Graphical Models (ch 8.1–2)

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

- **8.1 Bayesian networks**
  - Example: polynomial regression
  - Generative models
  - Discrete variables
  - Linear Gaussian models

- **8.2 Conditional independence**
  - Three example graphs
  - D–separation
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

\[ p(a, b, c) = p(c|a, b)p(a, b). \]  \hspace{1cm} (8.1)
\[ p(a, b, c) = p(c|a, b)p(b|a)p(a). \]  \hspace{1cm} (8.2)

\[ p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5). \]  \hspace{1cm} (8.4)

\[ p(x) = \prod_{k=1}^{K} p(x_k|\text{pa}_k) \]  \hspace{1cm} (8.5)
Example: Polynomial regression

\[ p(t, w) = p(w) \prod_{n=1}^{N} p(t_n \mid w). \]  \hspace{1cm} (8.6)

\[ p(t, w \mid x, \alpha, \sigma^2) = p(w \mid \alpha) \prod_{n=1}^{N} p(t_n \mid w, x_n, \sigma^2). \]

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

\[ p(\hat{t}, t, w \mid \hat{x}, x, \alpha, \sigma^2) = \left[ \prod_{n=1}^{N} p(t_n \mid x_n, w, \sigma^2) \right] p(w \mid \alpha) p(\hat{t} \mid \hat{x}, w, \sigma^2). \]  \hspace{1cm} (8.8)
Discrete variables

Discrete variables with K possible states

$$p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$p(x_1, x_2|\mu) = \prod_{k=1}^{K} \prod_{l=1}^{K} \mu_{kl}^{x_{1k}x_{2l}}.$$  

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)

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

\[ p(x_i|p_a_i) = \mathcal{N}\left(x_i \left| \sum_{j \in p_a_i} w_{ij} x_j + b_i, v_i \right. \right) \]  
(8.11)

\[
\ln p(x) = \sum_{i=1}^{D} \ln p(x_i|p_a_i) 
= - \sum_{i=1}^{D} \frac{1}{2v_i} \left( x_i - \sum_{j \in p_a_i} w_{ij} x_j - b_i \right)^2 + \text{const} 
\]  
(8.12)

Mean and covariance of the joint distribution:

\[ x_i = \sum_{j \in p_a_i} w_{ij} x_j + b_i + \sqrt{v_i} \varepsilon_i \]  
(8.14)

\[
\mathbb{E}[x_i] = \sum_{j \in p_a_i} w_{ij} \mathbb{E}[x_j] + b_i. \]  
(8.15)

\[
\text{cov}[x_i, x_j] = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])]
= \mathbb{E} \left[(x_i - \mathbb{E}[x_i]) \left\{ \sum_{k \in p_a_j} w_{jk} (x_k - \mathbb{E}[x_k]) + \sqrt{v_j} \varepsilon_j \right\} \right]
- \sum_{k \in p_a_j} w_{jk} \text{cov}[x_i, x_k] + I_{ij} v_j \]  
(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

\[ \mu = (b_1, b_2 + w_{21}b_1, b_3 + w_{32}b_2 + w_{32}w_{21}b_1)^T \]  
\[ \Sigma = \begin{pmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 \\ w_{21}v_1 & v_2 + w_{21}^2v_1 & w_{32}(v_2 + w_{21}^2v_1) \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2v_1) & v_3 + w_{32}^2(v_2 + w_{21}^2v_1) \end{pmatrix} \]

- Conditional distribution of node i:

\[ p(x_i|pa_i) = \mathcal{N} \left( x_i \left| \sum_{j \in pa_i} W_{ij}x_j + b_i, \Sigma_i \right. \right) \]
Conditional independence

- Conditional independence
  \( a \) is conditionally independent of \( b \) given \( c \)
  \[
p(a|b, c) = p(a|c).
  \] (8.20)

- \( a \) and \( b \) are statistically independent given \( c \):
  \[
p(a, b|c) = p(a|b, c)p(b|c)
  = p(a|c)p(b|c).
  \] (8.21)

  \( a \perp b \mid c \) \hspace{1cm} (8.22)

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

- **Example 1** (tail to tail)
  - **Without conditions:**
    \[
    p(a, b, c) = p(a|c)p(b|c)p(c). \tag{8.23}
    \]
    \[
    p(a, b) = \sum_c p(a|c)p(b|c)p(c). \tag{8.24}
    \]
    \[
    a \independent b \mid \emptyset \tag{8.25}
    \]
  - **With conditions:**
    \[
    p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b|c)
    \]
    \[
    a \independent b \mid c.
    \]
Three example graphs

- **Example 2 (head to tail)**
  - With conditions:
    \[
    p(a, b | c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c)
    \]
    \[
    a \perp b | c.
    \]

- **Example 3 (head to head)**
  - Without conditions:
    \[
    p(a, b, c) = p(a)p(b)p(c|a, b).
    \]
    \[
    p(a, b) = p(a)p(b)
    \]
    \[
    a \perp b | \emptyset.
    \]
  - With conditions:
    \[
    p(a, b | c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b)p(c|a, b)}{p(c)}
    \]
    \[
    a \perp b | c.
    \]
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

\[
\begin{align*}
 p(B = 1) &= 0.9 \\
 p(F = 1) &= 0.9 \\
 p(G = 1|B = 1, F = 1) &= 0.8 \\
 p(G = 1|B = 1, F = 0) &= 0.2 \\
 p(G = 1|B = 0, F = 1) &= 0.2 \\
 p(G = 1|B = 0, F = 0) &= 0.1 \\
\end{align*}
\]

\[
\begin{align*}
 p(G = 0) &= \sum_{B \in \{0, 1\}} \sum_{F \in \{0, 1\}} p(G = 0|B, F) p(B) p(F) = 0.315 \quad (8.30) \\
 p(G = 0|F = 0) &= \sum_{B \in \{0, 1\}} p(G = 0|B, F = 0) p(B) = 0.81 \quad (8.31) \\
 p(F = 0|G = 0) &= \frac{p(G = 0|F = 0) p(F = 0)}{p(G = 0)} \approx 0.257 \quad (8.32) \\
 p(F = 0|G = 0, B = 0) &= \frac{p(G = 0|B = 0, F = 0) p(F = 0)}{\sum_{F \in \{0, 1\}} p(G = 0|B = 0, F) p(F)} \approx 0.111 \quad (8.33)
\end{align*}
\]

- 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**

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

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

- **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

\[
p(x_i | x_{(j \neq i)}) = \frac{p(x_1, \ldots, x_D)}{\int p(x_1, \ldots, x_D) \, dx_i} = \frac{\prod_k p(x_k | pa_k)}{\int \prod_k p(x_k | pa_k) \, dx_i}
\]