Machine Learning

Graphical Models & Conclusion of Part II

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 30, 2017

http://www.carlhenrik.com

So far

- Part I
  - tools
- Part II
  - Models
- Part III
  - Inference

Today

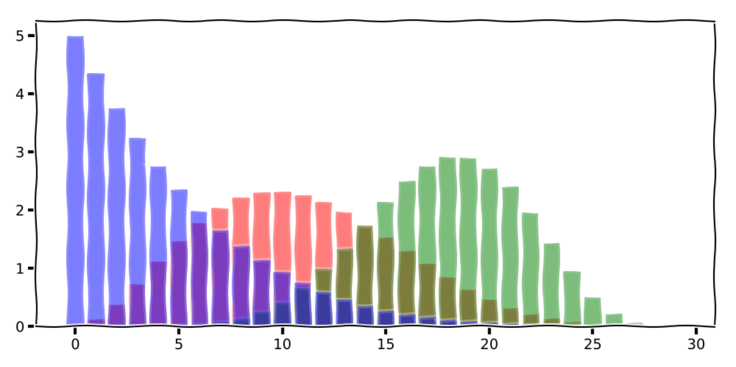
- Recap LDA
- Generalise Part II

Latent Dirichlet Allocation

Text Data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Bag-of-Words



Topic Data

Text collection of words

Images collection of visual words

Genomics collection of sequences

Bioinformatics collection of sequences

Anything that we can describe reliably as a random collection of things fits well into this modelling paradigm

Topic Model

Assumptions

- Documents contains a blend of topics
- Each topic has a characteristic distribution of words
- For simplicity lets assume a finite (known) number of words

Topic Distribution

Topic	p(Topic)
Computer Science	0.3
Pugs	0.5
Bristol City	0.1
C64	0.1

$$p(\text{Topic}) = p(\beta)$$

Word Topic Distribution

Topic	8-bit	Tammy	AI	Hugs
Computer Science	0.6	0.09	0.3	0.01
Pugs	0.1	0.1	0.3	0.5
Bristol City	0.1	0.8	0.05	0.05
C64	0.8	0.01	0.15	0.04

$$p(\text{Word}|\text{Topic}) = p(w_{d,n}|\beta_k)$$

Document Topic Distribution

Document	Computer Science	Pugs	Bristol City	C64
One Team in Bristol	0.1	0.2	0.69	0.01
Coursework report	0.8	0.1	0.07	0.03
Dr. Doobs	0.6	0.1	0.1	0.2
Pug Weekly	0.1	0.8	0.05	0.05

$$p(\theta_d)$$

Topic Model

- To generate the n:th word in the d:th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

Topic Model

- To generate the n:th word in the d:th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

- To generate a topic assignment $z_{d,n}$ I need to know the topic distribution for document d

$$p(z_{d,n}|\theta_d)$$

Topic Model

- To generate the n:th word in the d:th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

- To generate a topic assignment $z_{d,n}$ I need to know the topic distribution for document d

$$p(z_{d,n}|\theta_d)$$

- If we know the topic assignment and the topic-word distribution we can assume the words to be independent

$$p(w_d|\beta, z_d) = \prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n})$$

Topic Model

- We can assume that the topic-document distribution is not dependent between documents

$$p(\theta) = \prod_{d=1}^{D} p(\theta_d)$$

Topic Model

- We can assume that the topic-document distribution is not dependent between documents

$$p(\theta) = \prod_{d=1}^{D} p(\theta_d)$$

- We can assume that topic word distribution is independent

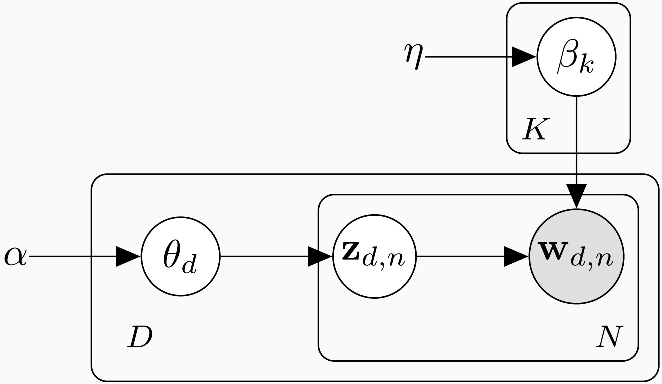
$$p(\beta) = \prod_{k=1}^{K} p(\beta_k)$$

Topic Model

$$p(w, z, \theta, \beta) = \underbrace{\prod_{k=1}^{K} p(\beta_k) \underbrace{\prod_{d=1}^{D} p(\theta_d) \underbrace{\prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n}) p(z_{d,n}|\theta_d)}_{\text{word}}}_{\text{document}}}_{\text{corpus}}$$

- This is the joint distribution of a specific topic model
- The assumptions we have made are called a Latent Dirichlet Allocation [1]
- This specifies a specific probability distribution by our assumptions

Graphical Model



Topic Model: Evidence

$$p(w) = \int p(w, z, \theta, \beta) \mathrm{d}\beta \mathrm{d}\theta dz$$
$$= \int \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n}) p(z_{d,n}|\theta_d) \mathrm{d}\beta \mathrm{d}\theta dz$$

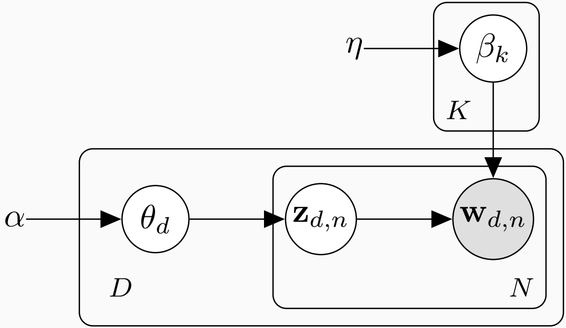
- Distribution over corpus for *any topic model* weighted by our assumptions

Topic Model: Distributions

- We still have to choose the specific form of the distribution
- Topic-word: $\beta_k \sim \text{Dir}(\eta)$
- Topic-proportions: $\theta_d \sim \text{Dir}(\alpha)$
- Topic assignment: $z_{d,n} \sim \text{Mult}(\theta_d)$
- Word assginment: $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

Graphical Models

Graphical Models



$$p(w, z, \theta, \beta|\eta, \alpha) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n}) p(z_{d,n}|\theta_d)$$

Lingo

node/vertice random variable

edge stochastic relationship

plate product

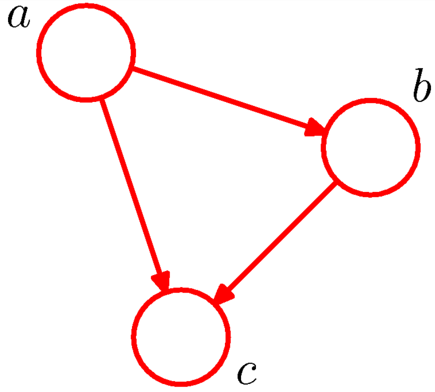
directed graph often known as Bayesian network

undirected graph often known as Markov Random Field

BayesNet

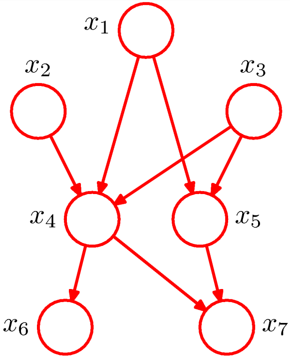
```
\usetikzlibrary{bayesnet}
\usetikzlibrary{positioning}
\begin{tikzpicture}
  \node [obs] (w) {$\mathbf{w}_{d,n}$};
  \node [latent,left=of w] (z) {$\mathbf{z}_{d,n}$};
  \node [latent,left=of z] (t) {$\theta_d$};
  \node [const,left=of t] (a) {$\alpha$};
  \node [latent,above=of w] (r) {$\beta_k$};
  \node [const,left=of r] (b) {$\eta$};
  \edge {a} {t};\edge {t} {z};\edge {z} {w};
  \edge {r} {w};\edge {b} {r};
  \plate {} {(z)(w)} {$N$};
  {\tikzset{plate caption/.append style={below=5pt of #1.sou
  \plate[inner sep=0.3cm] {} {(z)(w)(t)} {$D$};
  \plate {} {(r)} {$K$};
```

Directed Graphs



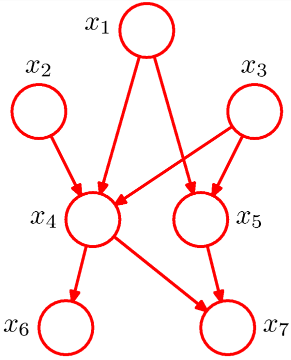
$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

Directed Graphs



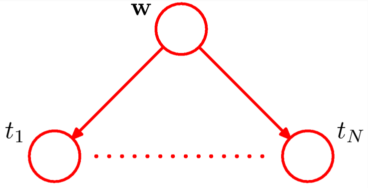
$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_5$$

Directed Graphs



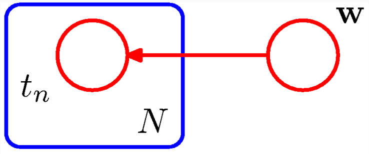
$$p(\mathbf{x}) = \prod_i p(x_i | \mathrm{pa}_i)$$

Linear Regression



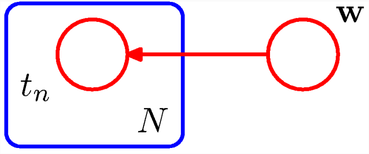
$$p(\mathbf{t}, \mathbf{W}|\mathbf{X}, \alpha, \beta) = \prod_{i}^{N} p(t_i|\mathbf{W}, x_i, \beta)p(\mathbf{W}|\alpha)$$

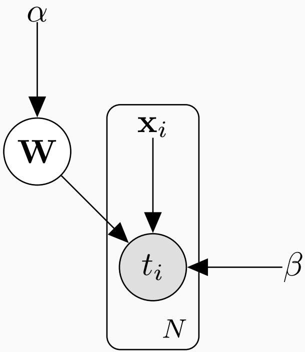
Linear Regression



$$p(\mathsf{t}, \mathsf{W} | \mathsf{X}, \alpha, \beta) = \prod_{i}^{N} p(t_i | \mathsf{W}, x_i, \beta) p(\mathsf{W} | \alpha)$$

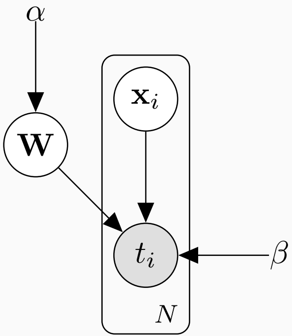
Linear Regression





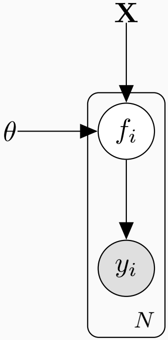
$$p(\mathbf{t}, \mathbf{W}|\mathbf{X}, \alpha, \beta) = \prod_{i}^{N} p(t_i|\mathbf{W}, x_i, \beta)p(\mathbf{W}|\alpha)$$

Unsupervised Linear Regression



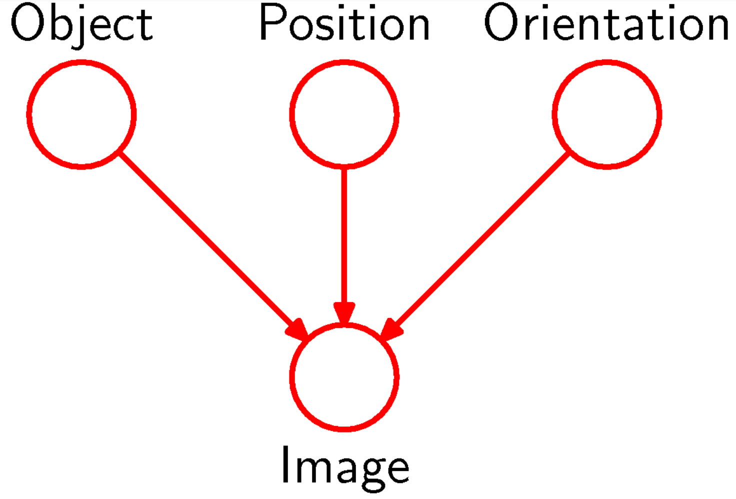
$$p(\mathbf{t}, \mathbf{x}, \mathbf{W}|\alpha, \beta) = \prod_i^N p(t_i|\mathbf{W}, x_i, \beta)p(\mathbf{W}|\alpha)p(\mathbf{x})$$

Gaussian process regression

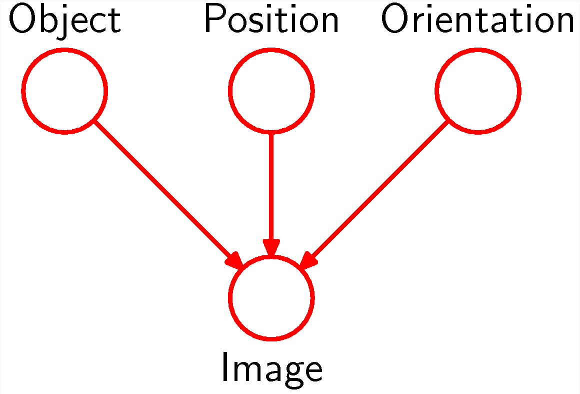


$$p(\mathbf{y}, \mathbf{f} | \mathbf{X}, \theta) = \prod_{i=1}^{N} p(y_i | f_i) p(f_i | \mathbf{X}, \theta)$$

Generative Models

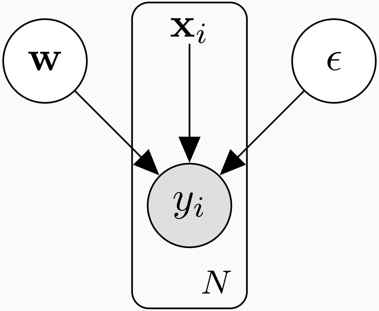


Explaining Away



- The Object variable *explains away* variance associated with objects from the image
  - $\rightarrow$ position won't contain object variations
  - $\rightarrow$ orientation won't contain object variations

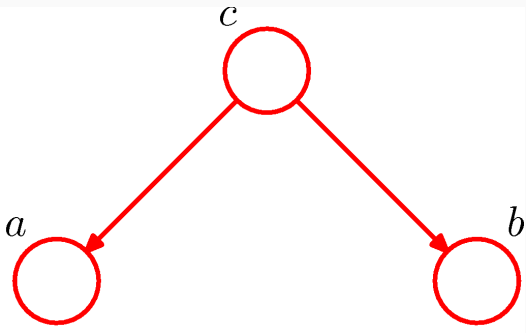
Explaining Away



$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \epsilon \quad y_i - \epsilon = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$$

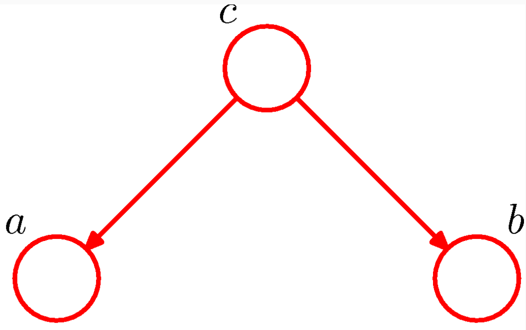
- ϵ *explains away* the noise from the data so that $\mathbf{w}$ can represent the signal

Independency



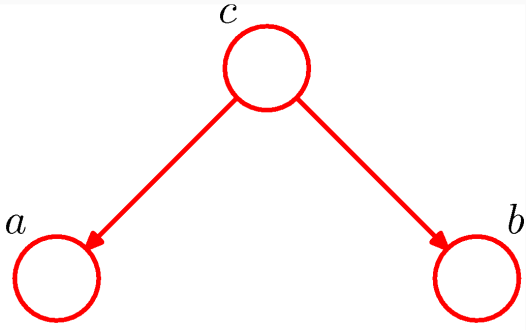
$$p(a|b,c) = p(a|c)$$

Independency



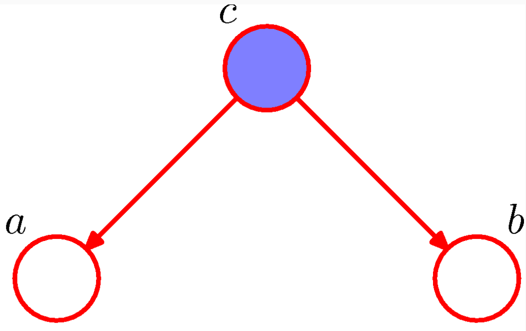
$$p(a, b|c) = p(a|b, c)p(b|c) = p(a|c)p(b|c)$$

Independency



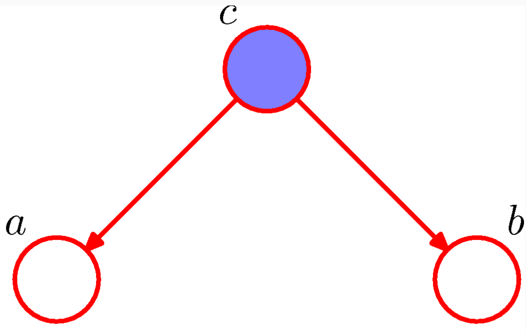
$$p(a, b) = \int p(a|c)p(b|c)p(c)\mathrm{d}c \neq p(a)p(b)$$

Conditional Independency



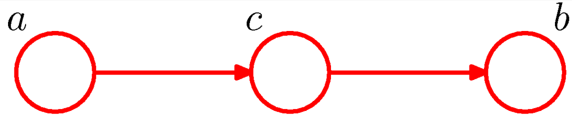
$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)$$

Conditional Independency



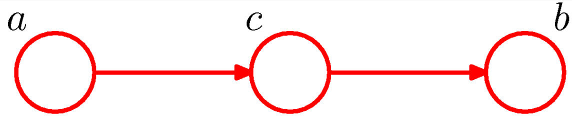
the path $a \leftrightarrow b$ is tail-tail in c therefore a and b are conditionally independent given c

Independency



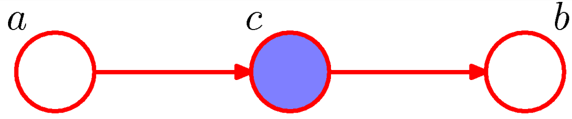
$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

Independency



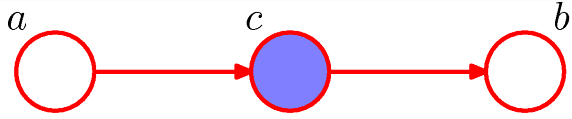
$$p(a, b) = p(a) \int p(c|a)p(b|c)\mathrm{d}c \neq p(a)p(b)$$

Conditional Independency



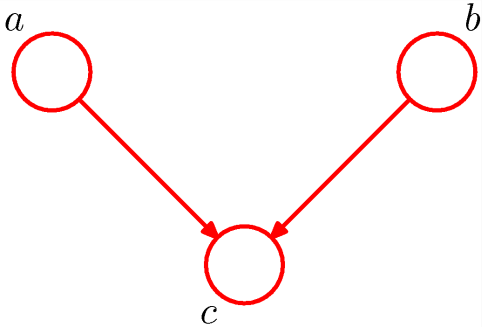
$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

Conditional Independency



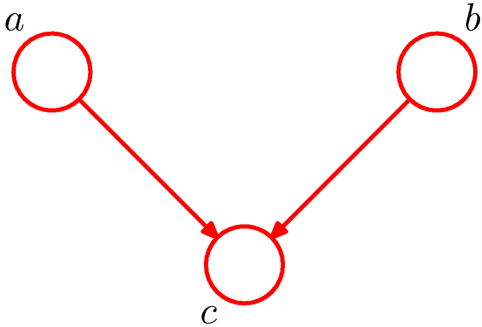
the path $a \leftrightarrow b$ *is head-tail in* c *therefore* a *and* b *are conditionally independent given* c *as* c *blocks the path*

Independency



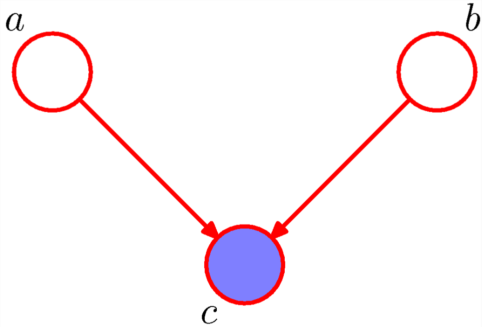
$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

Independency



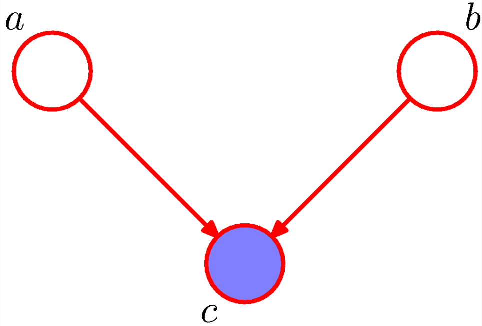
$$p(a, b) = \int p(c|a, b)p(a)p(b)\mathrm{d}c = p(a)p(b)$$

Conditional Independency



$$p(a, b|c) = \frac{p(c|a, b)p(a)p(b)}{p(c)} \neq p(a)p(b)$$

Conditional Independency



the path $a \leftrightarrow b$ is head-head in c therefore a and b are not conditionally independent given c as the conditioning "unblocks" the path

Rules

tail-to-tail when not observed the nodes are not independent, when observed makes them conditionally independent

Rules

tail-to-tail when not observed the nodes are not independent,
when observed makes them conditionally independent

head-to-tail when not observed the nodes are not independent,
when observed makes them conditionally independent

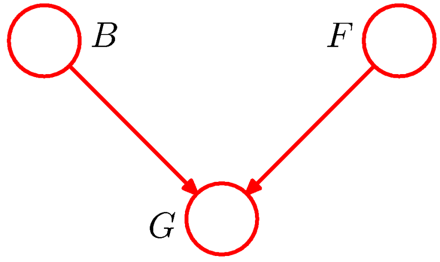
Rules

tail-to-tail when not observed the nodes are not independent,
when observed makes them conditionally independent

head-to-tail when not observed the nodes are not independent,
when observed makes them conditionally independent

head-to-head when not observed the nodes are independent,
when observed makes them dependent

Example p. 377 [2]

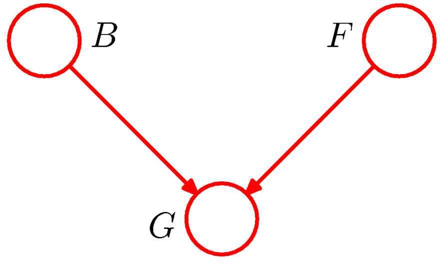


B Battery: 1 $\rightarrow$ Full, 0 $\rightarrow$ Empty

F Fuel Tank: 1 $\rightarrow$ Full, 0 $\rightarrow$ Empty

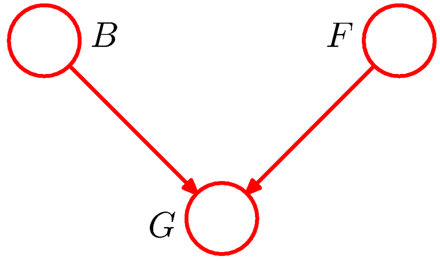
G Fuel Gauge: 1 $\rightarrow$ Indicates Full, 0 $\rightarrow$ Indicates Empty

Example p. 377 [2]



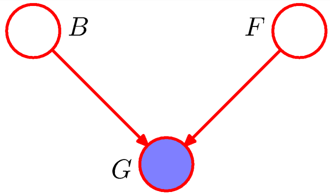
$$p(B = 1) = 0.9$$
$$p(F = 1) = 0.9$$

Example p. 377 [2]



$$p(G = 1|B = 1, F = 1) = 0.8$$
$$p(G = 1|B = 1, F = 0) = 0.2$$
$$p(G = 1|B = 0, F = 1) = 0.2$$
$$p(G = 1|B = 0, F = 0) = 0.1$$

Example p. 377 [2]



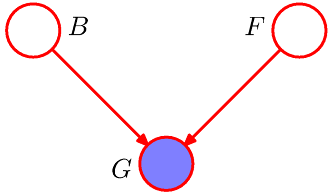
We observe an empty fuel tank $G = 0$

$$p(G=0) = \sum_{B \in \{0,1\}} \sum_{F \in \{0,1\}} p(G=0|B,F)p(B)p(F) = 0.315$$

$$p(G=0|F=0) = \sum_{B \in \{0,1\}} p(G=0|B,F=0)p(B) = 0.81$$

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \approx 0.257$$

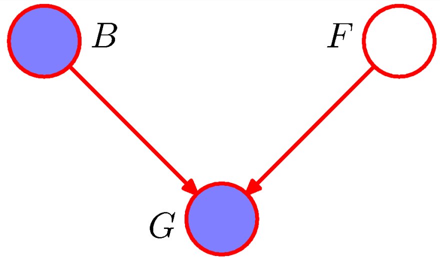
Example p. 377 [2]



$$p(F = 0|G = 0) > p(F = 0)$$

The gauge does provide information about the tank

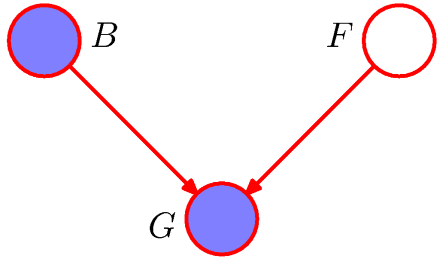
Example p. 377 [2]



We observe an empty fuel tank $G = 0$ and Battery empty $B = 0$

$$p(F = 0|G = 0, B = 0) = \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0|B = 0, F)p(F)} \approx 0.111$$

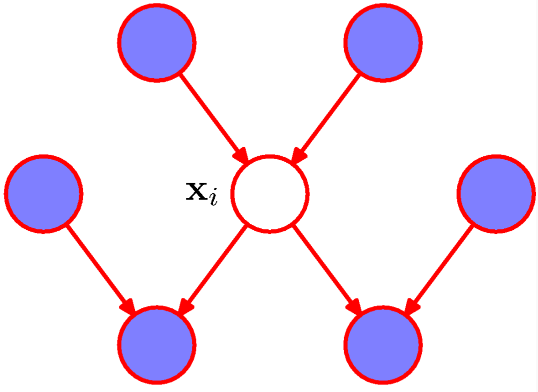
Example p. 377 [2]



$$p(F=0|G=0) > p(F=0|G=0,B=0) > P(F=0)$$

Knowing that the battery is empty explains away the information about the Gauge indicating empty

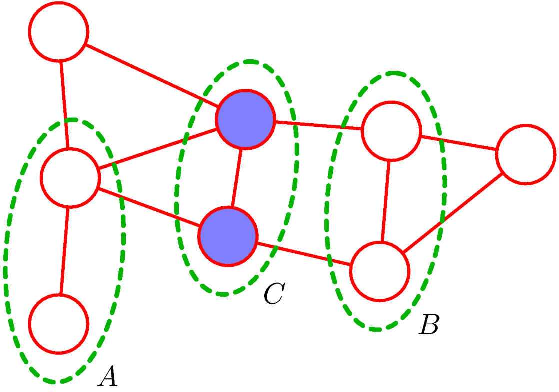
Markov Blanket p. 382 [2]



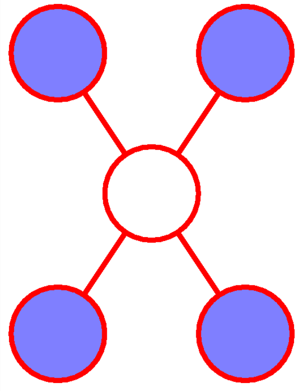
Definition (Markov Blanket)

Node x_i conditioned on all the remaining nodes in the graph can be written as conditioned on only its *parents* and *co-parents* these nodes make up the markov blanket

Undirected Graphs



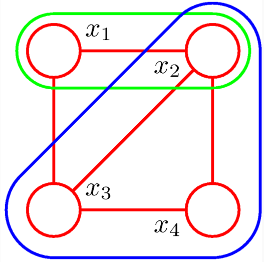
$$P(A, B|C) = P(A|C)p(B|C)$$



Definition (Markov Blanket)

The Markov Blanket for an undirected graph for node x_i contains only its direct neighbours

Undirected Graphs



Definition (Clique)

A subset of nodes such that there exists an edge between every pair of nodes in the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C = \Psi(\mathbf{x}_C)$$

Image Segmentation/Denoising

- Likelihood: each pixel is generated from its true pixel value

$$p(y_i|x_i)$$

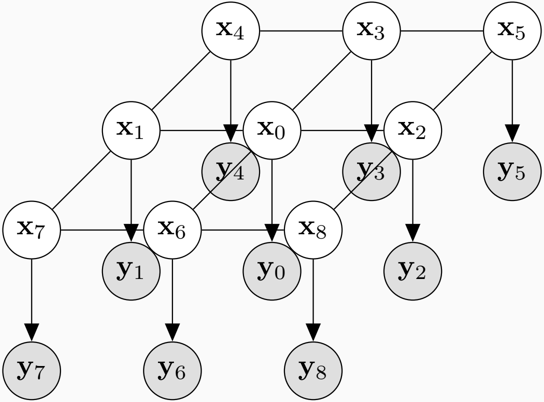
- Prior: neighbouring pixels are likely to be the same

$$p(x_i = 1, x_j = 1) > p(x_i = 0, x_j = 1) = p(x_i = 1, x_j = 0)$$

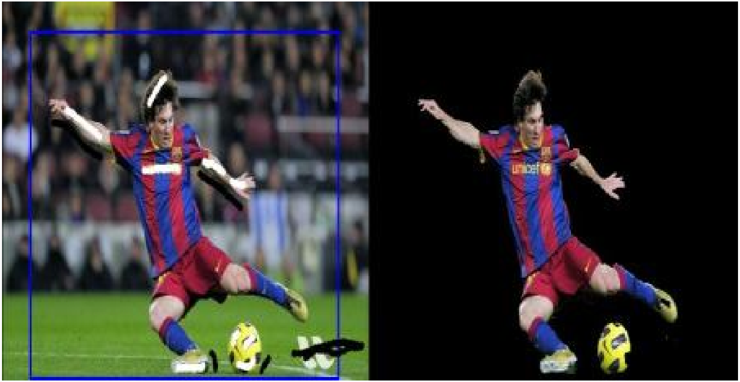
- Neighbourhood specifies a clique

$$p(\mathbf{x}) = \prod_{N} \Psi(\mathbf{x}_n)$$

Markov Random Field



GrabCut [3]



Summary

Summary

- Graphical models is just a language of what we have been doing
- Much easier to talk about when thinking of new models
- Directed graphical models implies building up conditional probabilities
- Undirected models are joint probabilities

Building Models

- we can now build models by trying to understand the generative process of the observed data
- this process leads to a formulation in terms of latent variables
- we have seen how one can formulate priors to make assumptions about these
- we know how to make inference in tractable models through conjugacy
- we know how to make inference using ML, MAP, Type-II ML and exact Bayesian

The Reminder

Tuesday Laplace approximation

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

Tuesday Variational inference

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

Tuesday Variational inference

- How can we approximate integrals with surrogate models

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

Tuesday Variational inference

- How can we approximate integrals with surrogate models

w9 Current topics

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

Tuesday Variational inference

- How can we approximate integrals with surrogate models

w9 Current topics

w10 Summary and Exam prep (1 Lecture)

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

Tuesday Variational inference

- How can we approximate integrals with surrogate models

w9 Current topics

w10 Summary and Exam prep (1 Lecture)

w11 Invited talks

The Reminder

Tuesday Laplace approximation

- What to do when we do regression to discrete output

Monday Sampling

- How can we approximate integrals with sums

Tuesday Variational inference

- How can we approximate integrals with surrogate models

w9 Current topics

w10 Summary and Exam prep (1 Lecture)

w11 Invited talks

w12 nothing

eof

References

David M Blei, Andrew Y Ng, and Michael I Jordan.
Latent dirichlet allocation.
3:993–1022, March 2003.

Christopher M. Bishop.
***Pattern Recognition and Machine Learning (Information Science and Statistics)*.**
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake.
GrabCut: interactive foreground extraction using iterated graph cuts.
ACM Transactions on Graphics (TOG), 23(3):309–314, August 2004.