HGEN 48600/STAT 35450: Fundamentals of Computational Biology

Lecture 2 — January 7

Lecturer: John Novembre *Scribe: Matt Bonakdarpour*

Note: These lecture notes are still rough, and have only have been mildly proofread.

2.1 Graphical Models: Introduction

Quite often, we are confronted with the task of understanding a complex system with many dependent components. In this context, probability theory gives us a framework to reason about a collection of possible outcomes and their associated likelihoods. For example, if we are trying to diagnose a patient, there are many potential diseases this patient could have. Also, associated with each patient, we may have hundreds of data points related to their symptoms, personal traits, and diagnostic tests. Each of these characteristics could be thought of as a random variable. Our goal is to reason about this patient given the values of one or more of these random variables. In the framework of probabilistic reasoning, we need to construct a joint distribution over the space of possible assignments to this set of random variables.

If we have five realizations of *binary* random variables for a patient, we must specify a joint distribution over 2^5 possible values – a potentially daunting task! Graphical models provide a framework to compactly encode a joint distribution in high dimensions. They also provide other benefits, including but not limited to:

- A tool for visualizing the structure of a probabilistic model
- Provides insights, such as conditional independence properties, by inspection of the graph
- Complex computations can be expressed as operations on the graph

In this lecture, we will consider directed graphical models, sometimes known as Bayesian networks.

2.1.1 Joint probabilities and independence

Using the definition of conditional probability, we have that the joint distribution between X_i and X_j is:

$$p(X_i, X_j) = P(X_i|X_j)P(X_j)$$

In general, if we have any n random variables $X_1, \ldots, X_n$, we can use the *chain rule* to factor the joint distribution as follows:

$$p(X_1, \ldots, X_n) = p(X_1)p(X_2|X_1)p(X_3|X_1, X_2) \cdots p(X_n|X_1, \ldots, X_{n-1})$$

Recall that X_i is independent of X_j if knowledge of X_j does not change our knowledge of X_i. That is, $p(X_i|X_j) = p(X_i)$. Plugging this in to the expression for the joint distribution above, we have that X_i is independent of X_j if and only if $p(X_i, X_j) = p(X_i)p(X_j)$.
We say that X_i is *conditionally independent* of X_j given X_k if

$$p(X_i, X_j|X_k) = p(X_i|X_k)p(X_j|X_k)$$

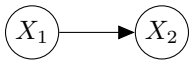
which means

$$p(X_i, X_j, X_k) = p(X_i|X_k)p(X_j|X_k)p(X_k)$$

Notice how the joint distribution factorizes nicely when we have conditional independence. Idea: when faced with a large number of features, use our prior knowledge of independence to simplify the joint distribution.

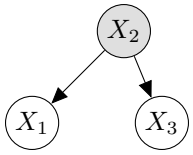
2.1.2 Some basics

A directed graph is a set of vertices (or nodes) along with a set of directed edges (or arrows) between nodes. Here is a simple graph with two vertices and one directed edge:



Here, each vertex represents a random variable. When a specific random variable in the graph is observed, we shade that particular vertex. For example, this is what the graph above would look like if we observed X_1, but X_2 was still latent:

A directed graphical model encodes conditional independence in the following way. A specific random variable (or vertex) is conditionally independent of all other nodes, given its parents (i.e. all nodes that points to it). For example, in the following graph X_1 is independent of X_3 given X_2:



The general form for the joint distribution for a directed graphical model is therefore:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\text{pa}_i) \tag{2.1}$$

where pa_i denotes the parents of X_i.

2.1.3 A motivating example

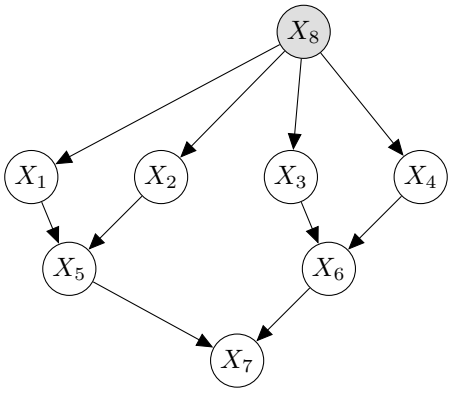
Assume we have 3 possible genotypes {AA,Aa,aa}. To each person, we will assign a random variable X_i according to their genotype:

Genotype	X_i
AA	0
Aa	1
aa	2

Let's also assume that the genotype of each person is binomially distributed, so that:

$$\begin{aligned} X_i|p &\sim \text{Bin}(2, p) \\ p(X_i = 2|p) &= p^2 \\ p(X_i = 1|p) &= 2p(1-p) \\ p(X_i = 0)|p) &= (1-p)^2 \end{aligned}$$

In this context, we represent inheritance in a 3-generation pedigree with the following graphical model. In what follows, we're assuming $X_8 = p$.



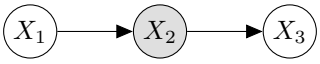
Assume X_8 is observed and that we are interested in $p(X_5|X_1, X_2)$. We could represent this conditional probability as a table, enumerating all possibly quantities that X_1, X_2 and X_5 take.

From the graphical model above, and using equation 2.1, we can factorize the joint distribution in the following way:

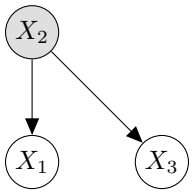
$$p(X_1, \ldots, X_8) = p(X_1|X_8)p(X_2|X_8)p(X_3|X_8)p(X_4|X_8)\times \\ p(X_5|X_1, X_2)p(X_6|X_3, X_4)p(X_7|X_5, X_6)p(X_8)$$

2.1.4 Some classes of 3-node graphs

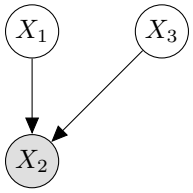
With the linear chain graph, we have that $X_1 \perp X_3|X_2$:



As we saw in a previous section, in this multiple offspring graph, we have that $X_1 \perp X_3|X_2$:



The v-structure graph has the following structure:



It is *not true* that $X_1 \perp X_3|X_2$. For example, assume X_2 was the event that your house alarm was going off, X_1 was the event that there was an earthquake, and X_3 was the event that there was a burglar in your house. Assuming your alarm was going off, the additional information about whether a burglar was in your house would affect your knowledge of whether or not an earthquake was happening.

2.1.5 D-separation

Take any undirected path (ignoring arrows) in the graph G. This path is called an *active trail* for observed variables $O \subset \{X_1, \ldots, X_n\}$, if for every consecutive triple of variables X, Y, Z on the path:

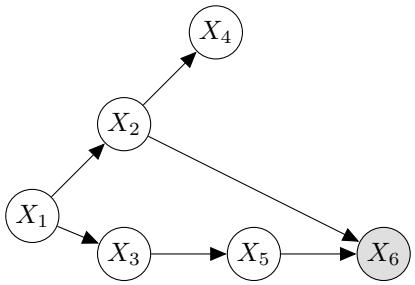
- $X \to Y \to Z$ and $Y \notin O$

- $X \leftarrow Y \leftarrow Z$ and $Y \notin O$
- $X \leftarrow Y \rightarrow Z$ and $Y \notin O$
- $X \rightarrow Y \leftarrow Z$ and Y or any of Y's descendants are in O

Any two variables X_i and X_j for which there does not exist an active trail for observations O are called *d-separated* by O, written d-sep$(X_i; X_j|O)$. Two sets of vertices A and B are d-separated by O if d-sep$(X, Y|O)$ for all $X \in A$ and $Y \in B$.
The key result is the following: if X and Y are d-separated by a set O, then $X \perp Y|O$. In words, X is conditionally independent of Y given O if there does not exist any active trail between X and Y for observations O. Intuition: active trails allow the dependencies to flow.

2.1.6 Elimination Algorithm

Assume we have a graph G, and we take two disjoint subsets of nodes X_E and X_Q. Our goal is to calculate $p(X_Q|X_E)$. In this section, we will focus on the case when X_Q is one node called the *query node*, and the set of nodes X_E are called *evidence nodes*.
Assume we have the following graph:



Let our evidence set be $\{X_6\}$ and our query set be $\{X_1\}$. We want to compute:

$$p(X_1|X_6) = \frac{p(X_1, X_6)}{p(X_6)} = \frac{p(X_1, X_6)}{\sum_{x_1} p(X_1 = x_1, X_6)}$$

The elimination algorithm provides an effective computational method for making these calculations.
From the expression above, it seems that to compute the conditional distribution, we just need to be able to compute the joint distribution $p(X_1, X_6)$. We could start off by marginalizing the full joint distribution for all variables:

$$p(X_1, X_6) = \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} p(X_1, X_2, \ldots, X_6)$$

In this case,if every X_i can take on one of k values, this expression would have k^4 terms. However, using the conditional independence relations from the graph, we get:

$$\begin{aligned} p(X_1, X_6) &= \sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2)p(X_5|X_3)p(X_6|X_2, X_5) \\ &= p(X_1)p\sum_{x_2}(X_2|X_1)\sum_{x_3} p(X_3|X_1)\sum_{x_4} p(X_4|X_2)\sum_{x_5} p(X_5|X_3)p(X_6|X_2, X_5) \end{aligned}$$

We can further simplify this summation by introducing the following notation. Let $m_i(x_{S_i})$ denote the expression that arises when performing the sum $\sum_{x_i}$ where x_{S_i} are the variables, other than x_i that appear in the summand. Employing this notation, we get:

$$\begin{aligned} p(X_1, X_6) &= p(X_1)p\sum_{x_2}(X_2|X_1)\sum_{x_3} p(X_3|X_1)\underbrace{\sum_{x_4} p(X_4|X_2)}_{m_4(x_2)}\underbrace{\sum_{x_5} p(X_5|X_3)p(X_6|X_2, X_5)}_{m_5(x_2,x_3)} \\ &= p(X_1)p\sum_{x_2}(X_2|X_1)m_4(x_2)\underbrace{\sum_{x_3} p(X_3|X_1)m_5(x_2, x_3)}_{m_3(x_1,x_2)} = p(X_1)p\sum_{x_2}(X_2|X_1)m_4(x_2)m_3(x_1, x_2) \\ &= p(x_1)m_2(x_1) \end{aligned}$$

Using this result, we get the desired conditional probability:

$$\Rightarrow p(X_1|X_6) = \frac{p(x_1)m_2(x_1)}{\sum_{x_1} p(x_1)m_2(x_1)}$$

This exercise gives rise to a general algorithm for computing marginal probabilities. To see the details of this algorithm, refer to the handout given in class entitled *The Elimination Algorithm.*
Relevant online course notes:

- `http://www.inf.ed.ac.uk/teaching/courses/pmr/slides/elim-2x2.pdf`
- `https://www.cs.cmu.edu/~aarti/Class/10701/readings/graphical_model_Jordan.pdf`
- `http://www.cs.columbia.edu/~blei/fogm/2015F/notes/inference.pdf`
- `http://www.cs.berkeley.edu/~jordan/courses/281A-fall04/`

For drawing graphical models

- `https://github.com/jluttine/tikz-bayesnet`