36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 30: November 16

Lecturer: Siva Balakrishnan

30.1 Review and Outline

In the last class we discussed causal inference:

1. The population average treatment effect and Neyman's null
2. Testing Neyman's null by estimating the variance of the population average treatment effect estimator
3. Causal inference in observational studies and selection bias
4. Correcting for selection bias by correcting for confounders

Today we will discuss directed graphical models. This is Chapter 17 of the Wasserman book.

Broadly, graphical models are a concise, graphical way of representing structured probability distributions. They are widely used in a variety of applications in machine learning and statistics, and there are entire courses devoted to the topic.

Our modest goal is to use a couple of lectures to get a sense for what the semantics of graphical models are and what are the main statistical questions of interest.

30.2 Parameter counting

Suppose we have a collection of d random variables: $X_1, \ldots, X_d$, each of which takes values in $\{1, \ldots, k\}$, and suppose we want to understand their joint distribution, i.e. I want to represent $P(X_1 = i_1, \ldots, X_d = i_d)$ for all possible values. How many parameters do I need?

The answer is $k^d - 1$. This is a gigantic number even for moderate sizes of k and d and this can be quite problematic.

Now, suppose I tell you that all the random variables are independent? You can then see that $d(k-1)$ parameters suffice to represent the entire distribution.

This is the basic idea behind graphical models: I want to represent the joint probability distribution over many random variables, but I want to do so compactly by explicitly modeling the dependence structure between the random variables.

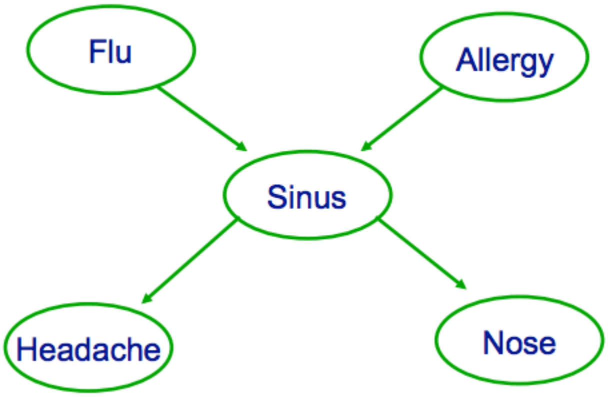
Lets see a slightly more interesting example: suppose that I have a system with three variables: "SAT score (S)", "High-School Grade (H)", "Intelligence (I)". Lets suppose these are binary variables.

A natural idea would be that both S and H are determined by I, and are *conditionally independent* given I. In this case, we can see that the joint probability:

$$\begin{aligned} P(S,H,I) &= P(I)P(S|I)P(H|S,I) \\ &= P(I)P(S|I)P(H|I), \end{aligned}$$

using the conditional independence. Now, again we can count the number of parameters needed to represent the distribution: it is $1+2+2=5$. With no independence assumptions we would need 7 parameters, while with full independence we would need only 3 parameters.

Here is another example of a simple Bayes net:



A Bayes net (or a directed graphical model) is a directed acyclic graph on a collection of random variables. It is a graphical way of representing a distribution that factorizes as:

$$P(X_1,\ldots,X_d) = \prod_{i=1}^{d} P(X_i|\text{Pa}(X_i)),$$

so in the above example, we are claiming that the joint distribution factorizes as:

$$P(F,S,H,N,A) = P(F)P(A)P(S|F,A)P(H|S)P(N|S).$$

Again supposing everything is binary we can count the number of parameters as: $1+1+4+2+2=10$ which is much less than the full 31 parameters. Just intuitively, the fact that there is a concise representation must mean that the distribution has many independence assumptions encoded in it.

30.3 Independencies

The most interesting fact about directed graphical models is that we can "read off" all the independence assumptions, and conditional independence assumptions.

Lets start off with independence assumptions.

In the flu example I claim that F and A are independent. Lets see how one might verify my claim:

$$\begin{aligned} P(F,A) &= \sum_{H,S,N} P(F,S,H,N,A) = P(F)P(A) \sum_{H,S,N} P(S|F,A)P(H|S)P(N|S) \\ &= P(F)P(A) \sum_{S} \left[P(S|F,A) \left[\sum_{H} P(H|S) \right] \left[\sum_{N} P(N|S) \right] \right] \\ &= P(F)P(A), \end{aligned}$$

so we can see that $F \perp\!\!\!\perp A$. There are actually no other marginal independencies implied by the graph. Lets try another pair: suppose we tried F and S:

$$\begin{aligned} P(F,S) &= \sum_{H,N,A} P(F,S,H,N,A) = P(F) \sum_{H,N,A} P(A)P(S|F,A)P(H|S)P(N|S) \\ &= P(F) \left[\sum_{H,N} \left[\sum_{A} P(S,A|F) \right] P(H|S)P(N|S) \right] \\ &= P(F)P(S|F) \left[\sum_{H,N} P(H|S)P(N|S) \right] = P(F)P(S|F), \end{aligned}$$

which does not tell us that they are independent.

The general rule is that in a graphical model, two variables are marginally independent if there is no directed path between them.

So we can see that $F \perp\!\!\!\perp A$ is the only marginal independence in our example. However, intuitively it seems like there are several other independence assumptions being encoded by the graph, since we have many fewer parameters than we would have if we added extra edges between $\{F, A\}$ and $\{H, N\}$.

The remaining independencies are actually conditional independencies. The graphical rule is: "variables are independent of their non-descendents conditioned on their parents". This is often called the local Markov property.

This graphical rule is equivalent to the factorization property. Lets verify one direction, i.e. that the local Markov property $\implies$ the factorization property, for the flu example.

We know by the chain rule:

$$P(F,A,S,H,N) = P(F)P(A|F)P(S|F,A)P(H|S,F,A)P(N|H,S,F,A).$$

Now you can apply the above rule to each of the terms and see that:

$$P(F, A, S, H, N) = P(F)P(A)P(S|F, A)P(H|S)P(N|S),$$

which is just the factorization property.