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)

Review

Discrete Distributions

- Sample space Ω
- 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 $X : \Omega \to \text{Val}(X)$ 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 Ω 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**

Joint Distributions

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.

Gender	Blood Pressure	Cholesterol	Heart Disease
Male	Med	Low	No
Male	Hi	Hi	Yes
Male	Med	Med	Yes
Male	Med	Hi	No
Female	Med	Low	No
Male	Low	Med	No

Joint Probability Distributions

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

$$p(x_1, \dots, x_N) = P(X_1 = x_1, \dots, X_N = x_N)$$

Joint Distributions: Heart Disease Example

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

Gender	BloodPressure	Cholesterol	HeartDisease	P
F	L	L	N	0.0127
F	L	L	Y	0.0007
F	L	M	N	0.0098
F	L	M	Y	0.0009
F	L	H	N	0.0087
F	L	H	Y	0.0010
...	...	...	...	...

Random Vectors

- It's convenient to use vector-valued random variables $\mathbf{X} = (X_1, ..., X_N)$ (or "random vectors") and assignments $\mathbf{x} = (x_1, ..., x_N)$:

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, ..., X_N = x_N)$$

- The PMF is $p_{\mathbf{X}}(\mathbf{x})$ or just $p(\mathbf{x})$
- This is just notation: it means the same thing as a joint distribution over $(X_1, \ldots, X_N)$
- **Notation**: use $\mathbf{X}_{-i}$ and $\mathbf{x}_{-i}$ for vectors excluding X_i or x_i

Rules of Probability

Marginal Distributions

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

$$P(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{y} \in \mathrm{Val}(\mathbf{Y})} P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$$
$$= \sum_{y_1 \in \mathrm{Val}(Y_1)} \cdots \sum_{y_M \in \mathrm{Val}(Y_M)} P(X_1 = x_1, ..., X_N = x_N, Y_1 = y_1, ..., Y_M = y_M)$$

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:

$$P(G = M, BP = H, C = H) = \sum_{h \in \{Y,N\}} P(G = M, BP = H, C = H, HD = h)$$
$$= P(G = M, BP = H, C = H, HD = Y)$$
$$+ P(G = M, BP = H, C = H, HD = N)$$
$$= 0.050 + 0.005$$

Gender	BloodPressure	Cholesterol	HeartDisease	P
M	H	H	Y	0.050
M	H	H	N	0.005
M	H	M	Y	0.045
M	H	M	N	0.008
...	...	...	...	...

Conditional Distributions

- Joint distributions are useful because we can use them to answer queries like "What is the probability that $\mathbf{Y} = \mathbf{y}$ given that I observed $\mathbf{X} = \mathbf{x}$?":

$$P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})}{P(\mathbf{X} = \mathbf{x})}$$

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

- Write $p(\mathbf{y}|\mathbf{x})$ to denote the PMF of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$

Conditional Distributions: Heart Disease Example

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

$$= \frac{0.050}{0.050 + 0.005} = 0.91$$

Gender	BloodPressure	Cholesterol	HeartDisease	P
M	H	H	Y	0.050
M	H	H	N	0.005
M	H	M	Y	0.045
M	H	M	N	0.008
...	...	...	...	...

Chain Rule

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

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

- Applying the chain rule repeatedly to a random vector $\mathbf{X}$ gives:

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

$$\vdots$$

$$= p(x_N | x_1, ..., x_{N-1}) p(x_{N-1} | x_1, ..., x_{N-2}) \cdots p(x_3 | x_2, x_1) p(x_2 | x_1) p(x_1)$$

Chain Rule: Heart Disease Example

We can apply the chain rule using any ordering of the variables:

$$p(g, bp, c, hd) = p(hd | c, bp, g) p(c | bp, g) p(bp | g) p(g)$$

$$p(g, bp, c, hd) = p(g | bp, c, hd) p(bp | c, hd) p(c | hd) p(hd)$$

$$p(g, bp, c, hd) = p(c | hd, g, bp) p(hd | g, bp) p(g | bp) p(bp)$$

Bayes' Rule

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

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$

- Bayes' rule lets us compute $p(\mathbf{x}|\mathbf{y})$ from a joint distribution specified by $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$.

Conditional Independence

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.

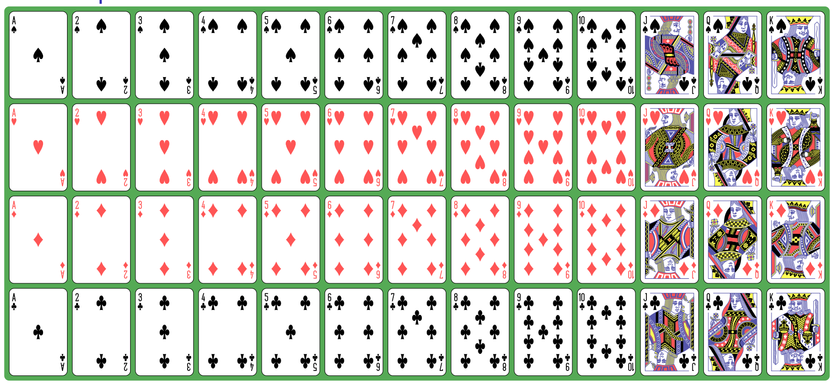
Marginal Independence

$$\mathbf{X} \perp \mathbf{Y} \iff p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

$$\mathbf{X} \perp \mathbf{Y} \iff p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x})$$

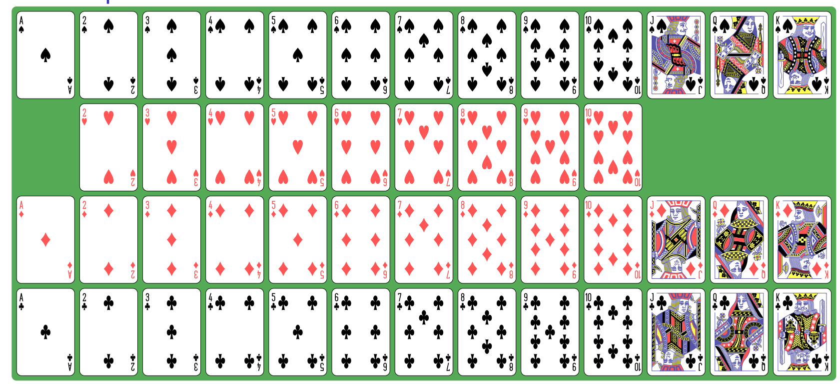
$$\mathbf{X} \perp \mathbf{Y} \iff p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})$$

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

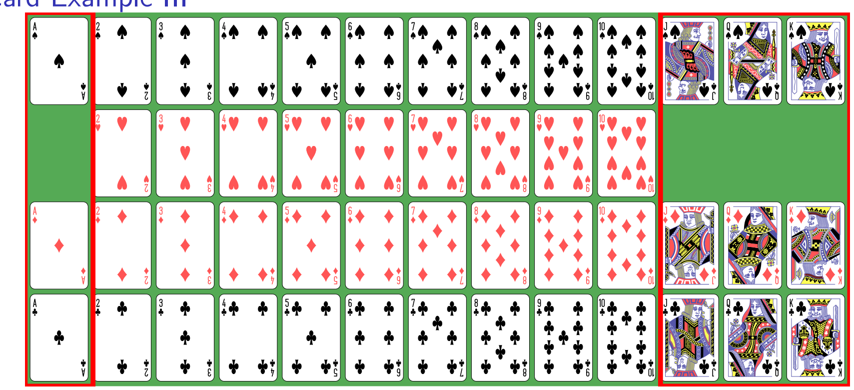
Conditional Independence

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \iff p(\mathbf{y}, \mathbf{x} | \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) p(\mathbf{y}|\mathbf{z})$$

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \iff p(\mathbf{x} | \mathbf{y}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})$$

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \iff p(\mathbf{y} | \mathbf{x}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z})$$

Card Example III



Is value $\perp$ color | facecard? Yes

Next Time

Next time, we'll discuss factorization and conditional independence in Bayesian networks.