COMPSCI 688: Probabilistic Graphical Models
Lecture 2: More Probability and Directed Graphical Models

Dan Sheldon
Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Discrete Distributions

- Sample space $\Omega$ for all $\omega \in \Omega$
- Atomic probability $p(\omega)$ for all $\omega \in \Omega$
  \[ p(\omega) \geq 0, \quad \sum_{\omega \in \Omega} p(\omega) = 1 \]
- Events $A \subseteq \Omega$ (only things that have probabilities!)
  \[ P(A) = \sum_{\omega \in A} p(\omega) \]
- Random variable $X : \Omega \rightarrow \text{Val}(X)$ has probability mass function (PMF)
  \[ p_X(x) = P(X(\omega) = x) = P(X = x) \]

Events vs Random Variables

- A random variable $X$ is a a mapping from $\Omega$ to $\text{Val}(X)$
- But: for any random variable $X$, we can also define the probability distribution with sample space $\Omega = \text{Val}(X)$ and atomic probabilities $p_X(x)$. This is the distribution of $X$.
  
  - If we only care about events involving $X$, it’s easier to just define the distribution of $X$ without using a different underlying probability space
  
  - If we care about multiple random variables, we can similarly define their joint distribution
Random Variables and Data Sets

In ML and stats, probability distributions are defined over records described by multiple attributes modeled as random variables. This leads to joint distributions.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Blood Pressure</th>
<th>Cholesterol</th>
<th>Heart Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Med</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Male</td>
<td>Hi</td>
<td>Hi</td>
<td>Yes</td>
</tr>
<tr>
<td>Male</td>
<td>Med</td>
<td>Med</td>
<td>Yes</td>
</tr>
<tr>
<td>Male</td>
<td>Med</td>
<td>Hi</td>
<td>No</td>
</tr>
<tr>
<td>Female</td>
<td>Med</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Male</td>
<td>Low</td>
<td>Med</td>
<td>No</td>
</tr>
</tbody>
</table>

Joint Distributions: Heart Disease Example

**Example**: The joint distribution over random variables Gender, BloodPressure, Cholesterol and HeartDisease is given by a table like this:

<table>
<thead>
<tr>
<th>Gender</th>
<th>BloodPressure</th>
<th>Cholesterol</th>
<th>HeartDisease</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>L</td>
<td>L</td>
<td>N</td>
<td>0.0127</td>
</tr>
<tr>
<td>F</td>
<td>L</td>
<td>L</td>
<td>Y</td>
<td>0.0007</td>
</tr>
<tr>
<td>F</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>0.0098</td>
</tr>
<tr>
<td>F</td>
<td>L</td>
<td>M</td>
<td>Y</td>
<td>0.0009</td>
</tr>
<tr>
<td>F</td>
<td>L</td>
<td>H</td>
<td>N</td>
<td>0.0087</td>
</tr>
<tr>
<td>F</td>
<td>L</td>
<td>H</td>
<td>Y</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

- The joint distribution of random variables $X_1, \ldots, X_N$ is a probability distribution over their canonical sample space.
- The canonical sample space $\Omega$ of $X_1, \ldots, X_N$ is the Cartesian product of their domains $\Omega = \text{Val}(X_1) \times \cdots \times \text{Val}(X_N)$.
- An element of $\Omega$ is a joint assignment $(x_1, \ldots, x_N)$.
- The joint probability mass function of $X_1, \ldots, X_N$ is $p(x_1, \ldots, x_N) = P(X_1 = x_1, \ldots, X_N = x_N)$.
Marginal Distributions

- Suppose we have a joint distribution $P(X = x, Y = y)$.
- $P(X = x)$ is called a marginal distribution. How can we find $P(X = x)$?

$$P(X = x) = \sum_{y \in \text{Val}(Y)} P(X = x, Y = y)$$

$$= \sum_{y_1 \in \text{Val}(Y_1)} \cdots \sum_{y_M \in \text{Val}(Y_M)} P(X_1 = x_1, \ldots, X_N = x_N, Y_1 = y_1, \ldots, Y_M = y_M)$$

Marginal Distributions: Heart Disease Example

Given a joint distribution on $G, BP, C, HD$, we obtain the marginal probability $P(G = M, BP = H, C = H)$ as follows:

$$P(G = M, BP = H, C = H) = \sum_{h \in \{Y, N\}} P(G = M, BP = H, C = H, HD = h)$$

$$= P(G = M, BP = H, C = H, HD = Y)$$

$$+ P(G = M, BP = H, C = H, HD = N)$$

$$= 0.050 + 0.005$$

<table>
<thead>
<tr>
<th>Gender</th>
<th>BloodPressure</th>
<th>Cholesterol</th>
<th>HeartDisease</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>H</td>
<td>H</td>
<td>Y</td>
<td>0.050</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>H</td>
<td>N</td>
<td>0.005</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>M</td>
<td>Y</td>
<td>0.045</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>M</td>
<td>N</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chain Rule

- **Joint distributions are useful because we can use them to answer queries like “What is the probability that Y = y given that I observed X = x?”**

\[
P(Y = y | X = x) = \frac{P(X = x, Y = y)}{P(X = x)}
\]

- **Write \( p(y|x) \) to denote the PMF of Y given X = x**

Conditional Distributions: Heart Disease Example

\[
P(HD = Y | G = M, BP = H, C = H) = \frac{P(G = M, BP = H, C = H, HD = Y)}{P(G = M, BP = H, C = H)}
\]

\[
= \frac{0.050}{0.050 + 0.005} = 0.91
\]

<table>
<thead>
<tr>
<th>Gender</th>
<th>BloodPressure</th>
<th>Cholesterol</th>
<th>HeartDisease</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>H</td>
<td>H</td>
<td>Y</td>
<td>0.050</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>H</td>
<td>N</td>
<td>0.005</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>M</td>
<td>Y</td>
<td>0.045</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>M</td>
<td>N</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Chain Rule: Heart Disease Example

- **By rearranging the definition of conditional probability, we get the chain rule:**

\[
p(x, y) = p(y|x)p(x)
\]

- **Applying the chain rule repeatedly to a random vector X gives:**

\[
p(x) = p(x_N | x_1, ..., x_{N-1})p(x_1, ..., x_{N-1})
\]

\[
= p(x_N | x_1, ..., x_{N-1})p(x_{N-1} | x_1, ..., x_{N-2})p(x_3 | x_2, x_1)p(x_2 | x_1)p(x_1)
\]
Bayes’ Rule

- By using the definition of conditional probability twice, we obtain one of the most important equations in probability theory, Bayes’ Rule:

\[ p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \]

- Bayes’ rule lets us compute \( p(x|y) \) from a joint distribution specified by \( p(x) \) and \( p(y|x) \).

Probabilistic Models

The solution to the problem of exponential-sized joint distributions is the use of compact probabilistic models.

- Bayesian networks achieve compactness by exploiting the chain rule and asserting (conditional) independence relations.

- As a result, Bayesian networks can express high-dimensional distributions as products of simpler factors.

Conditional Independence

\[ X \perp Y \iff p(x,y) = p(x)p(y) \]

\[ X \perp Y \iff p(x|y) = p(x) \]

\[ X \perp Y \iff p(y|x) = p(y) \]
Card Example I

Draw a random card: is value \( \perp \) color? Yes

Card Example II

What about with this deck? Is value \( \perp \) color? No

Card Example III

Is value \( \perp \) color | facecard? Yes

Conditional Independence

\[
X \perp Y | Z \iff p(y, x|z) = p(x|z)p(y|z)
\]

\[
X \perp Y | Z \iff p(x|y, z) = p(x|z)
\]

\[
X \perp Y | Z \iff p(y|x, z) = p(y|z)
\]
Compactness from Independence

Suppose we have a joint distribution \( p(a, b, c) \) and we know that the independence relation \( C \perp A | B \) holds. How can we exploit this fact to simplify \( p(a, b, c) \)?

\[
p(a, b, c) = p(a)p(b|a)p(c|a, b) \quad \text{chain rule}
\]

\[
= p(a)p(b|a)p(c|b) \quad \text{conditional independence}
\]

Bayesian Networks: Main Idea

- The main idea of Bayesian networks is conceptually simple:
  1. Order the variables and apply the chain rule
  2. Drop some dependencies, which corresponds to conditional independence assumptions

- **Example**: variables \( G, C, H, D, CP \), assume: (1) \( G \perp C \), (2) \( CP \perp G, C | H D \)
Bayesian Networks: Nodes

Formally, a Bayesian network consists of a directed acyclic graph (DAG) \( \mathcal{G} \) and a joint distribution \( p(\mathbf{x}) = p(x_1, \ldots, x_N) \) for random variables \( X_1, \ldots, X_N \).

The vertex set \( V \) has one node \( i \) for each random variable \( X_i \).

**Warning:** it’s also common to use the random variable itself, i.e., \( X_i \) as the node.

Example:

\[
\begin{align*}
\text{Gender} & \quad \text{Cholesterol} & \quad \text{BloodPressure} \\
\end{align*}
\]

\[
\text{HeartDisease}
\]

Bayesian Networks: Edges

The DAG constraint means that \( \mathcal{G} \) can’t contain any directed cycles \( i \rightarrow j \rightarrow \cdots \rightarrow i \).

Example:

\[
\begin{align*}
\text{Gender} & \quad \text{Cholesterol} & \quad \text{BloodPressure} \\
\text{HeartDisease} & \\
\end{align*}
\]

Not a valid DAG

Directed Cycle

Example:

\[
\begin{align*}
\text{Gender} & \quad \text{Cholesterol} & \quad \text{BloodPressure} \\
\text{HeartDisease} & \\
\end{align*}
\]

A valid DAG.

No directed cycle

Bayesian Networks: Parents/Children

If there is a directed edge \( i \rightarrow j \):

- \( i \) is a parent of \( j \)
- \( j \) is a child of \( i \)
- (sometimes: \( X_i \) is a parent of \( X_j \), and so on)

Define

- \( \text{pa}(i) = \) set of all parents of \( i \)
- \( \text{ch}(i) = \) set of all children of \( i \)

Example:

\[
\begin{align*}
\text{Gender} & \quad \text{Cholesterol} & \quad \text{BloodPressure} & \quad \text{Irritants} \\
\text{HeartDisease} & \\
\text{ChestPain} & \quad \text{Asthma} & \quad \text{Shortness} & \quad \text{OfBreath} \\
\end{align*}
\]

\[
\text{pa}(CP) = \{HD, A\}
\]

\[
\text{ch}(A) = \{CP, SB\}
\]

Bayesian Networks: Descendants/Non-Descendants

If there is a directed path from \( i \) to \( j \):

- \( j \) is a descendant of \( i \)
- Else \( j \) is a non-descendant of \( i \).

Define

- \( \text{de}(i) = \) set of all descendants of \( i \)
- \( \text{nd}(i) = \) set of all non-descendants of \( i \)

Example:

\[
\begin{align*}
\text{Gender} & \quad \text{Cholesterol} & \quad \text{BloodPressure} & \quad \text{Irritants} \\
\text{HeartDisease} & \\
\text{ChestPain} & \quad \text{Asthma} & \quad \text{Shortness} & \quad \text{OfBreath} \\
\end{align*}
\]

\[
\text{de}(I) = \{A, SB, CP\}
\]

\[
\text{nd}(BP) = \{G, C, I, A, SB\}
\]
Bayesian Networks: Joint Distribution

The joint distribution implied by a Bayesian network is factorized into a product of local conditional probability distributions.

\[ p(x) = \prod_{i=1}^{N} p(x_i | x_{pa(i)}) \]

The joint distribution is the product of the conditional distributions:

\[ p(x) = \prod_{i=1}^{N} p(x_i | x_{pa(i)}) \]

Figure 1: image

Bayesian Networks: CPT Example

<table>
<thead>
<tr>
<th>hd</th>
<th>g</th>
<th>bp</th>
<th>ch</th>
<th>p(hd,g,bp,ch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>M</td>
<td>Low</td>
<td>Low</td>
<td>0.95</td>
</tr>
<tr>
<td>Yes</td>
<td>M</td>
<td>Low</td>
<td>Low</td>
<td>0.05</td>
</tr>
<tr>
<td>No</td>
<td>F</td>
<td>Low</td>
<td>Low</td>
<td>0.99</td>
</tr>
<tr>
<td>Yes</td>
<td>F</td>
<td>Low</td>
<td>Low</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Bayesian Networks: CPDs and CPTs

- The individual factors \( p(x_i | x_{pa(i)}) \) in a Bayesian network are referred to as conditional probability distributions or CPDs.
- The CPD for node \( i \) must specify the probability that \( X_i \) takes any value \( x_i \) in its domain when conditioned on each joint assignment \( x_{pa(i)} \) of its parents.
- For discrete random variables, we can represent the CPD of each node using a look-up table called a conditional probability table or CPT.

Bayesian Networks: Storage Complexity

- What is the minimum amount of space needed to store the probability distribution for a single discrete random variable that takes \( V \) values? \( V - 1 \)
- How much space does it take to store the CPT for a binary-valued variable with \( D \) binary-valued parents? \( 2^D \)
- Suppose there are \( D \) binary variables connected in a chain \( X_1 \to X_2 \to \ldots \to X_D \). What is the total storage cost? \( 1 + 2(D - 1) \)
- How large is the full joint? \( 2^D - 1 \)
Next time, we'll discuss factorization and conditional independence in Bayesian networks.