CS281A/Stat241A: Statistical Learning Theory

Lecture 3: Conditional independence

Lecturer: Michael I. Jordan *Scribes: Beau Cronin, Alex Fabrikant, and Brighten Godfrey*

1 Overview

In this lecture we describe and compare which joint probability distributions can be modeled by undirected and by directed graphs. We also discuss conditional independence, parameterization, and uses of undirected graphs.

2 Which joint probability distributions can be modeled

2.1 Directed graphs

There are two equivalent representations for possible joint probability distributions associated with a directed graph:

Parametric: Those distributions which factor into the conditional distribution of each random variable given its parents. We'll call this set of distributions D_1:

$$D_1 = \left\{ p(x) : p(x) = \prod_{i=1}^{n} p(x_i | x_{\pi_i}) \right\}$$

Qualitative: Those distributions which meet all conditional independence statements associated with the graph. We'll call this set of distributions D_2.

A theorem, which we will prove later, states that $D_1 = D_2$.

Note that every possible joint distribution is allowed for a fully connected graph. For completely disconnected graphs, all variables are pairwise independent.

2.2 Undirected graphs

As there were for directed graphs, there are also two possible representations for the joint distributions associated with an undirected graph. Define:

$$\begin{aligned} U_1 &= \left\{ p(x) : p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Psi_C(x_C) \right\} \\ U_2 &= \{ p(x) : p(x) \text{ meets all conditional independence statements in the graph} \} \end{aligned}$$

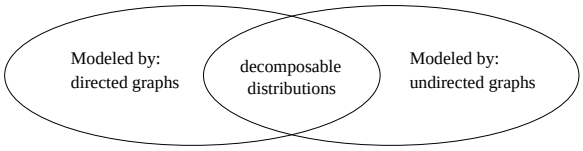
Also as in the directed case, a theorem [Hammersley-Clifford '70] states that $U_1 = U_2$.

2.3 Comparison

Contrary to what one might expect, directed graphs can't model everything that undirected graphs can model. Each particular graph G can model a family of distributions $F(G)$. (The particular distribution is selected by a parameterization of the graph.) Some undirected graphs can model families of distributions which can't be modeled by directed graphs, and vice versa. That is, if we let U be the set of all undirected graphs, and D be the set of all directed graphs:

$$\exists G_1 \in U. \forall G_2 \in D. F(G_1) \neq F(G_2)$$
$$\exists G_1 \in D. \forall G_2 \in U. F(G_1) \neq F(G_2)$$

Those families of distributions which are representable by both types of graphs are called *decomposable*. An example of a decomposable family is the family of all possible distributions. This can be modeled by a clique in either type of graph. The space of distributions thus looks like this:



3 Conditional independence in undirected graphs

Conditional independence in undirected graphs is simpler than in directed graphs. In an undirected graph, given node sets A, B and C, A is independent of B given C if C separates A from B in the graph. Put another way, all influence between A and B is mediated by C.

4 Parameterization of undirected graphs

We parameterize an undirected graph G with one potential function on each maximal clique of G. (It will be explained later explained why only the maximal cliques are needed.)

Let $\mathcal{C}$ be the set of all maximal cliques in G. Let $X_{\mathcal{C}}$ be the subset of random variables indexed by $\mathcal{C}$. For a particular maximal clique C, x_C is a specific configuration of X_C. A potential function $\Psi_C(x_C)$ maps these configurations to positive reals.

We compute the joint probability distribution of the nodes of G as the normalized product of the potential functions for all the maximal cliques. That is,

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Psi_C(x_C).$$

Z is a normalization factor which ensures that we get a valid probability distribution (one which sums to 1 over all possible configurations x):

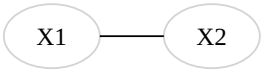
$$Z = \sum_x \prod_{C \in \mathcal{C}} \Psi_C(x_C)$$

Potential functions cannot be defined arbitrarily over sets which are not cliques, since it would then be possible to have a potential function which did not respect the independencies dictated by the graph structure. Hence, cliques are important.

For convenience of notation, we usually drop the subscript C from our potential function $\Psi_C(x_C)$, writing it simply as $\Psi(x_C)$. This is unambiguous since the potential function can be uniquely identified by its arguments.

4.1 Why we only need to consider maximal cliques

Consider, for example, the following graph G:



Using only maximal cliques, we get

$$p(x) = \frac{1}{Z}\Psi(x_1, x_2)$$

If we use all cliques, we include the singleton cliques $\{X_1\}$ and $\{X_2\}$, so we have

$$p'(x) = \frac{1}{Z}\Psi'(x_1, x_2)\Psi'(x_1)\Psi'(x_2)$$

But $p(x)$ can be made to mimic $p'(x)$ simply by defining $\Psi(x_1, x_2)$ as

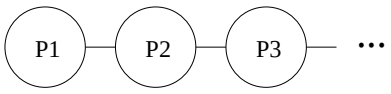
$$\Psi(x_1, x_2) = \Psi'(x_1, x_2)\Psi'(x_1)\Psi'(x_2)$$

This example can be easily extended to show that for any graph, for any parameterization using potential functions on all cliques, there is an equivalent parameterization using potential functions only on maximal cliques.

5 Uses of undirected graphs

Some problems are more naturally modeled by undirected graphs than by directed graphs.

An example is in computer vision: detecting edges in an image. We can construct one node for each pixel, and put edges between nodes representing adjacent pixels (thus forming a lattice). Potential functions for this lattice would be defined so as to return higher values for those mutual orientations which indicate an edge which "passes through" these neighboring nodes. Undirected graphs are a natural choice here as opposed to directed graphs, because there is no direct causality relationship – that is, a parent/child relationship between pixels does not make sense – yet there is a correlation between pixels.



Another example is in part-of-speech tagging in natural language processing. The undirected graph above may have a variable P_i for the part of speech of the i'th word in a sentence. If P_2 is "noun", for instance, that would correspond to a high probability of P_1 being set to "determiner" (like the English article "the"), and a lower probability of it being set to "adverb."

Other areas where undirected graphs are used include error correcting codes and phylogeny.