Graphical Models

TTT4185 Machine Learning for Signal Processing

Giampiero Salvi

Department of Electronic Systems
NTNU

HT2021

Course plan (So far)

- Speech Perception and Production
- Speech Analysis and Feature Extraction (Ex. 1)
- Intro to Machine Learning/Probability and Information Theory
- Linear Models for Regression and Classification
- Kernel Methods and Support Vector Machines (Ex. 2)
- Deep Neural Networks (Ex. 3)

Course plan (Rest of the lectures)

• Graphical Models (Bayesian Networks)	2h
• Unsupervised Learning (k-means and Mixture Models)	2h
• Sequences and Hidden Markov Models	4h
• Dimensionality Reduction	2h
• Summary	1h

Guest lecture!
Pablo Ortiz, Telenor Research 2020-11-24

Computer Exercises: Orals

- You will be able to book a time from week 46
- The focus will be on understanding rather than implementation
- More information will be given by the TAs

Graphical Models: Motivation

So far (supervised learning)

- $\mathbf{x}_i$ input, t_i output
- goal: estimate $f(\mathbf{x})$ that approximates relationship between $\mathbf{x}$ and t

We would like to consider probabilistic models more in general

- given a set of random variables $\mathbf{x}_1, \dots, \mathbf{x}_K$
- describe the joint probability distribution $p(\mathbf{x}_1, \dots, \mathbf{x}_K)$

Graphical Models:

- simple way to visualize structure in probabilistic models
- insights into the properties of the model (conditional independence)
- complex computations expressed in terms of graphical manipulations

Graphical Models

Definition

- Each node corresponds to a random variable
- Each link (edge or arc) represents probabilistic relationship between variables
- Graph: how to decompose the joint distribution into factors

Bayesian Networks:

- Directed graphs
- useful to express causal relationships

Markov Random Fields:

- Undirected graphs

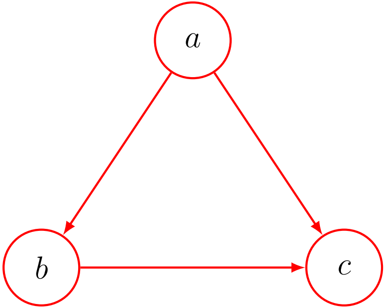
Factor Graphs:

- Mainly used for calculations

Bayesian Network simple example

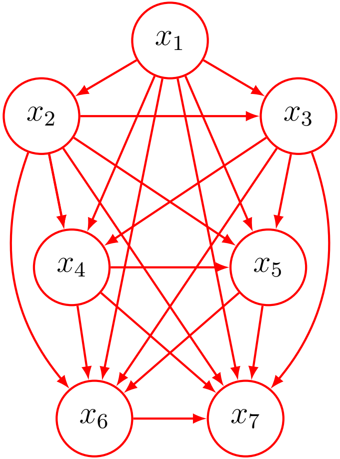
Three variables a, b, and c.

$$
\begin{aligned}
p(a,b,c) &= p(c|a,b)p(a,b) \\
&= p(c|a,b)p(b|a)p(a)
\end{aligned}
$$



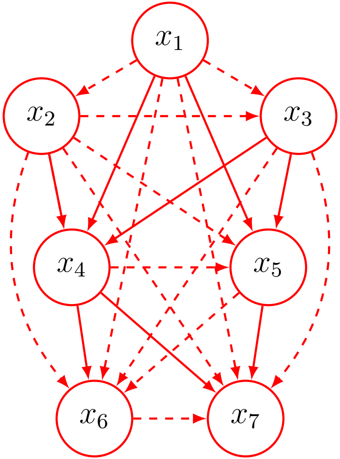
Bayesian Networks (example)

$$
\begin{aligned}
p(x_1, \ldots, x_7) = \\
& p(x_1) \\
& p(x_2|x_1) \\
& p(x_3|x_1, x_2) \\
& p(x_4|x_1, x_2, x_3) \\
& p(x_5|x_1, x_2, x_3, x_4) \\
& p(x_6|x_1, x_2, x_3, x_4, x_5) \\
& p(x_7|x_1, x_2, x_3, x_4, x_5, x_6)
\end{aligned}
$$



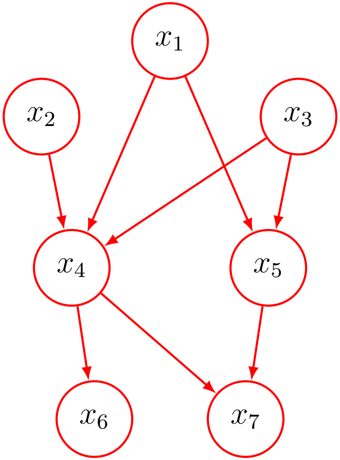
Bayesian Networks (example)

$$
\begin{aligned}
p(x_1, \ldots, x_7) = \\
& p(x_1) \\
& p(x_2 | x_1) \\
& p(x_3 | x_1, x_2) \\
& p(x_4 | x_1, x_2, x_3) \\
& p(x_5 | x_1, x_2, x_3, x_4) \\
& p(x_6 | x_1, x_2, x_3, x_4, x_5) \\
& p(x_7 | x_1, x_2, x_3, x_4, x_5, x_6)
\end{aligned}
$$



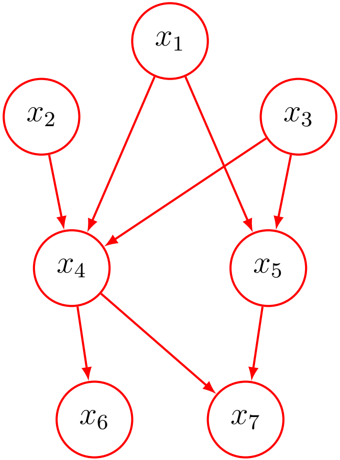
Bayesian Networks (example)

$$
\begin{aligned}
p(x_1, \ldots, x_7) = \\
& p(x_1) \\
& p(x_2) \\
& p(x_3) \\
& p(x_4 | x_1, x_2, x_3) \\
& p(x_5 | x_1, x_3) \\
& p(x_6 | x_4) \\
& p(x_7 | x_4, x_5)
\end{aligned}
$$



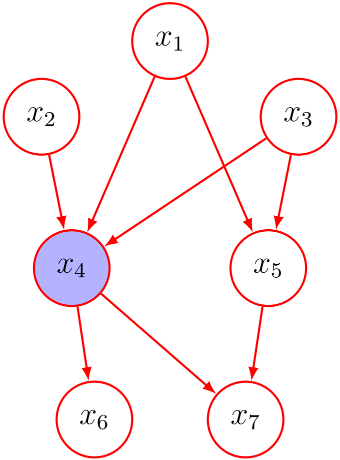
Bayesian Networks (example)

$$p(x_1, \dots, x_7) = \prod_{k=1}^{K} p(x_k | \mathsf{pa}_k)$$



Bayesian Networks (example)

If we observe x_4...



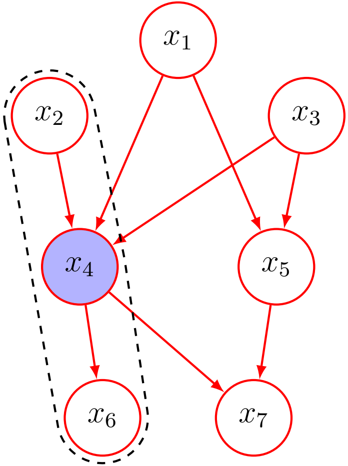
Bayesian Networks (example)

If we observe $x_4 \ldots$

d-separation:

Head-to-tail:

x_2 and x_6 conditionally independent



Bayesian Networks (example)

If we observe $x_4 \ldots$

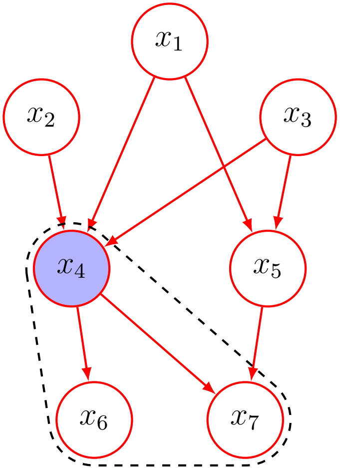
d-separation:

Head-to-tail:

x_2 and x_6 conditionally independent

Tail-to-tail:

x_6 and x_7 conditionally independent



Bayesian Networks (example)

If we observe $x_4 \ldots$

d-separation:

Head-to-tail:

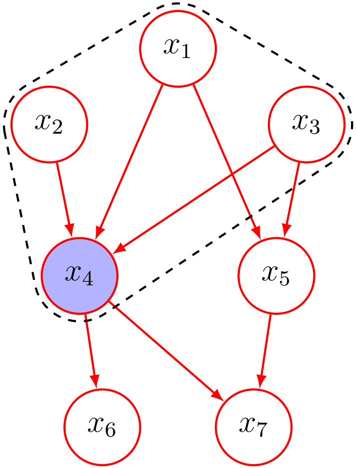
x_2 and x_6 conditionally independent

Tail-to-tail:

x_6 and x_7 conditionally independent

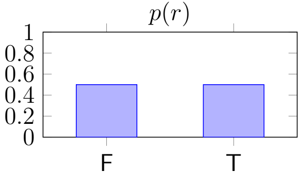
Head-to-head:

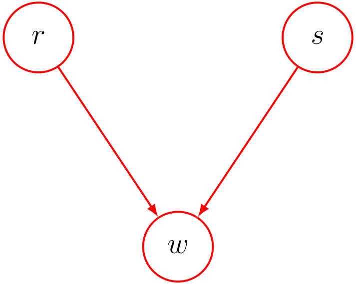
x_1, x_2 and x_3 dependent
(explaining away)

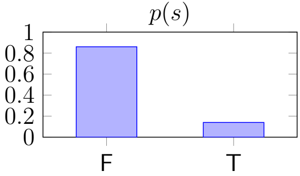


Explaining Away: Intuition

r: it rains, s: sprinkle is on, w: the grass is wet

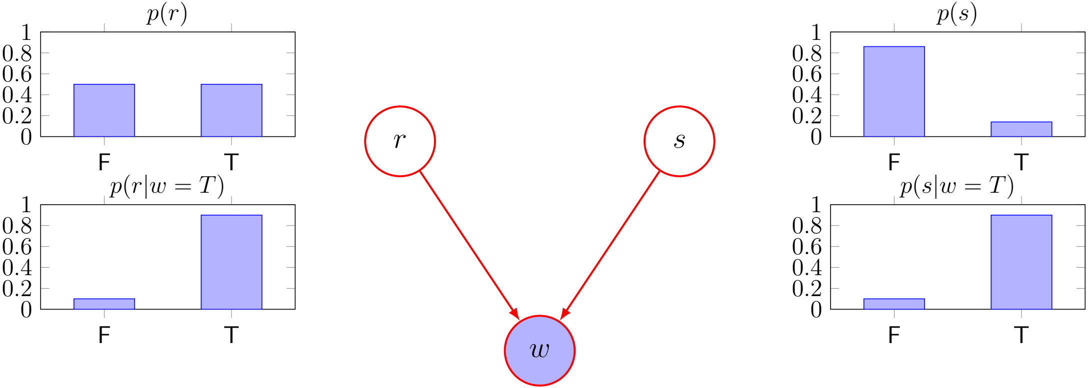






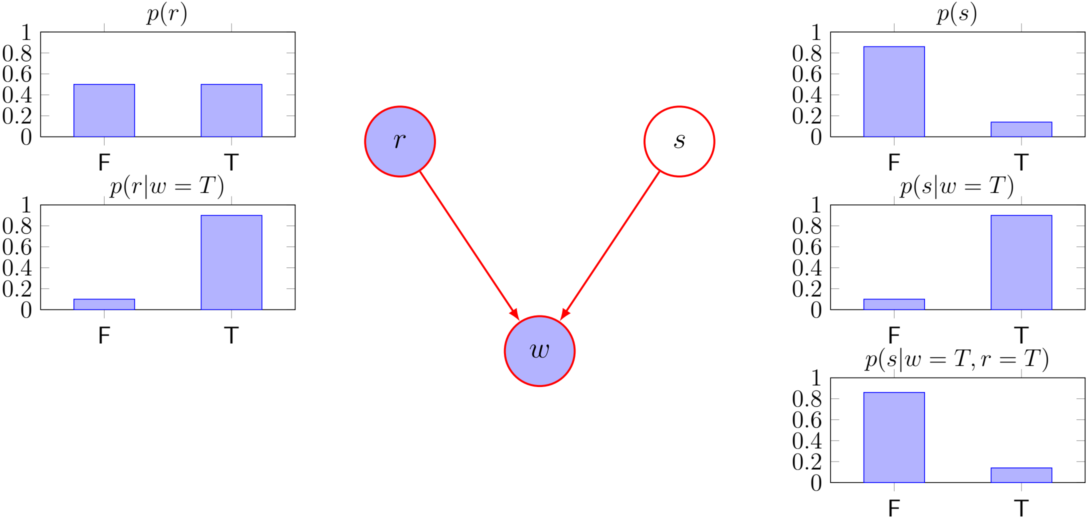
Explaining Away: Intuition

r: it rains, s: sprinkle is on, w: the grass is wet



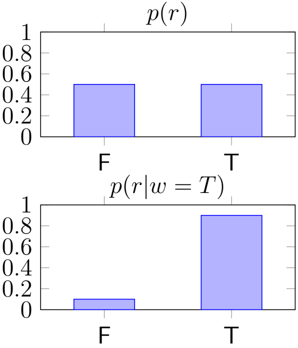
Explaining Away: Intuition

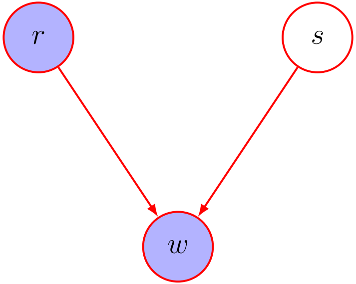
r: it rains, s: sprinkle is on, w: the grass is wet



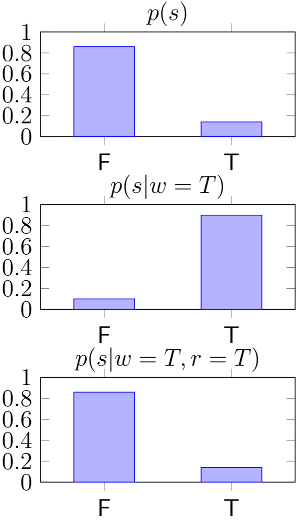
Explaining Away: Intuition

r: it rains, s: sprinkle is on, w: the grass is wet





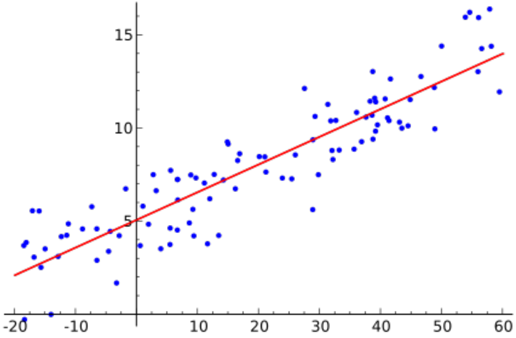
$$p(s|w=T, r=T) \neq$$
$$p(s|w=T)$$



Example: MAP linear regression

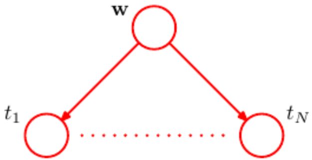
Model:

$$
\begin{aligned}
p(t|\mathbf{w}, \mathbf{x}, \sigma^2) &= \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2) \\
p(\mathbf{w}|\alpha) &= \mathcal{N}(0, \frac{1}{\alpha}\mathbf{I})
\end{aligned}
$$



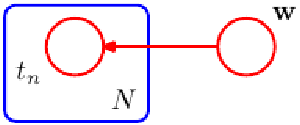
Example: MAP linear regression

$$\mathbf{t} = (t_1, \ldots, t_n)^T$$
$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$$
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$



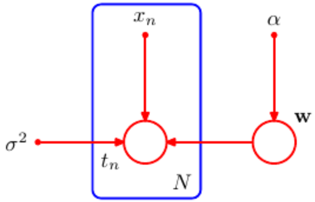
Example: MAP linear regression

$$
\begin{aligned}
\mathbf{t} &= (t_1, \dots, t_n)^T \\
\mathbf{X} &= (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \\
p(\mathbf{t}, \mathbf{w}) &= p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})
\end{aligned}
$$



Example: MAP linear regression

$$
\begin{aligned}
\mathbf{t} &= (t_1, \ldots, t_n)^T \\
\mathbf{X} &= (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \\
p(\mathbf{t}, \mathbf{w}) &= p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}) \\
p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \alpha, \sigma^2) &= p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, \mathbf{x}_n, \sigma^2)
\end{aligned}
$$



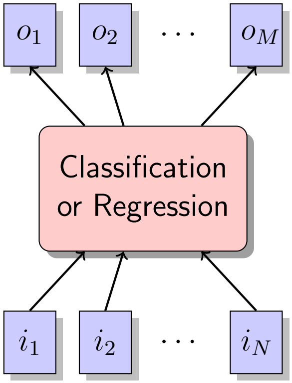
Generative Models: Ancestral Sampling

- consider variables $x_1, \dots, x_k$ with joint $p(x_1, \dots, x_k)$
- order them in such a way that there is no link from any node x_i to any node x_j if $j < i$
- we want a sample $\hat{x}_1, \dots, \hat{x}_k$
- first sample $\hat{x}_1$ from $p(x_1)$
- then for every $n = [2, k]$ sample $\hat{x}_n$ from $p(x_n|\text{pa}_n)$

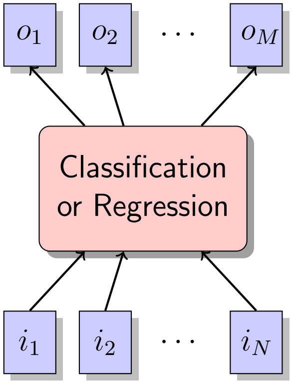
Inference

1. Given a set of observed variables, compute marginal for the other nodes
   - junction tree algorithm: efficient algorithm based on dynamic programming
2. given a set of data points eliminate arcs
   - K2 algorithm

Graphical Models and Latent Variables



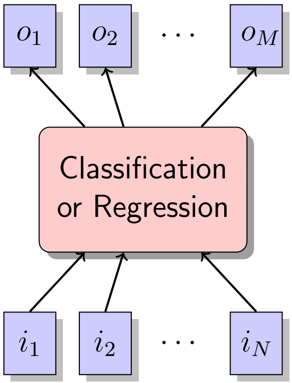
Graphical Models and Latent Variables



conditional distributions

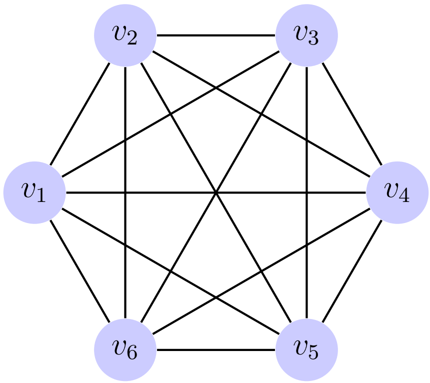
$$P(o_1, \cdots, o_M | i_1, \cdots, i_N)$$
$$P(i_1, \cdots, i_M | o_1, \cdots, o_N)$$

Graphical Models and Latent Variables

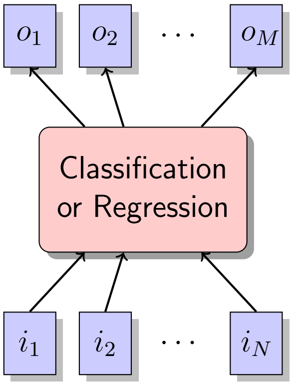


conditional distributions

$$P(o_1, \cdots, o_M | i_1, \cdots, i_N)$$
$$P(i_1, \cdots, i_M | o_1, \cdots, o_N)$$

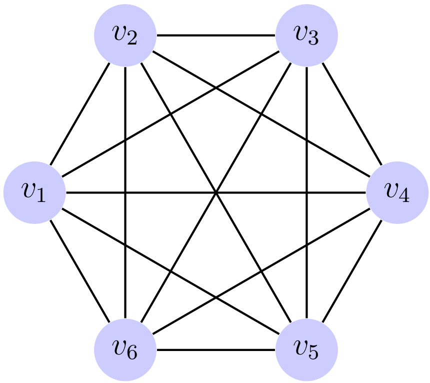


Graphical Models and Latent Variables



conditional distributions

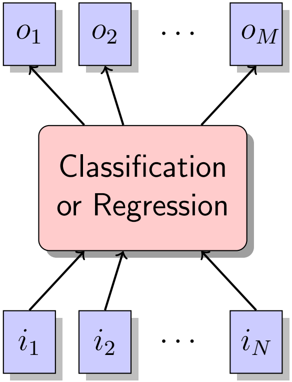
$$P(o_1, \cdots, o_M | i_1, \cdots, i_N)$$
$$P(i_1, \cdots, i_M | o_1, \cdots, o_N)$$



joint distribution

$$P(v_1, v_2, \cdots, v_K)$$

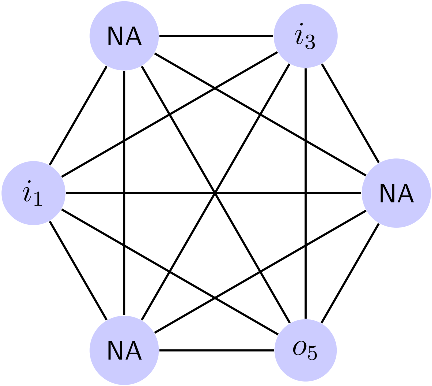
Graphical Models and Latent Variables



conditional distributions

$$P(o_1, \cdots, o_M | i_1, \cdots, i_N)$$

$$P(i_1, \cdots, i_M | o_1, \cdots, o_N)$$

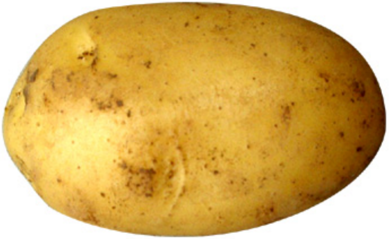


joint distribution

$P(v_1, v_2, \cdots, v_K)$ marginalisation

$$P(o_5 | i_1, i_3)$$

Multi-modality: analogy



Use of Graphical Models

- interpret probabilistic methods
- use directly as ML method (junction tree and K2 algorithms)
- Pros: much more flexible than SVMs and DNNs
- Cons: learning more difficult
    - not feasible for complex (noisy, continuous) perception tasks
    - more suitable for higher cognitive functions (discrete)