CSCI 590: Machine Learning

Lecture 17: Bayesian networks
Instructor: Murat Dundar

Acknowledgement:
These slides are taken from course textbook website
http://research.microsoft.com/~cmbishop/prml/
Bayesian Networks

Directed Acyclic Graph (DAG)

\[
p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)
\]

\[
p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)
\]
Bayesian Networks

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

General Factorization

\[ p(x) = \prod_{k=1}^{K} p(x_k|pa_k) \]
Bayesian Curve Fitting (1)

\[ y(x, w) = \sum_{j=0}^{M} w_j x^j \]

\[ p(t, w) = p(w) \prod_{n=1}^{N} p(t_n | y(w, x_n)) \]
Bayesian Curve Fitting (2)

\[ p(t, w) = p(w) \prod_{n=1}^{N} p(t_n | y(w, x_n)) \]
Bayesian Curve Fitting (3)

Input variables and explicit hyperparameters

\[ p(t, w | x, \alpha, \sigma^2) = p(w | \alpha) \prod_{n=1}^{N} p(t_n | w, x_n, \sigma^2). \]
Bayesian Curve Fitting—Learning

Condition on data

\[ p(w | t) \propto p(w) \prod_{n=1}^{N} p(t_n | w) \]
Bayesian Curve Fitting—Prediction

Predictive distribution: \( p(t | \hat{x}, x, t, \alpha, \sigma^2) \propto \int p(\hat{t}, t, w | \hat{x}, x, \alpha, \sigma^2) \, dw \)

where

\[
p(\hat{t}, t, w | \hat{x}, x, \alpha, \sigma^2) = \left[ \prod_{n=1}^{N} p(t_n | x_n, w, \sigma^2) \right] p(w | \alpha)p(t | \hat{x}, w, \sigma^2)
\]
Discrete Variables (1)

General joint distribution: $K^2 - 1$ parameters

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

Independent joint distribution: $2(K - 1)$ parameters

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

General joint distribution over $M$ variables:

$K^M - 1$ parameters

$M$-node Markov chain: $K - 1 + (M - 1)K(K - 1)$ parameters
Discrete Variables: Bayesian Parameters (1)

\[ p(\{x_m, \mu_m\}) = p(x_1 | \mu_1) p(\mu_1) \prod_{m=2}^{M} p(x_m | x_{m-1}, \mu_m) p(\mu_m) \]

\[ p(\mu_m) = \text{Dir}(\mu_m | \alpha_m) \]
Discrete Variables: Bayesian Parameters (2)

\[ p(\{x_m\}, \mu_1, \mu) = p(x_1 | \mu_1) p(\mu_1) \prod_{m=2}^{M} p(x_m | x_{m-1}, \mu) p(\mu) \]
Parameterized Conditional Distributions

If $x_1, \ldots, x_M$ are discrete, $K$-state variables, $p(y = 1|x_1, \ldots, x_M)$ in general has $O(K^M)$ parameters.

The parameterized form

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

requires only $M + 1$ parameters
Linear-Gaussian Models

Directed Graph

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

Each node is Gaussian, the mean is a linear function of the parents.

Vector-valued Gaussian Nodes

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

\(a\) is independent of \(b\) given \(c\)

\[ p(a|b, c) = p(a|c) \]

Equivalently

\[ p(a, b|c) = p(a|b, c)p(b|c) = p(a|c)p(b|c) \]

Notation

\(a \perp b \mid c\)
Conditional Independence: Example 1

\[ p(a, b, c) = p(a|c)p(b|c)p(c) \]

\[ p(a, b) = \sum_c p(a|c)p(b|c)p(c) \]

\[ a \independent b \mid \emptyset \]
Conditional Independence: Example 1

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

\[ a \perp b \mid c \]
Conditional Independence: Example 2

\[ p(a, b, c) = p(a)p(c|a)p(b|c) \]

\[ p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a) \]

\[ a \perp b \mid \emptyset \]
Conditional Independence: Example 2

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

\[ = \frac{p(a)p(c|a)p(b|c)}{p(c)} \]

\[ = p(a|c)p(b|c) \]

\[ a \perp b \mid c \]
Conditional Independence: Example 3

Note: this is the opposite of Example 1, with $c$ unobserved.

\[ p(a, b, c) = p(a)p(b)p(c|a, b) \]

\[ p(a, b) = p(a)p(b) \]

\[ a \perp b \mid \emptyset \]
Conditional Independence: Example 3

Note: this is the opposite of Example 1, with $c$ observed.

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

\[ a \perp b \mid c \]
“Am I out of fuel?”

\[
p(G = 1|B = 1, F = 1) = 0.8 \\
p(G = 1|B = 1, F = 0) = 0.2 \\
p(G = 1|B = 0, F = 1) = 0.2 \\
p(G = 1|B = 0, F = 0) = 0.1
\]

\[
p(B = 1) = 0.9 \\
p(F = 1) = 0.9 \\
\text{and hence} \\
p(F = 0) = 0.1
\]

\[B = \text{Battery (0=flat, 1=fully charged)}\]
\[F = \text{Fuel Tank (0=empty, 1=full)}\]
\[G = \text{Fuel Gauge Reading} \quad (0=empty, \ 1=full)\]
“Am I out of fuel?”

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

Probability of an empty tank increased by observing \( G = 0 \).
Probability of an empty tank reduced by observing $B = 0$.
This referred to as “explaining away”.

\[
p(F = 0|G = 0, B = 0) = \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0, 1\}} p(G = 0|B = 0, F)p(F)} \\ \approx 0.111
\]
D-separation

- $A$, $B$, and $C$ are non-intersecting subsets of nodes in a directed graph.
- A path from $A$ to $B$ is blocked if it contains a node such that either
  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, are in the set $C$.
- If all paths from $A$ to $B$ are blocked, $A$ is said to be d-separated from $B$ by $C$.
- If $A$ is d-separated from $B$ by $C$, the joint distribution over all variables in the graph satisfies $A \perp B \mid C$. 
D-separation: Example

\[
\begin{align*}
& a \perp b \mid c \\
& a \not\perp b \mid c
\end{align*}
\]
D-separation: I.I.D. Data

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

\[ p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu) \, d\mu \neq \prod_{n=1}^{N} p(x_n) \]
The Markov Blanket

\[ p(x_i | x_{\{j \neq i\}}) = \frac{p(x_1, \ldots, x_M)}{\int p(x_1, \ldots, x_M) \, dx_i} \times \prod_k \frac{p(x_k | p_{a_k})}{\int \prod_k p(x_k | p_{a_k}) \, dx_i} \]

Factors independent of \(x_i\) cancel between numerator and denominator.
Cliques and Maximal Cliques

Clique

Maximal Clique
Joint Distribution

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

where \( \psi_C(x_C) \) is the potential over clique \( C \) and

\[ Z = \sum_x \prod_C \psi_C(x_C) \]

is the normalization coefficient; note: \( MK \)-state variables \( \to K^M \) terms in \( Z \).

Energies and the Boltzmann distribution

\[ \psi_C(x_C) = \exp \{-E(x_C)\} \]
Illustration: Image De-Noising (1)

Original Image

Noisy Image
Illustration: Image De-Noising (2)

\[ E(x, y) = h \sum_{i} x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_{i} x_i y_i \]

\[ p(x, y) = \frac{1}{Z} \exp\{-E(x, y)\} \]
Illustration: Image De-Noising (3)

Noisy Image

Restored Image (ICM)
Illustration: Image De-Noising (4)

Restored Image (ICM)  Restored Image (Graph cuts)