SCRIBE NOTES FOR COS513 LECTURE 3

THIAGO PEREIRA

1. Introduction

In the previous lecture, we discussed how a directed graphical model represents a factorization of the joint $p(x_1, ..., x_n) = \prod_i p(x_i|x_{\pi_i})$. By inspecting this factorization, we were able to make some basic conditional independence statements about the model. However, the question of whether there are other possible conditional independence statements remained. In this lecture, we introduce the Bayes ball algorithm. It will let us find all statements associated with a directed graphical model.

An important remark is that the lack of an independence statement does not imply that the variables are dependent. It only means that they are not necessarily independent.

2. Bayes ball algorithm

2.1. **Three canonical graphs.** In this section, we study the conditional independence statements associated with three very simple graphs. Knowledge of these particular cases will enable us to build an algorithm for general undirected graphs in the next subsection, the *Bayes ball algorithm*. At each of the graphs, we are going to observe how the concepts of graph separability and conditional independence are interconnected.

Our first graph is a small chain (Figure 1). It represents the factorization $p(x, y, z) = p(x)p(y|x)p(z|y)$. From this formula, we can see there is one independence statement implied by this graph: $X \perp Z | Y$. This graph can be seen as a ”past, present and future” chain. The intuition here is that the future is independent of the past given the present. In fact, this is the only statement implied by this graph.

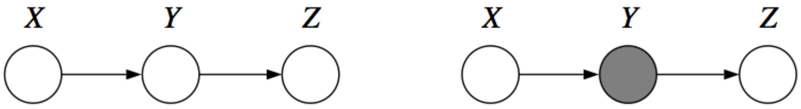

FIGURE 1. A small chain graph.

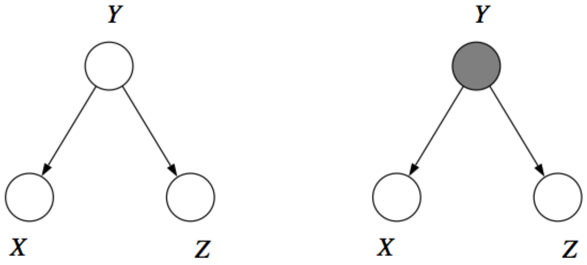

FIGURE 2. A small tree.

These results can be interpreted from a graph separability point of view. Whenever node Y is given, we remove it from the graph. By removing Y, we disconnect X and Z, i.e. there is no path connecting X and Z. In this example, separability and conditional independence agree since X and Z are independent given Y. However, if Y is not given, then we do not remove it from the graph. In which case, X and Z are still connected. This is consistent with the fact that $X \perp Z$ in general is *not* a conditional independence statement implied by the graph.

Our second graph is a small tree (Figure 2). It represents the factorization $p(x, y, z) = p(y)p(x|y)p(z|y)$. Once again, there is only one implied independence statement: $X \perp Z|Y$. Y can be interpreted as a hidden variable. In this example, X and Z may be dependent variables. However, once the hidden variable Y is given, X and Z become independent.

As an example of this tree, we may consider the random variables amount of gray hair G and shoe size S. In general, these variables are dependent, since older people tend to have more gray hair and larger shoe size at the same time. However, we can break this dependency once we condition on the random variable age A, the hidden cause of both shoe size and gray hair. This means $G \perp S|A$, but $G \perp S$ is not be true.

Once again the concept of separability and independence agree. When we remove (condition on) Y in Figure 2, X and Z cannot reach each other (they are independent).

In our final example, separability and independence no longer agree. This is why we cannot simply use naive separability to determine independence statements. In Figure 3, we see an inverse tree graphical model. It represents the factorization $p(x, y, z) = p(x)p(z)p(y|x, z)$. It can be shown, as is intuitive, that $X \perp Z$. In other words, X is marginally independent of Z. However, the statement $X \perp Z|Y$ is not necessarily true. In this example, exactly when Y is kept in the graph, they are independent, even though X and Z can reach each other.

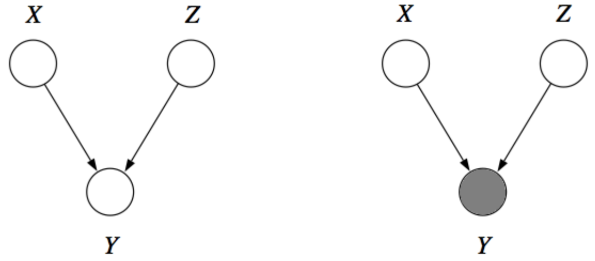

FIGURE 3. An inverse tree.

An example of this inverse tree situation occurs if we consider the following random variables: Bob's watch is broken Z, Alice does not arrive on time for lunch Y and Alice was abducted by aliens X. In this example, $X \perp Z$ means the fact that Bob's watch is broken and whether Alice was or not captured by aliens are completely independent events, as is intuitive. However, given that Alice does not arrive on time, X and Z become dependent. For example, assuming Alice late, if Bob discovers that his watch is broken, it means Alice was probably not abducted by aliens.

This last example show that the concept of separation we are interested in is different from naive graph separation. They agree on the first two examples, but disagree on the inverse tree. Therefore, we define a new concept of separation called *d-separation*. This will let us determine any conditional independence statement in a graphical model.

2.2. **Bayes ball algorithm.** We can define d-separation from an algorithmic point of view, we shall call it the *Bayes ball* algorithm. This algorithm let us determine if a conditional statement of the form $E \perp F | G$, where E, F and G are sets of random variables, necessarily holds. It consists of the following steps:

- Shade nodes that are being conditioned on.
- Place balls on one set of variables (either E or F).
- If any ball can reach a member of the other set then they are not conditionally independent, otherwise they are.

However, the propagation of the Bayes balls follows some specific rules. In general, these rules tell you if a ball can go from node X through node Y to node Z. To cover all the cases, we must consider different scenarios according to whether edges are coming in or out of Y, either to X or Z. In fact, all combinations fall in one of the three simpler graphs we have already considered. We depic these three cases, now as separability rules, in Figures 4, 5 and 6. Both in the chain and tree cases, the Bayes ball can only pass if Y is not conditioned on. The ball get blocked when conditioning on Y,

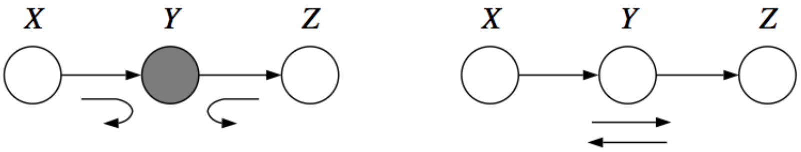

FIGURE 4. The rules in the chain case.

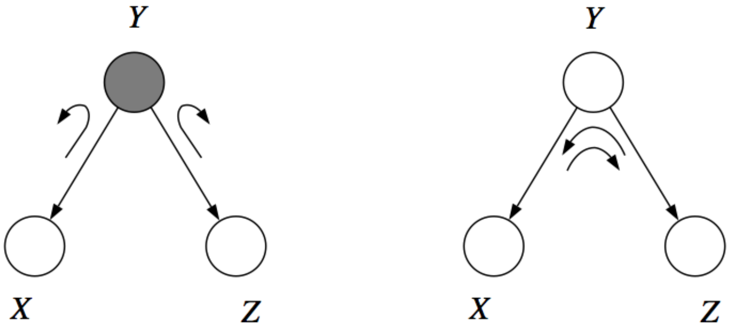

FIGURE 5. The rules in the tree case.

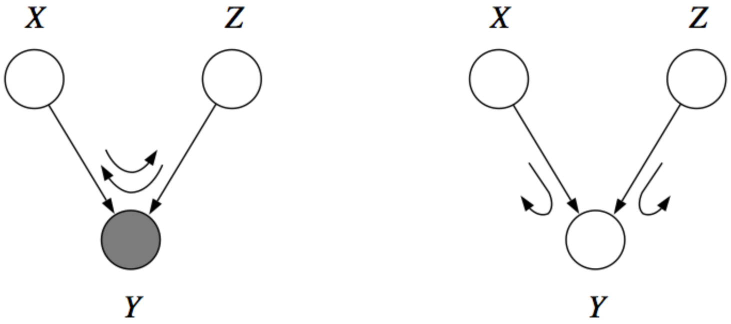

FIGURE 6. The rules in the inverse tree case.

as shown in Figures 4 and 5. The opposite happens when edges form an inverse tree. The ball can only pass when Y is shaded, as shown in Figure 6.

Besides these previous cases, we must also consider some degenerate cases (Figure 7) when the ball wants to go from X through Y back to X. This rules can be designed by considering X and Z to be the same node and applying the above presented rules. For example, Figure 7 shows the degenerate inverse tree case, both X and Z (remember that Z is the same as X here) point towards Y. In this case, if Y is shaded, it can bounce back

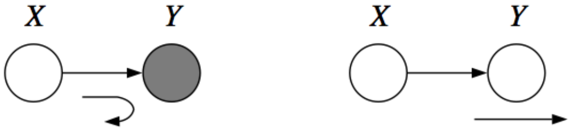

FIGURE 7. The rules in degenerate case.

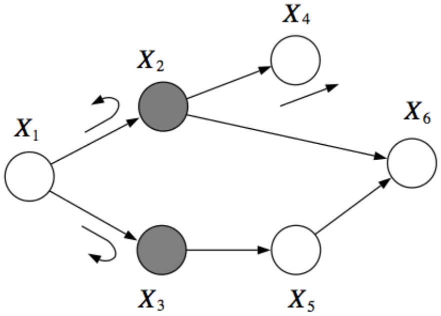

FIGURE 8. A ball trying to get from X_1 to X_6 cannot pass through either X_2 nor X_3. This is because of the rule in the case of the chain graph.

to X. If Y is not shaded, the ball stops at Y. In case the edge was oriented from Y to X, we would be in the regular tree case and the inverse of the above would happen.

Consider Figure 8 and let us use the Bayes ball algorithm to evaluate some conditional independence statements. Is $X_1 \perp X_6 | \{X_2, X_3\}$ true? If we put a Bayes ball in X_1 and it tries to reach X_6 it must pass through one shaded node. However, the configuration of edges around either X_2 or X_3 leads us to use the rule of the chain graph. As these nodes are shaded, this means the ball cannot pass. As a consequence we can state that $X_1 \perp X_6 | \{X_2, X_3\}$.

As another example, consider Figure 9. Is $X_2 \perp X_3 | \{X_1, X_6\}$ implied by the graph? The answer is no. Indeed, we can show a path between X_2 and X_3. The ball starts at X_2 and passes through X_6 arriving at X_5 using the inverse tree case. Next, it uses a simple chain path to get to X_3. Since X_2 and X_3 are not separated we cannot make the above independence statement.

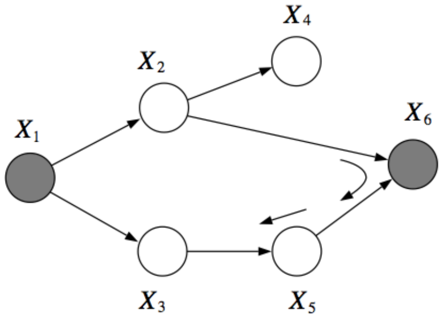

FIGURE 9. A ball can get from X_2 to X_3 first using the inverse tree case through X_6 and then using the chain case.

3. CHARACTERIZATION OF UNDIRECTED GRAPHICAL MODELS

Now that we know how to find all the independence statements associated with a graph, we can provide a new characterization of graphical models. Consider two families of propability distributions.

- Those found by ranging over all possible selections of conditional probability *distributions* associated with each node.
- All joint probability distribution that respect all conditional independence *statements* implied by a directed graphical model using d-separation.

The *Hammersley-Clifford theorem* states that these two families are the same.

Up to this point, we have only considered directed graphical models. However, it is also possible to work with undirected graphical models which we consider next.

4. UNDIRECTED GRAPHICAL MODELS

An undirected graphical model is an undirected graph together with a set of *potential functions* defined on cliques of the graph, i.e. a fully connected subgraph. These potentials are real-valued positive functions and they define a joint distribution:

$$p(x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

The set C contains cliques of the graph and it must cover all random variables and edges. Z is a normalizing constant, it ensures that p sums to 1 when ranging over all values of x.

The concept of separation in undirected graphs is much simpler. Here naive separation provides us with all necessary conditional independence statements associated with the graph. It is not necessary to use d-separation.

Finally, it is important to remark that the set of joint distributions that can be represented in an undirected graphical model is different than those from the directed graphical model.

5. The Elimination Algorithm

In this section, our objective is presenting a general algorithm for computing conditional and marginal probabilities. While, many algorithms exist for this problem, we start our study by the simplest one, known as the *Elimination Algorithm*.

Let f be a node and E be a set of nodes not containing f. Let R be the remaining nodes. In the *probabilistic inference* problem, we want to calculate $p(x_f|x_E)$. We can do it in three steps:

1) Compute marginal

$$p(x_f, x_E) = \sum_{x_R} p(x_f, x_E, x_R)$$

2) Compute

$$p(x_E) = \sum_{x_f} p(x_f, x_E)$$

3) Calculate the ratio

$$p(x_f|x_E) = \frac{p(x_f, x_E)}{p(x_E)}$$

Notice that calculating $p(x_E)$ by summing over x_f is more efficient that using the original joint distribution.