Graphical Models
(ch 8.1-2)

from “Pattern Recognition and Machine Learning” by C. M. Bishop

Contents

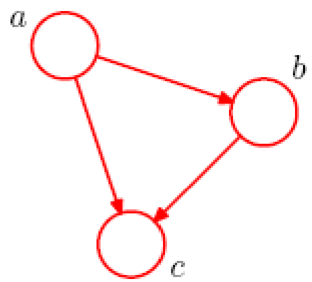
Bayesian Networks

- A simple way to visualize the structure of a probabilistic model.
- Insights (including conditional independence properties) can be obtained
- Complex computations can be expressed in terms of graphical manipulations
- Each node represents a random variable and links represent probabilistic relationships between these variables
- The graph captures the decomposition of the joint distributions over the set of random variables into a product of factors
  - Directed graphical models: Bayesian networks
  - Undirected graphical models: Markov random fields

Bayesian Networks

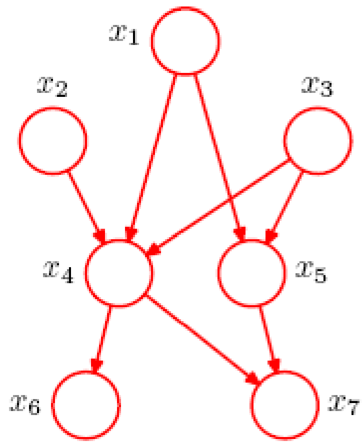
- $$p(a, b, c) = p(c|a, b)p(a, b). \tag{8.1}$$

  $$p(a, b, c) = p(c|a, b)p(b|a)p(a). \tag{8.2}$$



- $$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5). \tag{8.4}$$

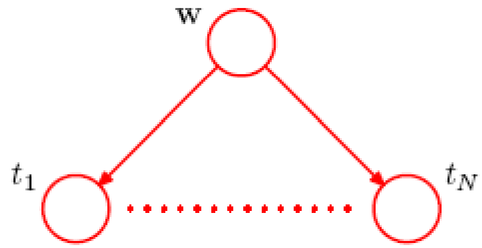
  $$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k) \tag{8.5}$$

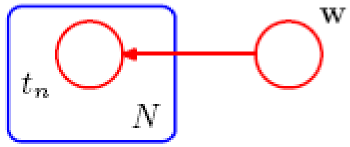


Example: Polynomial regression

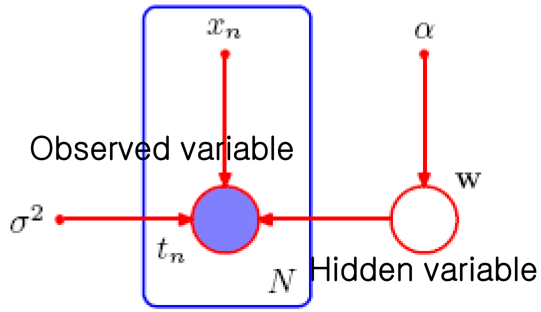
- $$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}). \qquad (8.6)$$

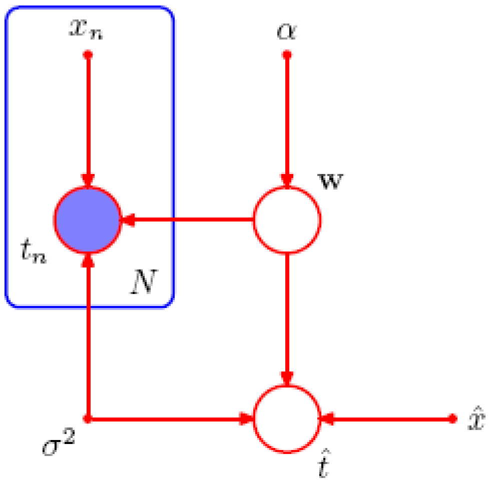
$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2).$$





- Addition of a new input value $\hat{x}$ and the corresponding variable $\hat{t}$





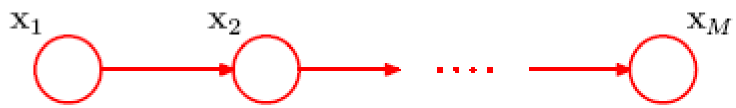
$$p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha) p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2). \qquad (8.8)$$

Discrete variables

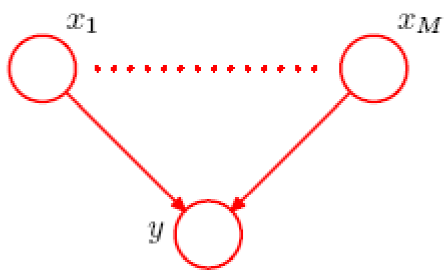
- Discrete variables with K possible states

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad\qquad p(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\mu}) = \prod_{k=1}^{K}\prod_{l=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}.$$

- Number of independent parameters:
  - M independent discrete variables: M(K−1)
  - M fully connected discrete variables: $K^M - 1$
  - A chain of M discrete nodes: K−1+(M−1)K(K−1)



- Controlling the number of parameters:
  - 2^M vs logistic sigmoid function (linear with M)



$$p(y = 1|x_1, \ldots, x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) \qquad (8.10)$$

Linear Gaussian models

- $$p(x_i|\mathrm{pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j\in\mathrm{pa}_i} w_{ij}x_j + b_i, v_i\right) \tag{8.11}$$

$$\ln p(\mathbf{x}) = \sum_{i=1}^{D} \ln p(x_i|\mathrm{pa}_i) \tag{8.12}$$

$$= -\sum_{i=1}^{D} \frac{1}{2v_i}\left(x_i - \sum_{j\in\mathrm{pa}_i} w_{ij}x_j - b_i\right)^2 + \mathrm{const} \tag{8.13}$$

- Mean and covariance of the joint distribution:

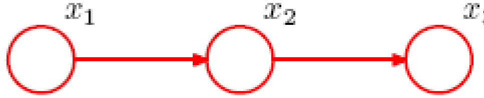
$$x_i = \sum_{j\in\mathrm{pa}_i} w_{ij}x_j + b_i + \sqrt{v_i}\epsilon_i \tag{8.14}$$

$$\mathbb{E}[x_i] = \sum_{j\in\mathrm{pa}_i} w_{ij}\mathbb{E}[x_j] + b_i. \tag{8.15}$$

$$\begin{aligned}\mathrm{cov}[x_i, x_j] &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])\right] \\ &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\left\{\sum_{k\in\mathrm{pa}_j} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\right\}\right] \\ &= \sum_{k\in\mathrm{pa}_j} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij}v_j \end{aligned} \tag{8.16}$$

Linear Gaussian models

- Two extreme cases, intermediate case
  - no links: D isolated nodes, total of D+D parameters (diagonal covariance)
  - Fully connected: each node has all lower numbered nodes as parents. D(D−1)/2 + D parameters
  - Intermediate case: partially connected



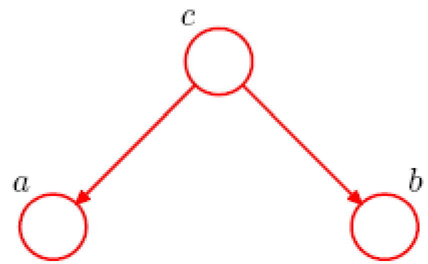
$$\boldsymbol{\mu} = (b_1, b_2 + w_{21}b_1, b_3 + w_{32}b_2 + w_{32}w_{21}b_1)^{\mathrm{T}} \tag{8.17}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 \\ w_{21}v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2(v_2 + w_{21}^2 v_1) \end{pmatrix}. \tag{8.18}$$

- Conditional distribution of node i:

$$p(\mathbf{x}_i|\mathrm{pa}_i) = \mathcal{N}\left(\mathbf{x}_i \,\middle|\, \sum_{j\in\mathrm{pa}_i} \mathbf{W}_{ij}\mathbf{x}_j + \mathbf{b}_i, \boldsymbol{\Sigma}_i\right) \tag{8.19}$$

Conditional independence



- Conditional independence
(*a* is conditionally independent of *b* given *c*)

$$p(a|b,c) = p(a|c). \tag{8.20}$$

- *a* and *b* are statistically independent given *c* :

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c). \end{aligned} \tag{8.21}$$

$$a \perp\!\!\!\perp b \mid c \tag{8.22}$$

- D-separation: conditional independence properties can be checked without analytical manipulations

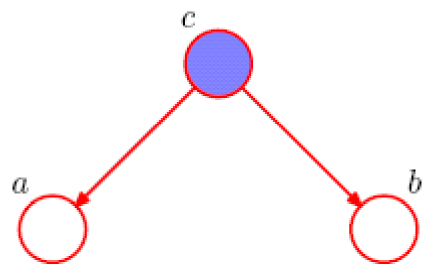
Three example graphs

- Example 1 (tail to tail)
  - Without conditions:

$$p(a, b, c) = p(a|c)p(b|c)p(c). \quad (8.23)$$

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c). \quad (8.24)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset \quad (8.25)$$



  - With conditions:

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

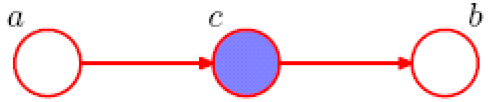
$$a \perp\!\!\!\perp b \mid c.$$

Three example graphs

- Example 2 (head to tail)
  - With conditions:

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

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



- Example 3 (head to head)
  - Without conditions:

$$p(a,b,c) = p(a)p(b)p(c|a,b). \qquad (8.28)$$

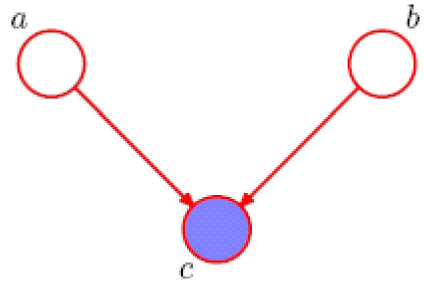
$$p(a,b) = p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset. \qquad (8.29)$$

  - With conditions:

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

$$a \not\perp\!\!\!\perp b \mid c.$$

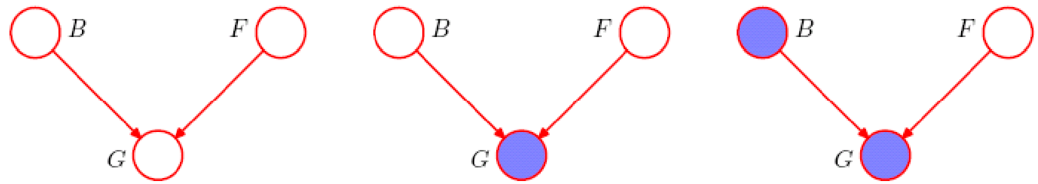


Three example graphs

- A head to head node blocks a path if it is unobserved, but once the node or at least one of its descendents is observed, the path becomes unblocked
- B: state of battery, F: state of fuel tank, G: state of fuel gauge

$$\begin{aligned} p(B=1) &= 0.9 \\ p(F=1) &= 0.9. \end{aligned}$$

$$\begin{aligned} p(G=1|B=1,F=1) &= 0.8 \\ p(G=1|B=1,F=0) &= 0.2 \\ p(G=1|B=0,F=1) &= 0.2 \\ p(G=1|B=0,F=0) &= 0.1 \end{aligned}$$



$$p(G=0) = \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G=0|B,F)p(B)p(F) = 0.315 \qquad (8.30)$$

$$p(G=0|F=0) = \sum_{B\in\{0,1\}} p(G=0|B,F=0)p(B) = 0.81 \qquad (8.31)$$

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \simeq 0.257 \qquad (8.32)$$

$$p(F=0|G=0,B=0) = \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)} \simeq 0.111 \qquad (8.33)$$

- the state of fuel tank and battery are dependent after observing fuel gage

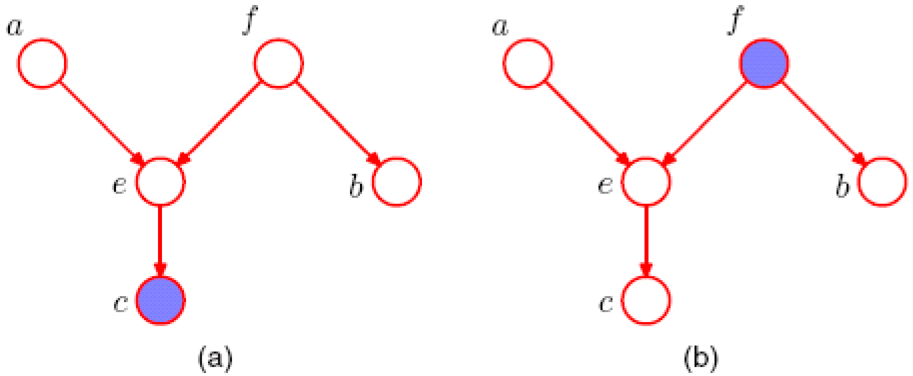
D-separation

- Conditions for D-separation : if all paths are blocked, then A is d-separated from B by C

(a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or

(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C.

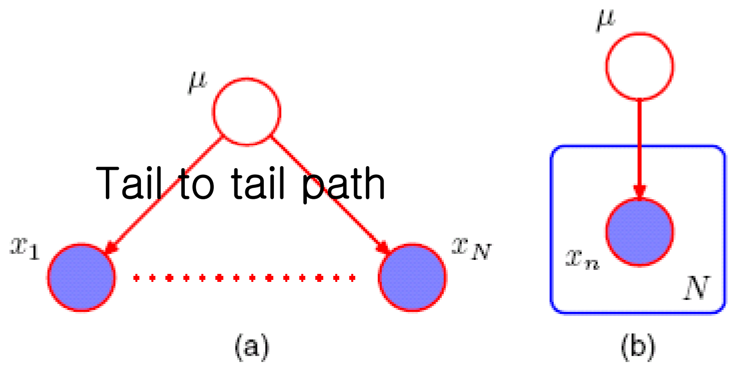
  - In (a), a is not d-separated from b by f
  - In (b), a is d-separated from b by f



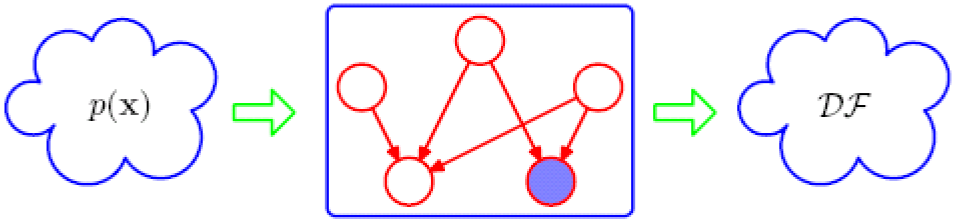
- Example of i.i.d. samples

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu). \quad (8.34)$$

$$p(\mathcal{D}) = \int_0^{\infty} p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n). \quad (8.35)$$



- Graphical model as a filter (equivalence between d-separation property and graph factorization)



- Markov blanket, Markov boundary: set of nodes for parents, children, coparents
  - Minimal set of nodes to isolate x_i from the rest of the graph

$$
\begin{aligned}
p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) &= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_D)}{\int p(\mathbf{x}_1,\ldots,\mathbf{x}_D)\,\mathrm{d}\mathbf{x}_i} \\
&= \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_k)\,\mathrm{d}\mathbf{x}_i}
\end{aligned}
$$