Pattern Recognition And Machine Learning

8.2 Conditional Independence

Bohyeon Park

University of Seoul, Statistics

2019.01.31

Outline

- Conditional Independence
- Three example graphs
  - Diverging Connections.
  - Serial Connections.
  - Converging Connections.
- D-separation

Conditional Independence

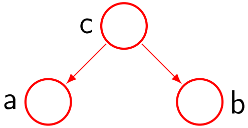
Conditional Independence

- Definition
  - a, b, c: variables.
  - a and b are conditionally independent given c.

$$
\begin{aligned}
p(a, b \mid c) &= p(a \mid b, c)p(b \mid c). \\
&= p(a \mid c)p(b \mid c). \\
\Leftrightarrow \quad & p(a \mid b, c) = p(a \mid c). \\
\Leftrightarrow \quad & a \perp\!\!\!\perp b \mid c.
\end{aligned}
$$

Three Example Graphs

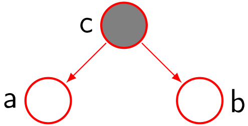
Diverging Connections



- a, b, c: variables.
- c: tail-to-tail.
- Three variables aren't observed.
  - $\emptyset$: empty set.
  - $a \not\perp\!\!\!\perp b \mid \emptyset$: a and b are dependent.

$$
\begin{aligned}
p(a, b) = \sum_c p(a, b, c) &= \sum_c p(a \mid c) p(b \mid c) p(c). \\
&\neq p(a)p(b).
\end{aligned}
$$

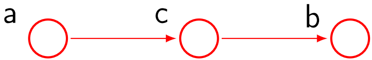
Diverging Connections



- The variable c is observed.
  - $a \perp\!\!\!\perp b \mid c$: a and b are conditionally independent given c.

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

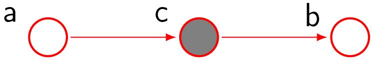
Serial Connections



- a, b, c: variables.
- c: head-to-tail.
- Three variables aren't observed.
  - $\emptyset$: empty set.
  - $a \not\perp\!\!\!\perp b \mid \emptyset$: a and b are dependent.

$$
\begin{aligned}
p(a, b) &= \sum_c p(a)p(c \mid a)p(b \mid c) = p(a)p(b \mid a). \\
&\neq p(a)p(b).
\end{aligned}
$$

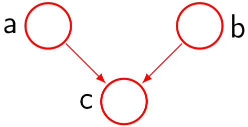
Serial Connections



- The variable c is observed.
  - $a \perp\!\!\!\perp b \mid c$: a and b are conditionally independent given c.

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

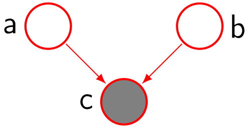
Converging Connections



- a, b, c: variables.
- c: head-to-head.
- Three variables aren't observed.
  - $\emptyset$: empty set.
  - $a \perp\!\!\!\perp b \mid \emptyset$: a and b are independent.

$$p(a, b) = \sum_c p(a)p(b)p(c \mid a, b) = p(a)p(b).$$

Converging Connections

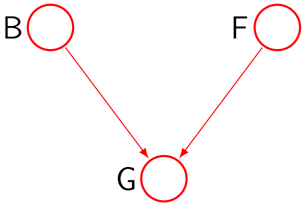


- The variable c is observed.
  - a, b, c: Variables.
  - $a \not\perp\!\!\!\perp b \mid c$: a and b are conditionally dependent given c.

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

Example: The fuel system on a car

- B, F, G: three binary random variables.
- B: the state of a battery (charged: B=1 or flat: B=0).
- F: the state of the fuel tank (full: F=1 or empty: F=0).
- G: the state of an electric fuel gauge (full: G=1 or empty: G=0).



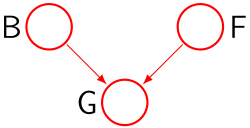
Example: The fuel system on a car

- The prior probabilities.
  - p(B=1)=0.9.
  - p(F=1)=0.9.

- Given the state of B and F, the fuel gauge(G) reads full with probabilities given by.
  - $p(G = 1 \mid B = 1, F = 1) = 0.8.$
  - $p(G = 1 \mid B = 1, F = 0) = 0.2.$
  - $p(G = 1 \mid B = 0, F = 1) = 0.2.$
  - $p(G = 1 \mid B = 0, F = 0) = 0.1.$

- Compare $p(F = 0), p(F = 0 \mid G = 0), p(F = 0 \mid G = 0, B = 0)$.

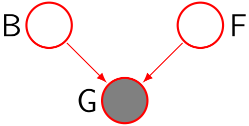
Example: The fuel system on a car



1. Before we observe any data,

$$p(F = 0) = 0.1.$$

Example: The fuel system on a car



2. Suppose that we observed G=0,

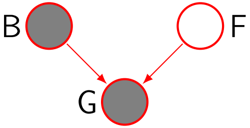
$$p(F = 0 \mid G = 0) \simeq 0.257.$$

$$\Rightarrow \quad p(G = 0) = \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G = 0 \mid B, F)p(B)p(F) = 0.315.$$

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

$$p(F = 0 \mid G = 0) = \frac{p(G = 0 \mid F = 0)p(F = 0)}{p(G = 0)} \simeq 0.257.$$

Example: The fuel system on a car



3. Suppose that we observed G=0 and B=0,

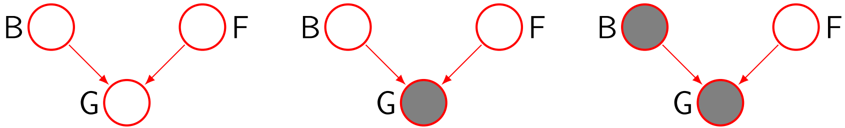
$$p(F = 0 \mid G = 0, B = 0) \simeq 0.111.$$

$$\Rightarrow \quad p(G = 0 \mid B = 0) = \sum_{F \in \{0,1\}} p(G = 0 \mid B = 0, F) p(F) = 0.81.$$

$$p(F = 0 \mid G = 0, B = 0) = \frac{p(G = 0 \mid B = 0, F = 0) p(F = 0)}{p(G = 0 \mid B = 0)}.$$

$$\simeq 0.111.$$

Example: The fuel system on a car



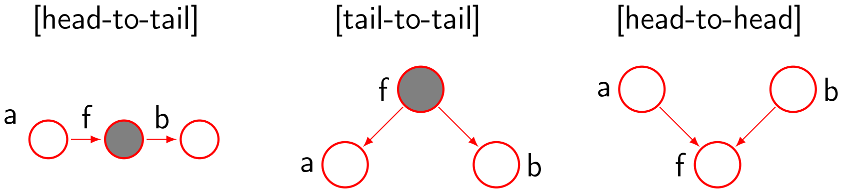
- Results
  - $p(F=0) = 0.1, p(F=0 \mid G=0) \simeq 0.257,$
    $p(F=0 \mid G=0, B=0) \simeq 0.111.$
  - $p(F=0 \mid G=0) > p(F=0).$
  - $p(F=0 \mid G=0) > p(F=0 \mid G=0, B=0).$
    $\rightarrow \quad B \not\perp\!\!\!\perp F \mid G.$
  - $p(F=0 \mid G=0, B=0) > p(F=0).$

D-Separation

D-separation

- A, B, C: arbitrary nonintersecting sets of nodes.
- We wish to ascertain whether a particular $A \perp\!\!\!\perp B \mid C$ is implied by a given DAG.
- We consider all possible paths from any node in A to any node in B.

D-separation



blocked

- Any such path is said to be *blocked* if it includes a node such that either
  - 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. e.g. $f \in C$.
  - the arrows on the path meet head-to-head at the node, and neither the node, nor any of its descendants, is in the set C. e.g. $f \notin C$.

D-separation

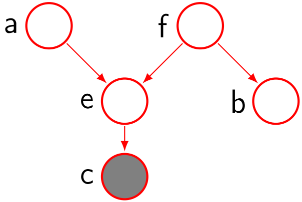
D-separation

If all paths are blocked,

then A is said to be d-separated from B by C.

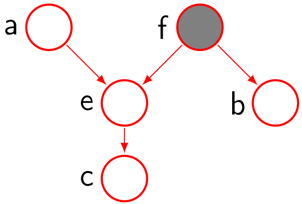
Example 1

- The path from a to b is not blocked by f.
  - f: tail-to-tail, not observed.
- The path from a to b is not blocked by e.
  - e: head-to-head, not observed.
  - c: observed.

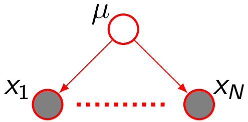


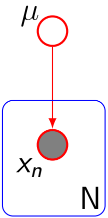
Example 2

- The path from a to b is blocked by f.
  - f: tail-to-tail, observed.
- The path from a to b is blocked by e.
  - e: head-to-head, not observed.
  - c: not observed.



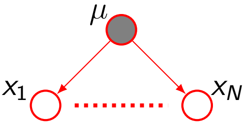
Example 3

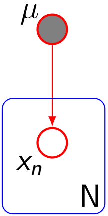




- Consider the problem of finding the posterior distribution for the mean of a univariate Gaussian distribution.
  - $D = \{x_1, \ldots, x_N\}$.
  - p(μ): prior distribution.
  - p($x_n \mid \mu$): conditional distributions, $n = 1, \ldots, N$.
  - p($\mu \mid D$): posterior distribution.
  - Our goal: $p(\mu \mid D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu}$?.

Example 3

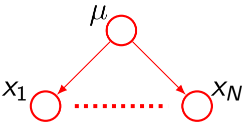


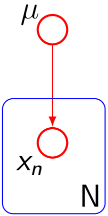


- Consider p$(D \mid \mu)$.
  - $D = \{x_1, \ldots, x_N\}$.
  - this path is tail-to-tail with respect to the observed node μ.
  - this path is blocked.
  - the observations D are independent given μ.

$$p(D \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu).$$

Example 3

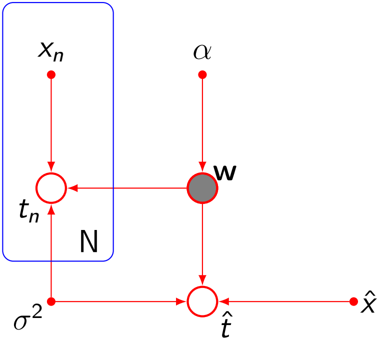




- Consider p(D).

$$p(D) = \int_0^\infty p(D \mid \mu)p(\mu)d\mu \neq \prod_{n=1}^{N} p(x_n).$$

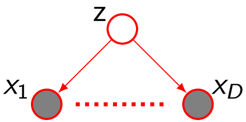
Example 4: Bayesian polynomial regression



- t_n, $\mathbf{w}$, $\hat{t}$: stochastic nodes.
- $\mathbf{w}$: tail-to-tail.

$$t_n \perp\!\!\!\perp \hat{t} \mid \mathbf{w}.$$

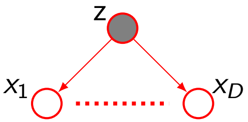
Example 5: Naive Bayes model



- We use conditional independence assumptions to simplify the model structure.
  - We wish to assign observed values of x to one of K classes.
  - $\mathbf{x}$: observed vector, $\mathbf{x} = (x_1 \ldots, x_D)^t$.
  - $\mathbf{z}$: K-dimensional binary vector (One-hot encoding).
  - μ_k: the prior probability of class C_k.

$$p(\mathbf{z} \mid \mu) \to p(\mathbf{x} \mid \mathbf{z}).$$

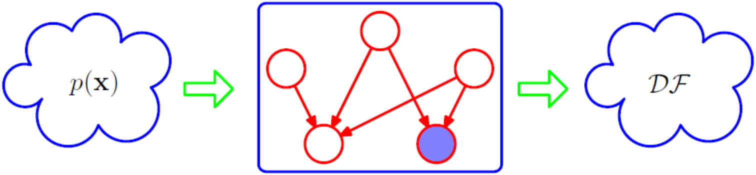
Example 5: Naive Bayes model



- The key assumption:

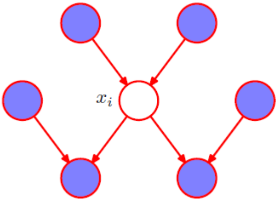
  The distributions of the input variables $x_1, \ldots, x_D$ are independent, given z.

Directed factorization, DF



- Directed factorization, DF
  - If we present to the filter the set of all possible distributions p(x) over the set of variables x,
    then the subset of distributions that are passed by the filter will be denoted DF, for directed factorization.

Markov blanket(Markov boundary)



- Markov blanket
  - The set of nodes comprising the parents, the children and the co-parents is called the Markov blanket.

*co-parents: variables corresponding to parents of node x_k other than node x_i.

Summary

- Conditional Independence
- Three example graphs
  - Diverging Connections: $a \not\perp\!\!\!\perp b \mid \emptyset, \quad a \perp\!\!\!\perp b \mid c$.
  - Serial Connections: $a \not\perp\!\!\!\perp b \mid \emptyset, \quad a \perp\!\!\!\perp b \mid c$.
  - Converging Connections: $a \perp\!\!\!\perp b \mid \emptyset, \quad a \not\perp\!\!\!\perp b \mid c$.
- D-separation
  - blocked.
- Directed factorization, DF.
- Markov blanket(Markov boundary).

Reference

- Christopher M.Bishop, Pattern Recognition and Machine Learning.


THANK YOU

Bayes' theorem

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}.$$

Additional explain

- Converging Connections example: The fuel system on a car.
- $p(F = 0 \mid G = 0)$:
  - p(B=1)=0.9.
  - p(B=0)=0.1.
  - p(F=1)=0.9.
  - p(F=0)=0.1.

  - $p(G = 0 \mid B = 1, F = 1) = 0.2.$
  - $p(G = 0 \mid B = 1, F = 0) = 0.8.$
  - $p(G = 0 \mid B = 0, F = 1) = 0.8.$
  - $p(G = 0 \mid B = 0, F = 0) = 0.9.$

Additional explain

- $$\begin{aligned} p(G=0) &= \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G=0 \mid B,F)p(B)p(F). \\ &= \quad p(G=0 \mid B=1, F=1)p(B=1)p(F=1). \\ &\quad + p(G=0 \mid B=1, F=0)p(B=1)p(F=0). \\ &\quad + p(G=0 \mid B=0, F=1)p(B=0)p(F=1). \\ &\quad + p(G=0 \mid B=0, F=0)p(B=0)p(F=0). \\ &= 0.2 * 0.9 * 0.9 + 0.8 * 0.9 * 0.1. \\ &\quad + 0.8 * 0.1 * 0.9 + 0.9 * 0.1 * 0.1. \\ &= 0.315. \end{aligned}$$

Additional explain

- $$p(G=0 \mid F=0) = \sum_{B\in\{0,1\}} p(G=0 \mid B, F=0)p(B).$$
  $$= \quad p(G=0 \mid B=0, F=0)p(B=0).$$
  $$+ p(G=0 \mid B=1, F=0)p(B=1).$$
  $$= 0.9 * 0.1 + 0.8 * 0.9.$$
  $$= 0.81.$$
- $$p(F=0|G=0) = \frac{p(G=0 \mid F=0)p(F=0)}{p(G=0)}.$$
  $$= \frac{0.81 * 0.1}{0.315}.$$
  $$\simeq 0.257.$$

Additional explain

- $p(F = 0 \mid G = 0, B = 0)$:
  - $$\begin{aligned} p(G = 0 \mid B = 0) &= \sum_{F \in \{0,1\}} p(G = 0 \mid B = 0, F)p(F). \\ &= p(G = 0 \mid B = 0, F = 0)p(F = 0). \\ &\quad + p(G = 0 \mid B = 0, F = 1)p(F = 1). \\ &= 0.9 * 0.1 + 0.8 * 0.9. \\ &= 0.81. \end{aligned}$$
  - $$\begin{aligned} p(F = 0 \mid G = 0, B = 0) &= \frac{p(G = 0 \mid B = 0, F = 0)p(F = 0)}{p(G = 0 \mid B = 0)}. \\ &= \frac{0.9 * 0.1}{0.81}. \\ &\simeq 0.111. \end{aligned}$$