb{R}^n$

Marginalization

$$p(x_1) = \sum_{x_2} p(x_1, x_2)$$

Factorization

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_n|x_1, \ldots, x_{n-1})$$

Entropy and Kullback-Leibler divergence

Entropie

$$H(p) = -\sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

$\rightarrow$ Expectation of the negative log-likelihood

Entropy and Kullback-Leibler divergence

Entropie

$$H(p) = -\sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

$\rightarrow$ Expectation of the negative log-likelihood

Kullback-Leibler divergence

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p\Big[\log \frac{p(X)}{q(X)}\Big]$$

$\rightarrow$ Expectation of the log-likelihood ratio

Entropy and Kullback-Leibler divergence

Entropie

$$H(p) = -\sum_x p(x)\log p(x) = \mathbb{E}[-\log p(X)]$$

$\rightarrow$ Expectation of the negative log-likelihood

Kullback-Leibler divergence

$$KL(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \mathbb{E}_p\Big[\log\frac{p(X)}{q(X)}\Big]$$

$\rightarrow$ Expectation of the log-likelihood ratio

$\rightarrow$ Property: $KL(p\|q) \geq 0$

Independence concepts

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

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

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

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:

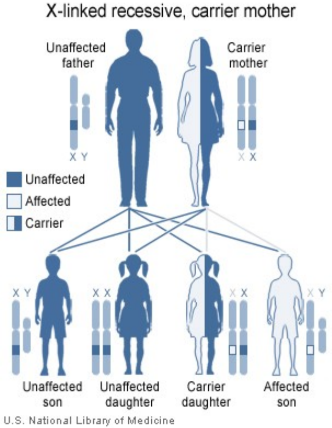
$\forall x, y, z,$

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

Conditional Independence exemple

Example of
"X-linked recessive inheritance":

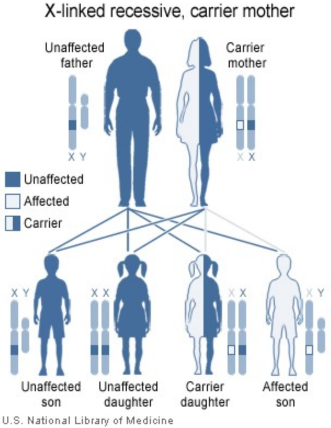
Transmission of the gene
responsible for hemophilia



Conditional Independence exemple

Example of
"X-linked recessive inheritance":

Transmission of the gene
responsible for hemophilia



Risk for sons from an unaffected father:

- dependance between the situation of the two brothers.
- conditionally independent given that the mother is a carrier of the gene or not.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \dots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \dots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.
So $\boldsymbol{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

Indicator variable coding for multinomial variables

Let C a r.v. taking values in $\{1, \ldots, K\}$, with

$$\mathbb{P}(C = k) = \pi_k.$$

We will code C with a r.v. $Y = (Y_1, \ldots, Y_K)^\top$ with

$$\boxed{Y_k = 1_{\{C=k\}}}$$

For example if $K = 5$ and $c = 4$ then $\boldsymbol{y} = (0, 0, 0, 1, 0)^\top$.
So $\boldsymbol{y} \in \{0, 1\}^K$ with $\sum_{k=1}^K y_k = 1$.

$$\mathbb{P}(C = k) = \mathbb{P}(Y_k = 1) \quad \text{and} \quad \mathbb{P}(Y = y) = \prod_{k=1}^K \pi_k^{y_k}.$$

Bernoulli, Binomial, Multinomial

$Y \sim \text{Ber}(\pi)$	$(Y_1, \ldots, Y_K) \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$
$p(y) = \pi^y (1-\pi)^{1-y}$	$p(\mathbf{y}) = \pi_1^{y_1} \ldots \pi_K^{y_K}$
$N_1 \sim \text{Bin}(n, \pi)$	$(N_1, \ldots, N_K) \sim \mathcal{M}(n, \pi_1, \ldots, \pi_K)$
$p(n_1) = \binom{n}{n_1} \pi^{n_1} (1-\pi)^{n-n_1}$	$p(\mathbf{n}) = \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} \pi_1^{n_1} \ldots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!\,i!} \qquad \text{and} \qquad \begin{pmatrix} & n & \\ n_1 & \ldots & n_K \end{pmatrix} = \frac{n!}{n_1! \ldots n_K!}$$

Gaussian model

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

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

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

Gaussian model

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

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

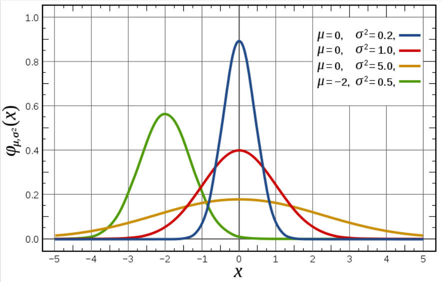
$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

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

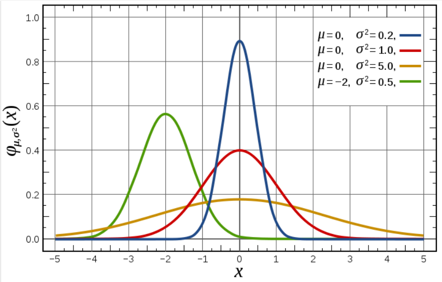
X takes values in $\mathbb{R}^d$. Si $\mathcal{K}_n$ is the set of $n \times n$ positive definite matrices, and $\theta = (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \mathcal{K}_n$.

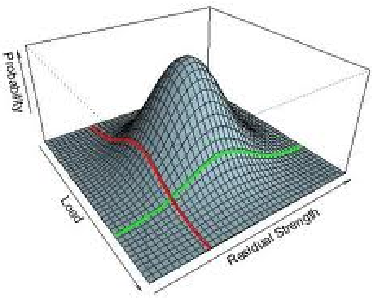
$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Gaussian densities



Gaussian densities





Maximum likelihood principle

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

Maximum likelihood principle

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

Likelihood:

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

Maximum likelihood principle

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

Likelihood:

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

Maximum likelihood estimator:

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

Sir Ronald Fisher
(1890-1962)

Maximum likelihood principle

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

Sir Ronald Fisher
(1890-1962)

Likelihood:

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

Maximum likelihood estimator:

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

Case of i.i.d. data

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

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

Bayesian estimation

Parameters θ are modelled as a random variable.

A priori

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

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.

Outline

1. Preliminary concepts
2. Directed graphical models
3. Markov random field
4. Operations on graphical models

Notations for graphical models

Graphs

$G = (V, E)$ is a graph with vertex set V and edge set E.

Notations for graphical models

Graphs

$G = (V, E)$ is a graph with vertex set V and edge set E.
The graph will be

- either a **directed acyclic graph (DAG)**
  - ≫ then $(i, j) \in E \subset V \times V$ means $i \to j$.

Notations for graphical models

Graphs

$G = (V, E)$ is a graph with vertex set V and edge set E.
The graph will be

- either a **directed acyclic graph (DAG)**
  - $\gg$ then $(i,j) \in E \subset V \times V$ means $i \to j$.
- or a an undirected graph
  - $\gg$ then $\{i,j\} \in E$ means i and j are adjacent.

Notations for graphical models

Graphs

$G = (V, E)$ is a graph with vertex set V and edge set E.
The graph will be

- either a **directed acyclic graph (DAG)**
  - ≫ then $(i, j) \in E \subset V \times V \quad$ means $\quad i \to j$.
- or a an undirected graph
  - ≫ then $\{i, j\} \in E \quad$ means $\quad i$ and j are adjacent.

Variables of the graphical model

- To each node $i \in V$, we associate a graphical variable X_i.

Notations for graphical models

Graphs

$G = (V, E)$ is a graph with vertex set V and edge set E.
The graph will be

- either a **directed acyclic graph (DAG)**
  - ≫ then $(i, j) \in E \subset V \times V$ means $i \to j$.
- or a an undirected graph
  - ≫ then $\{i, j\} \in E$ means i and j are adjacent.

Variables of the graphical model

- To each node $i \in V$, we associate a graphical variable X_i.
- Observations/values of X_i are denoted x_i.

Notations for graphical models

Graphs

$G = (V, E)$ is a graph with vertex set V and edge set E.
The graph will be

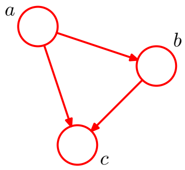
- either a **directed acyclic graph (DAG)**
  - ≫ then $(i,j) \in E \subset V \times V$ means $i \to j$.
- or a an undirected graph
  - ≫ then $\{i,j\} \in E$ means i and j are adjacent.

Variables of the graphical model

- To each node $i \in V$, we associate a graphical variable X_i.
- Observations/values of X_i are denoted x_i.
- If $A \subset V$ is a set of nodes we will write $X_A = (X_i)_{i \in A}$ et $x_A = (x_i)_{i \in A}$.

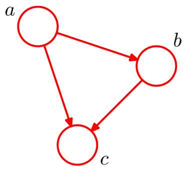
Directed graphical model or Bayesian network

$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$



Directed graphical model or Bayesian network

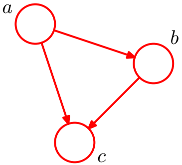
$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$



$$p(x_1, x_2) = p(x_1)p(x_2)$$

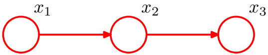
Directed graphical model or Bayesian network

$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$



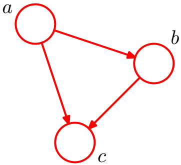
$$p(x_1, x_2) = p(x_1)p(x_2)$$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$$



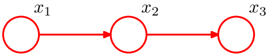
Directed graphical model or Bayesian network

$$p(a, b, c) = p(a)\, p(b|a)\, p(c|b, a)$$

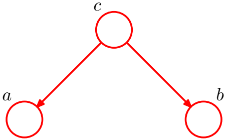


$$p(x_1, x_2) = p(x_1)p(x_2)$$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$$



$$a \perp\!\!\!\perp b \mid c$$

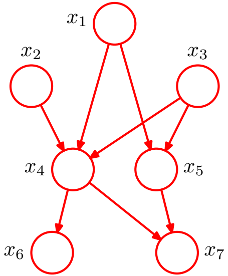


Directed graphical model or Bayesian network

Factorization according to a directed graph

Definition: a distribution factorizes according to a directed graph

$$\prod_{j=1}^{p} p(x_j | x_{\Pi_j})$$

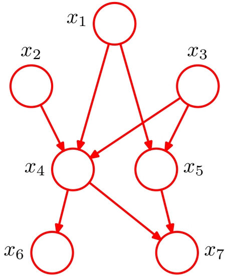


Directed graphical model or Bayesian network

Factorization according to a directed graph

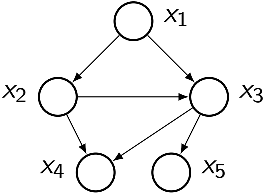
Definition: a distribution factorizes according to a directed graph

$$\prod_{j=1}^{p} p(x_j | x_{\Pi_j})$$



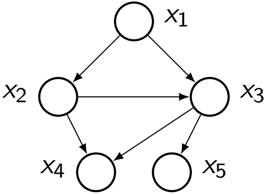
$$p(x_1) \prod_{j=2}^{M} p(x_j | x_{j-1})$$

How to parameterize an Oriented graphical model?



$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \theta_1)\, p(x_2|x_1; \theta_2)\, p(x_3|x_2, x_1; \theta_3)\, p(x_4|x_3, x_2; \theta_4)\, p(x_5|x_3; \theta_5)$$

How to parameterize an Oriented graphical model?

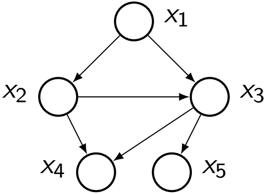


Conditional Probability tables

- $x_1 \in \{0, 1\}$
- $x_2 \in \{0, 1, 2\}$
- $x_3 \in \{0, 1, 2\}$

$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \theta_1)\, p(x_2 | x_1; \theta_2)\, p(x_3 | x_2, x_1; \theta_3)\, p(x_4 | x_3, x_2; \theta_4)\, p(x_5 | x_3; \theta_5)$$

How to parameterize an Oriented graphical model?



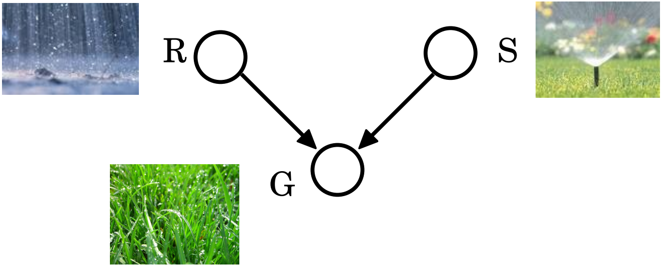
Conditional Probability tables

- $x_1 \in \{0, 1\}$
- $x_2 \in \{0, 1, 2\}$
- $x_3 \in \{0, 1, 2\}$

		$p(x_3 = k)$		
x_1	x_2	0	1	2
0	0	1	0	0
0	1	1	0	0
0	2	0.1	0	0.9
1	0	1	0	0
1	1	0.5	0.5	0
1	2	0.2	0.3	0.5

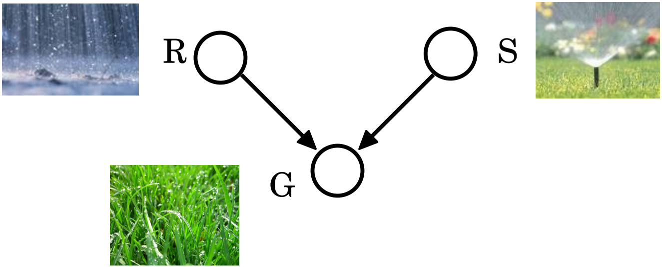
$$p(\mathbf{x}; \boldsymbol{\theta}) = p(x_1; \theta_1)\, p(x_2|x_1; \theta_2)\, p(x_3|x_2, x_1; \theta_3)\, p(x_4|x_3, x_2; \theta_4)\, p(x_5|x_3; \theta_5)$$

The Sprinkler



- $R = 1$: it has rained
- $S = 1$: the sprinkler worked
- $G = 1$: the grass is wet

The Sprinkler



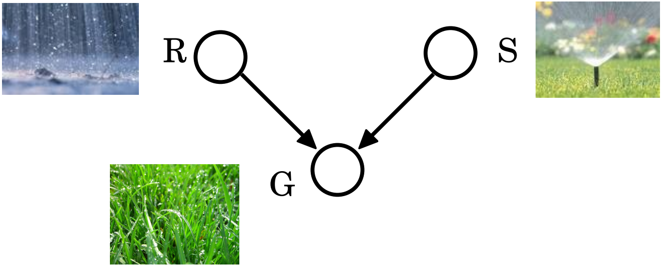
- $R = 1$: it has rained
- $S = 1$: the sprinkler worked
- $G = 1$: the grass is wet

$$P(S = 1) = 0.5$$

$$P(R = 1) = 0.2$$

$P(G = 1 \mid S, R)$	R=0	R=1
S=0	0.01	0.8
S=1	0.8	0.95

The Sprinkler



$$P(S=1)=0.5$$

$$P(R=1)=0.2$$

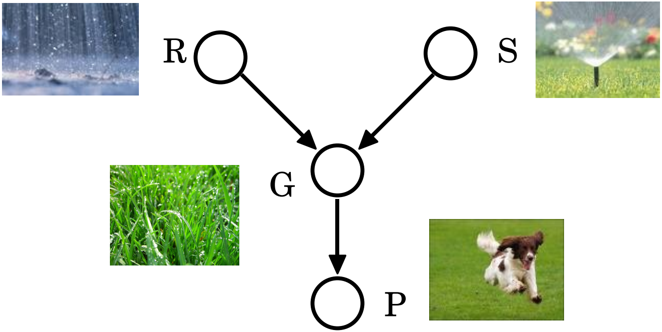
- $R=1$: it has rained
- $S=1$: the sprinkler worked
- $G=1$: the grass is wet

$P(G=1\|S,R)$	R=0	R=1
S=0	0.01	0.8
S=1	0.8	0.95

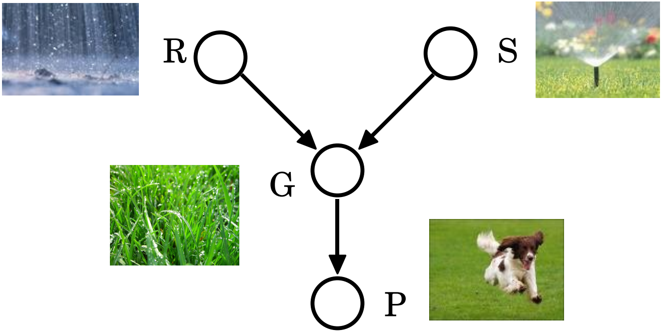
- Given that we observe that the grass is wet, are R and S independent?

The Sprinkler II

The Sprinkler II



The Sprinkler II



- $R = 1$: it has rained
- $S = 1$: the sprinkler worked
- $G = 1$: the grass is wet
- $P = 2$: the paws of the dog are wet

$P(S = 1) = 0.5 \quad P(R = 1) = 0.2$

$P(G = 1 \mid S, R)$	R=0	R=1
S=0	0.01	0.8
S=1	0.8	0.95

$P(P = 1 \mid G)$	G=0	G=1
	0.2	0.7

Factorization and Independence

- A factorization imposes independence statements

$$\forall x,\ p(x) = \prod_{j=1}^{p} p(x_j | x_{\Pi_j}) \quad \Leftrightarrow \quad \forall j,\ X_j \perp\!\!\!\perp X_{\{1,\dots,j-1\}\setminus\Pi_j} \mid X_{\Pi_j}$$

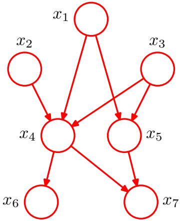
Factorization and Independence

- A factorization imposes independence statements

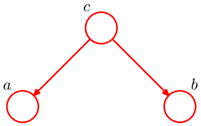
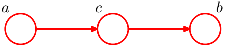
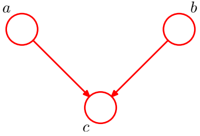
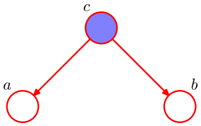
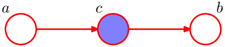
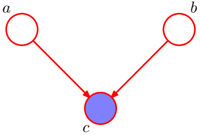
$$\forall x,\ p(x) = \prod_{j=1}^{p} p(x_j | x_{\Pi_j}) \quad \Leftrightarrow \quad \forall j,\ X_j \perp\!\!\!\perp X_{\{1,\ldots,j-1\}\setminus\Pi_j} \mid X_{\Pi_j}$$

- Is it possible to read from the graph the (conditional) independence statements that hold given the factorization.

$$X_5 \overset{?}{\perp\!\!\!\perp} X_2 \mid X_4$$

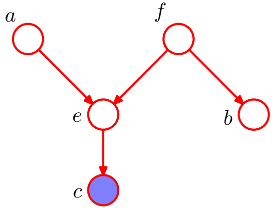


Blocking nodes

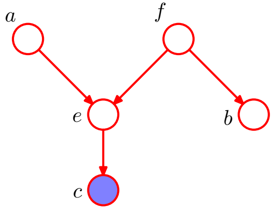
diverging edges	head-to-tail	converging edges
		
		$\not\leftrightarrow$
$a \not\perp\!\!\!\perp b$	$a \not\perp\!\!\!\perp b$	$a \perp\!\!\!\perp b$
		
$\not\leftrightarrow$	$\not\leftrightarrow$	
$a \perp\!\!\!\perp b \mid c$	$a \perp\!\!\!\perp b \mid c$	$a \not\perp\!\!\!\perp b \mid c$

The configuration with converging edges is called a v-structure

d-separation



d-separation



Theorem

If A, B and C are three disjoint sets of node, the statement $X_A \perp\!\!\!\perp X_B | X_C$ holds if all paths joining A to B go through at least one *blocking node*. A node j is blocking a path

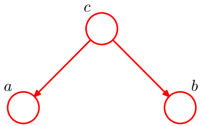
- if the edges of the paths are diverging/following and $j \in C$
- if the edges of the paths are converging (i.e. form a v-structure) and neither j nor any of its descendants is in C

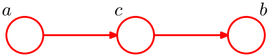
Factorization et Independence II

- Several graphs can induce the same set of conditional independences .

Factorization et Independence II

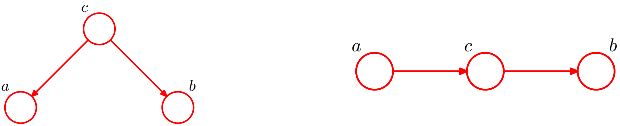
- Several graphs can induce the same set of conditional independences .





Factorization et Independence II

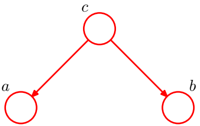
- Several graphs can induce the same set of conditional independences .

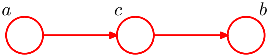


$p(c)p(a|c)p(b|c)$

Factorization et Independence II

- Several graphs can induce the same set of conditional independences .

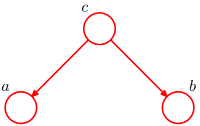


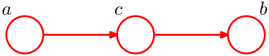


$$p(c)p(a|c)p(b|c) = p(a)p(c|a)p(b|c)$$

Factorization et Independence II

- Several graphs can induce the same set of conditional independences .



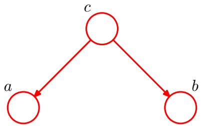


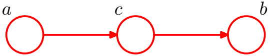
$$p(c)p(a|c)p(b|c) = p(a)p(c|a)p(b|c)$$

- Some combinations of conditional independences cannot be faithfully represented by a graphical model

Factorization et Independence II

- Several graphs can induce the same set of conditional independences .



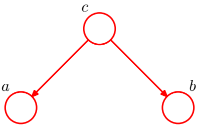


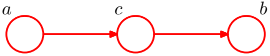
$$p(c)p(a|c)p(b|c) = p(a)p(c|a)p(b|c)$$

- Some combinations of conditional independences cannot be faithfully represented by a graphical model
  - Ex1: $X \sim \text{Ber}\frac{1}{2} \qquad Y \sim \text{Ber}\frac{1}{2} \qquad Z = X \oplus Y.$

Factorization et Independence II

- Several graphs can induce the same set of conditional independences .





$$p(c)p(a|c)p(b|c) = p(a)p(c|a)p(b|c)$$

- Some combinations of conditional independences cannot be faithfully represented by a graphical model
  - Ex1: $X \sim \text{Ber}\frac{1}{2} \qquad Y \sim \text{Ber}\frac{1}{2} \qquad Z = X \oplus Y.$
  - Ex2: $X \perp\!\!\!\perp Y \mid Z = 1$ but $X \not\perp\!\!\!\perp Y \mid Z = 0$

Outline

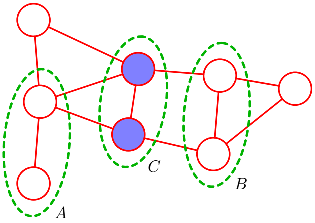
1. Preliminary concepts
2. Directed graphical models
3. Markov random field
4. Operations on graphical models

Markov random field (MRF) or *Oriented graphical model*

Is it possible to associate to each graph a family of distribution so that conditional independence coincides exactly with the notion of separation in the graph?

Global Markov Property

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad \Leftrightarrow C \text{ separates } A \text{ et } B$$



Gibbs distribution

Clique Set of nodes that are all connected to one another.

Gibbs distribution

Clique Set of nodes that are all connected to one another.

Potential function The potential $\psi_C(x_C) \geq 0$ is associated to clique C.

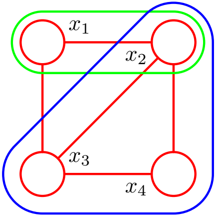
Gibbs distribution

Clique Set of nodes that are all connected to one another.

Potential function The potential $\psi_C(x_C) \geq 0$ is associated to clique C.

Gibbs distribution

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$$



Partition function

$$Z = \sum_x \prod_C \psi_C(x_C)$$

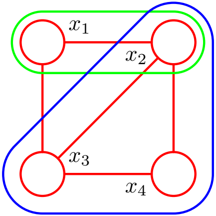
Gibbs distribution

Clique Set of nodes that are all connected to one another.

Potential function The potential $\psi_C(x_C) \geq 0$ is associated to clique C.

Gibbs distribution

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$$



Partition function

$$Z = \sum_x \prod_C \psi_C(x_C)$$

Writing potential in exponential form $\psi_C(x_C) = \exp\{-E(x_C)\}$.
$E(x_C)$ is an *energy*.
This a *Boltzmann distribution*.

Example 1: Ising model

$X = (X_1, \ldots, X_d)$ is a collection of binary variables, whose joint probability distribution is

$$\begin{aligned} p(x_1, \ldots, x_d) &= \frac{1}{Z(\eta)} \exp\Big(\sum_{i \in V} \eta_i x_i + \sum_{\{i,j\} \in E} \eta_{ij} x_i x_j\Big) \\ &= \frac{1}{Z(\eta)} \prod_{i \in V} e^{\eta_i x_i} \prod_{\{i,j\} \in E} e^{\eta_{ij} x_i x_j} \\ &= \frac{1}{Z(\eta)} \prod_{i \in V} \psi_i(x_i) \prod_{\{i,j\} \in E} \psi_i(x_i, x_j) \end{aligned}$$

with $\psi_i(x_i) = e^{\eta_i x_i}$ and $\psi_{ij}(x_i, x_j) = e^{\eta_{ij} x_i x_j}$.

Example 2: Directed graphical model

Consider a distribution p that factorizes according to a directed graph $G = (V, E)$, then

$$
\begin{aligned}
p(x_1, \dots, x_d) &= \prod_{i=1}^{d} p(x_i \mid x_{\pi_i}) \\
&= \prod_{i=1}^{d} \psi_{C_i}(x_{C_i}) \qquad \text{with} \quad C_i = \{i\} \cup \pi_i
\end{aligned}
$$

Consequence: A distribution that factorizes according to a directed model is a Gibbs distribution for the cliques $C_i = \{i\} \cup \pi_i$. As a consequence, it factorizes according to an undirected graph in which C_i are cliques.

Theorem of Hammersley and Clifford (1971)

A distribution p, which is such that $p(x) > 0$ for all x satisfies the *global Markov property* for graph G if and only if it is a Gibbs distribution associated with G.

- Gibbs distribution: $\mathcal{P}_G : \; p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}_G} \psi_C(x_C)$
- Global Markov property:

$$\mathcal{P}_M : \; X_A \perp\!\!\!\perp X_B \mid X_C \quad \text{si} \quad C \text{ separated } A \text{ and } B \text{ in } G$$

Theorem

We have $\quad \mathcal{P}_G \Rightarrow \mathcal{P}_M \quad$ and $\quad$ (HC): if $\forall x, \; p(x) > 0$, then $\mathcal{P}_M \Rightarrow \mathcal{P}_G$

Markov Blanket in an undirected graph

Definition

The Markov Blanket B of a node i is the smallest set of nodes B such that

$$X_i \perp\!\!\!\perp X_R \mid X_B, \qquad \text{with} \quad R = V \backslash (B \cup \{i\})$$

Markov Blanket in an undirected graph

Definition

The Markov Blanket B of a node i is the smallest set of nodes B such that

$$X_i \perp\!\!\!\perp X_R \mid X_B, \qquad \text{with} \quad R = V \backslash (B \cup \{i\})$$

or equivalently such that

$$p(X_i \mid X_{-i}) = p(X_i \mid X_B).$$

Markov Blanket in an undirected graph

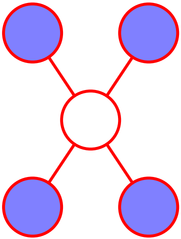
Definition

The Markov Blanket B of a node i is the smallest set of nodes B such that

$$X_i \perp\!\!\!\perp X_R \mid X_B, \qquad \text{with} \quad R = V \backslash (B \cup \{i\})$$

or equivalently such that

$$p(X_i \mid X_{-i}) = p(X_i \mid X_B).$$

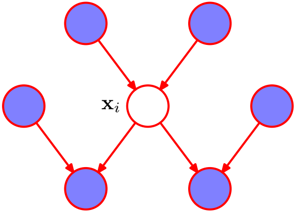


Markov Blanket for a directed graph?

What is the Markov Blanket in a directed graph? By definition: the smallest set C of nodes such that conditionally on X_C, j is independent of all the other nodes in the graph?

Markov Blanket for a directed graph?

What is the Markov Blanket in a directed graph? By definition: the smallest set C of nodes such that conditionally on X_C, j is independent of all the other nodes in the graph?



Moralization

For a given oriented graphical model

- is there an unoriented graphical model which is equivalent?
- is there a smallest unoriented graphical which contains the oriented graphical model?

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C) \quad \text{vs} \quad \prod_{j=1}^{M} p(x_j | x_{\Pi_j})$$

Moralization

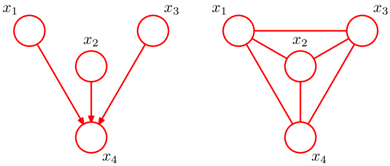
Given a directed graph G, its moralized graph G_M is obtained by

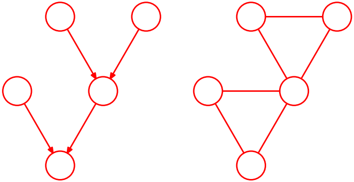
1. For any node i, add undirected edges between all its parents
2. Remove the orientation of all the oriented edges

Moralization

Given a directed graph G, its moralized graph G_M is obtained by

1. For any node i, add undirected edges between all its parents
2. Remove the orientation of all the oriented edges

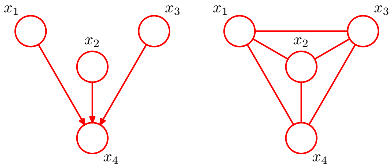


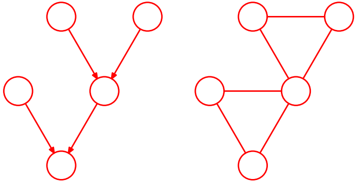


Moralization

Given a directed graph G, its moralized graph G_M is obtained by

1. For any node i, add undirected edges between all its parents
2. Remove the orientation of all the oriented edges





Proposition

If a probability distribution factorizes according to a directed graph G then it factorizes according to the undirected graph G_M.

Directed vs undirected trees

Definition: directed tree

A directed tree is a DAG such that each node has at most one parent

Directed vs undirected trees

Definition: directed tree

A directed tree is a DAG such that each node has at most one parent

Remark: By definition a directed tree has no v-structure.

Directed vs undirected trees

Definition: directed tree

A directed tree is a DAG such that each node has at most one parent

Remark: By definition a directed tree has no v-structure.

Moralizing trees

- What is the moralized graph for a directed tree?
- The corresponding undirected tree!

Proposition (Equivalence between directed and undirected tree)

A distribution factorizes according to a directed tree if and only if it factorizes according to its undirected version.

Directed vs undirected trees

Definition: directed tree

A directed tree is a DAG such that each node has at most one parent

Remark: By definition a directed tree has no v-structure.

Moralizing trees

- What is the moralized graph for a directed tree?
- The corresponding undirected tree!

Proposition (Equivalence between directed and undirected tree)

A distribution factorizes according to a directed tree if and only if it factorizes according to its undirected version.

Corollary

All orientations of the edges of a tree that do not create v-structure are equivalent.

Outline

1 Preliminary concepts

2 Directed graphical models

3 Markov random field

4 Operations on graphical models

Operations on graphical models

Probabilistic inference

Compute a marginal distribution $p(x_i)$ or a *conditional marginal* $p(x_i|x_1 = 3, x_7 = 0)$

Operations on graphical models

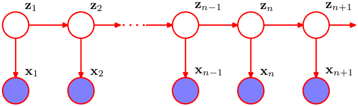
Probabilistic inference

Compute a marginal distribution $p(x_i)$ or a *conditional marginal* $p(x_i|x_1 = 3, x_7 = 0)$

Decoding (aka MAP Inference)

Finding what is the most probable configuration for the set of random variables?

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



Learning/ estimation in graphical models

Frequentist learning

The main *frequentist* learning principle for graphical model is the *maximum likelihood principle* of R. Fisher. Let $p(x;\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_C \psi(x_C, \theta_C)$, we would like to find

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

Bayesian learning

Graphical models can also learn using *bayesian inference*.

The "Naive Bayes" model for classification

Data

- Class label: $C \in \{1, \dots, K\}$
- Class indicator vector $Z \in \{0, 1\}^K$
- Features $X_j, \quad j = 1, \dots, D$
  (e.g. word presence)

The "Naive Bayes" model for classification

Data

- Class label: $C \in \{1, \ldots, K\}$
- Class indicator vector $Z \in \{0,1\}^K$
- Features $X_j, \quad j = 1, \ldots, D$
  (e.g. word presence)

Model

$$p(\mathbf{z}) = \prod_k \pi_k^{z_k}$$

Model

Which model for

$$p(x_1, \ldots, x_D | z_k = 1)\ ?$$

The "Naive Bayes" model for classification

Data

- Class label: $C \in \{1, \ldots, K\}$
- Class indicator vector $Z \in \{0, 1\}^K$
- Features $X_j, \quad j = 1, \ldots, D$
  (e.g. word presence)

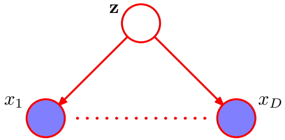
Model

Which model for

$$p(x_1, \ldots, x_D | z_k = 1)\ ?$$

Model

$$p(\boldsymbol{z}) = \prod_k \pi_k^{z_k}$$



"Naive" hypothesis

$$p(x_1, \ldots, x_D | z_k = 1) = \prod_{j=1}^{D} p(x_j \mid z_k = 1; b_{jk}) = \prod_{j=1}^{D} b_{jk}^{x_j} (1 - b_{jk})^{1 - x_j}$$

with $b_{jk} = \mathbb{P}(x_j = 1 \mid z_k = 1)$.

Naive Bayes (continued)

Learning (estimation) with the maximum likelihood principle

$$\hat{\pi} = \underset{\pi:\pi^\top \mathbf{1}=1}{\operatorname{argmax}} \prod_{k,i} \pi_k^{z_k^{(i)}} \qquad \hat{b}_{jk} = \underset{b_{jk}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(x_j^{(i)} | z^{(i)} = k; b_{jk})$$

Prediction:

$$\hat{z} = \operatorname{argmax}_z \frac{\prod_{j=1}^{D} p(x_j|z)p(z)}{\sum_{z'} \prod_{j=1}^{D} p(x_j|z')p(z')}$$

Naive Bayes (continued)

Learning (estimation) with the maximum likelihood principle

$$\hat{\pi} = \underset{\pi:\pi^\top \mathbf{1}=1}{\text{argmax}} \prod_{k,i} \pi_k^{z_k^{(i)}} \qquad \hat{b}_{jk} = \underset{b_{jk}}{\text{argmax}} \sum_{i=1}^{n} \log p(x_j^{(i)}|z^{(i)} = k; b_{jk})$$

Prediction:

$$\hat{z} = \text{argmax}_z \frac{\prod_{j=1}^{D} p(x_j|z)p(z)}{\sum_{z'} \prod_{j=1}^{D} p(x_j|z')p(z')}$$

Properties

- Ignores the correlation between features
- Prediction requires only to use Bayes rule
- The model can be learnt in parallel
- Complexity in $\mathcal{O}(nD)$