STA561: Probabilistic machine learning

Directed Graphical Models (Bayes Nets) (9/4/13)

Lecturer: Barbara Engelhardt Scribes: Richard (Fangjian) Guo, Yan Chen, Siyang Wang, Huayang Cui

1 Introduction

For a joint distribution $P(X|\theta)$, how can we compactly represent it? How can we predict a set of variables based on their distributions? How can we learn the parameters in a reasonable amount of time with a reasonable amount of data? Directed graphical models provide promising answers to these questions in machine learning. The advantages of DAG are mainly summarized as follows.

1. Compactness: We can have a small number of parameters.
2. Predictive: We can use this model to tell us about our next sample.
3. Parameter estimation: We can efficiently (in terms of the computational time) and effectively (in terms of a small number of samples) estimate the model parameters given available data.

2 Weather pattern example

Let X represent the weather every day over one month. So, X might be

$$X = \underbrace{SSRSSRRR\cdots S}_{30} \quad (S : Sunny, R : Rainy).$$

There exist 2^{30} different situations in total. The parameterization of the joint distribution of this space will have $2^{30} - 1$ values, all between zero and one, and they have to sum to less than or equal than one (i.e., we put a probability of each event on that event).

Based on the *chain rule*, we can start to factor the joint probability as

$$P(X = x \mid \theta) = P(x_1 \mid \theta)P(x_2 \mid x_1, \theta)P(x_3 \mid x_2, x_1, \theta)\cdots P(x_{30} \mid x_{29}, \cdots, x_1, \theta).$$

We can represent each of these probabilities in *conditional probability tables* or *CPTs*. The constraints of the tables are: all the entries should be between 0 and 1, and the entries in each row should sum to 1. We see that there are $O(K^V)$ parameters in the model for K variables.

3 Conditional independence

Given we have observed Z, X is conditionally independent of Y (denoted $X \perp Y|Z$), if and only if the conditional joint probability can be written as a product of conditional marginal probabilities, i.e.,

$$X \perp Y|Z \Leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z).$$

Based on this equation, we can derive the following:

$$X \perp Y|Z \Leftrightarrow P(X|Y,Z) = \frac{P(X,Y|Z)}{P(Y|Z} = P(X|Y).$$

4 The Markov Assumption

How can we use conditional independence to make our model more concise? We will see how random variables that satisfy the *Markov assumption* can be represented in a more compact form. The Markov assumption (first order) is based on the idea that the future is independent of the past given the present. In other words, it means that what happens in the future is conditionally independent of what happened in the past given what is happening now. More formally,

$$X_{t+1} \perp X_1, ...X_{t-1} \mid X_t$$

In the weather example, let us assume further that X satisfies the Markov assumption; then we are able to rewrite these conditional probabilities as

$$\begin{aligned} P(X = x \mid \theta) &= P(x_1 \mid \theta)P(x_2 \mid x_1, \theta)P(x_3 \mid x_2, \theta) \cdots P(x_{30} \mid x_{29}, \theta) \\ &= P(x_1 \mid \theta) \prod_{i=1}^{29} P(x_{i+1} \mid x_i, \theta). \end{aligned}$$

Now, we will introduce a state transition diagram (Figure 1). Suppose X follows a *stationary* Markov chain, then the transition matrix, or CPT, is independent of time t:

$$P(x_{t+1} \mid x_t) = P(x_{s+1} \mid x_s) \quad (t = 1, 2, \cdots, T-1), (s = 1, 2, \cdots, T-1).$$

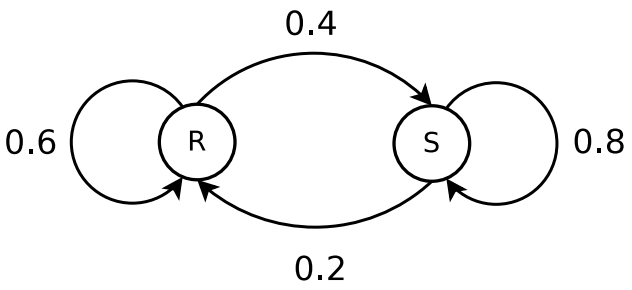

Figure 1: **State transition diagram.** This figure represents how the different states transition in the Markov chain. This does not represent a graphical model: each node is not a random variable.

The state transition diagram is a visual way to represent the probability of one state transitioning to another state (or the same state, a *self transition*) at the next time point. From the state transition diagram, we have

$$\begin{aligned} P(x_{t+1} = S \mid x_t = S) &= 0.8, \\ P(x_{t+1} = R \mid x_t = S) &= 0.2, \\ P(x_{t+1} = S \mid x_t = R) &= 0.4, \\ P(x_{t+1} = R \mid x_t = R) &= 0.6. \end{aligned}$$

Then, based on Markov assumption, if we know the state at time t (i.e., the weather for day t), it will be easy for us to compute the probability of each state at time $t+1$ (i.e., the state of the weather for day $t+1$). Furthermore, if we know the initial distribution π_0 and the state transition matrix CPT, we can compute the probability of the state of X at any time in the future by using the following expression:

$$\pi_{t+h} = \pi_t T^h \quad (t = 1, 2, \cdots, T-1), (h > 0).$$

Intuitively, as h grows, we are less and less certain about the future state given the state at time t. As $h \to \infty$, this produces the stationary distribution π_s of X, which is the marginal probability of every state, and satisfies the form:

$$\pi_s = \pi_s T.$$

Returning to our weather example, we see that the original representation that enumerates all possible states requires $(2^{30} - 1)$ parameters. Here, we have seen that we only need 2 parameters (or four for π_0 also) to characterize a stationary first order Markov chain (noting the normalization). Next, we will introduce directed graphical models as a more general way for compactly representing joint probabilities.

5 Directed Acyclic Graphs: Bayesian networks

5.1 Definitions

Directed Acyclic Graphs (DAGs) are directed graphs that do not have cycles (i.e., no node is an ancestor or descendant of itself). As a probabilistic model, every node in a DAG corresponds to a random variable, and the edges capture the conditional indepedencies (or, more precisely, the lack of conditional independencies) among the nodes. More precisely, a DAG $\mathcal{G}$ is composed of vertices (nodes) V and directed edges E. This is usually denoted as

$$\mathcal{G} = \{V, E\},$$

where $V = \{1, \cdots, m\}$ is the set of vertices, and $E = \{(s, t) : s, t \in V\}$ is the set of directed edges. The acyclic condition can be more formally stated as no node can be its own ancestor, namely

$$\forall i, i \notin anc(i).$$

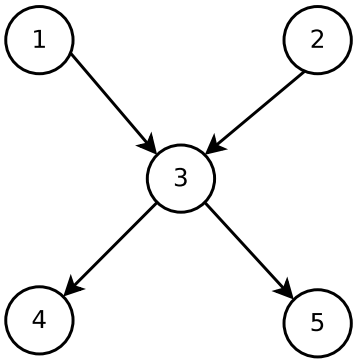

Figure 2: DAG

Below are some definitions for DAG:

1. Parent: If there is an edge from node a to node b, then node a is a parent of node b. In Figure 2, node 1 and node 2 are the parents of node 3, or we can write $pa(3) = \{1, 2\}$.

2. Child: If there is an edge from node a to node b, then node b is a child of node a. In Figure 2, node 4 and node 5 are the children of node 3, or we can write $child(3) = \{4, 5\}$.

3. Root: A node with no parents is called a root. It only has outgoing edges. In Figure 2, there are 2 roots, which are node 1 and node 2. $root = \{1, 2\}$.

4. Leaves: A node with no children is called a leave. It only has incoming edges. In Figure 2, there are 2 leaves, which are node 4 and node 5. $leaves = \{4, 5\}$.

5. Ancestors and Descendants: If a directed path from node a to node b exists, then node a is an ancestor of node b and node b is a descendent of node a. In Figure 2, $anc(5) = \{1, 2, 3\}$, and $desc(1) = \{3, 4, 5\}$.

6. acyclic: For each node i, $anc(i)$ does not contain i, i.e. $\forall i, i \notin anc(i)$.

5.2 Tree and Plate Notation

5.2.1 Tree

A DAG is a tree when each node has at most one parent, i.e. $\forall i, |pa(i)| \leq 1$. As shown in Figure 3, node 1 is the root, and nodes 4, 5, 6, and 7 are the leaves.

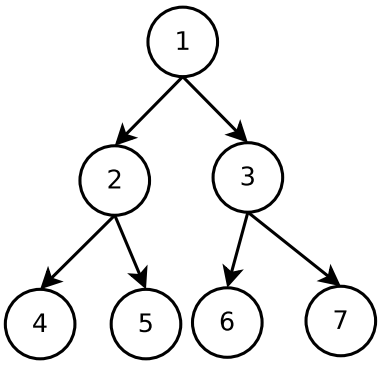

Figure 3: Tree: no node has more than one parent.

5.2.2 Plate Notation

Because the number of IID samples may be large, in which case using one node to represent each sample will make the graph unnecessarily large, we will use plate notation to represent copies of nodes and subnetworks.

For example, Y is a random variable, which each observation X depends on, and all of the m observation X's are conditionally independent when Y is known. The DAG can be simplified to two nodes plus a plate, representing m IID copies of the X random variable (Figure 4).

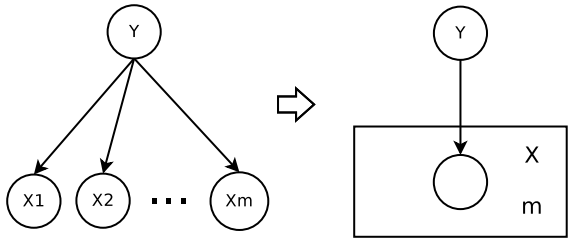

Figure 4: Plate notation (right panel).

5.3 How do graphical models encode a joint probability?

For a DAG, the joint probability of m random variables in the graph $\mathcal{G}$ can be *factorized* as

$$P(X_1, \cdots, X_m|G) = \prod_{i=1}^{m} P(X_i|X_{pa(i)}).$$

Using this formula, the joint probability for the DAG in Figure 2 can be rewritten as:

$$P(X_1, X_2, X_3, X_4, X_5|G) = P(X_1) \cdot P(X_2) \cdot P(X_3|X_1, X_2) \cdot P(X_4|X_3) \cdot P(X_5|X_3).$$

This *causal network* significantly reduces the parameters in the joint probability by exploiting conditional independencies. Assuming each random variable takes one of two values, 0 and 1, the joint probability includes $2^5 = 32$ parameters. However, if we factorize, or decompose, the joint probability as above, $P(X_1)$ and $P(X_2)$ each need one parameter, $P(X_4|X_3)$ and $P(X_5|X_3)$ each need two parameters, and $P(X_3|X_1, X_2)$ needs 4 parameters. The factorized joint probability only needs $2 * 1 + 2 * 2 + 4 = 10$ parameters to represent the entire joint probability. We define *in-degree* as the number of parents of a node, and *out-degree* as the number of children of a node. The size of a CPT is determined by the in-degree.

6 Conditional independence and canonical networks

Given a DAG graphical model, then how do we determine the conditional independent relations among nodes (variables)? It happens that there are three basic graphical patterns, called *canonical networks*, which can be used as building blocks for determining the conditional independence among nodes in a DAG model.

6.1 Canonical networks

There are three canonical networks, each of which corresponds to a basic three node structure and corresponding conditional independence relationships within that structure. These three canonical networks form the building blocks of all Bayesian networks, and we will use them in combination to build DAGs and to determine the conditional independencies in a graph with multiple nodes and edges.

6.1.1 Chain

A *chain* is composed of three nodes connected in a line along the same direction (Figure 5). For X, Y and Z in the figure, the conditional independencies are:

$$X \perp Z|Y.$$

It is a common practice to shade the nodes that are *observed*. When we condition on node Y, nodes X and Z are independent of each other.

To see this conditional independence, we factorize the joint distribution as

$$P(X, Y, Z) = P(X)P(Y|X)P(Z|Y).$$

By conditioning on Y, we have

$$P(X, Z|Y) = \frac{P(X)P(Y|X)P(Z|Y)}{P(Y)} = \frac{P(X, Y)P(Z|Y)}{P(Y)} = P(X|Y)P(Z|Y),$$

which proves that $X \perp Z|Y$.

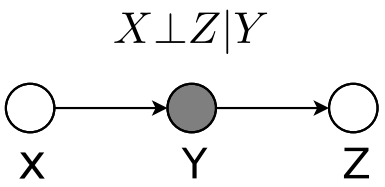

Figure 5: Canonical networks: the chain.

6.1.2 Fork

The second canonical network is called the *fork*, which is in the inverted V-shape (Figure 6). It follows from the definition of the graphical model that

$$P(X, Y, Z) = P(Y)P(X|Y)P(Z|Y).$$

Therefore, the conditional probability $P(X, Z|Y)$ can be rewritten as

$$P(X, Y|Z) = \frac{P(X, Y, Z)}{P(Y)} = P(X|Y)P(Z|Y),$$

which shows that $X \perp Z|Y$.

6.1.3 V-structure

The v-structure is the last canonical network (Figure 7). The conditional independencies it implies are different from the previous two structures: in this case, we have independence $X \perp Z$, which is **not conditioned** on Y.

Moreover, the conditional independence $X \perp Z|Y$ **does not hold**, because X and Z become related to each other once their outcome Y is observed. For example, supposing X refers to the event "Joe's watch stops"

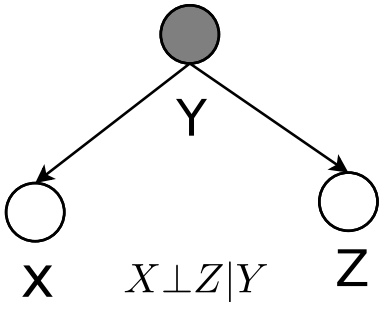

Figure 6: Canonical networks: the fork.

and Z refers to the event "Joe is abducted by an alien" and the event "Joe does not show up today" is denoted by Y. When we do not know whether Joe has shown up today or not, X and Z are independent – a watch stopping and alien abduction have nothing to do with each other – however, if we observe that $Y = 1$ (Joe does not show up), then $X = 1$ will be very unlikely if you are abducted by alien ($Z = 1$). In other words, knowing about one of the latent causes changes the probability of the other. Conditioning on a common child in a v-structure makes their parents become dependent, and this effect is called *"explaining away"*.

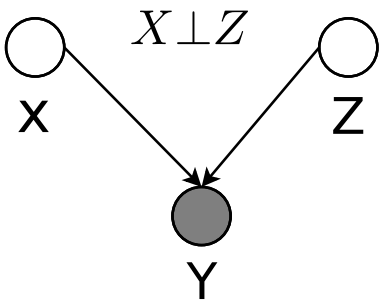

Figure 7: Canonical networks: the v-structure.

For the v-structure, the joint probability can be factorized as

$$P(X, Y, Z) = P(X)P(Z)P(Y|X, Z).$$

Then we can write

$$P(X, Z|Y) = \frac{P(X)P(Z)P(Y|X, Z)}{P(Y)},$$

but we can see there is no way to get this to $P(X|Y)P(Z|Y)$.
To show $X \perp Z$, we have

$$P(X, Z) = \sum_Y P(X)P(Z)P(Y|X, Z) = P(X)P(Z)\sum_Y P(Y|X, Z) = P(X)P(Z).$$

6.2 The Bayes Ball Algorithm

6.2.1 Definition

The Bayes Ball Algorithm is a method to test for conditional independencies in a network using a *reachability* argument. Let us suppose we place a ball on a node in the graph and the ball can reach other nodes by

moving along edges (in both directions) according to some rules. If the ball can reach from *start* to *end* with *observations* (*shaded*) in the graph, then we can conclude that *start* is **not** conditionally independent of *end* given the *observations*. If on the other hand, the ball cannot reach *end* from *start*, they are conditionally independent.

The rules for ball movement are different with regard to the type of node that the ball bounces into.

1. When the ball bounces into an *unshaded* node, then
   (a) if it moves in through an edge in, it can only go out of the node through an edge out;
   (b) if it moves in through an edge out, it can go out of the node through an edge either in or out.
2. When the ball bounces into a *shaded* node (observation), then
   (a) if it moves in through an edge in, it can go out of the node only through an edge in;
   (b) if it moves in through an edge out, then it cannot exit from the node.

6.2.2 Examples

For example, in the case of Figure 8, by the rule 1 above, the ball can reach B and C if starting from A (as shown in the left), from which we can conclude that $A \perp D$. Meanwhile, as shown in the right figure, C is non-independent to any other node.

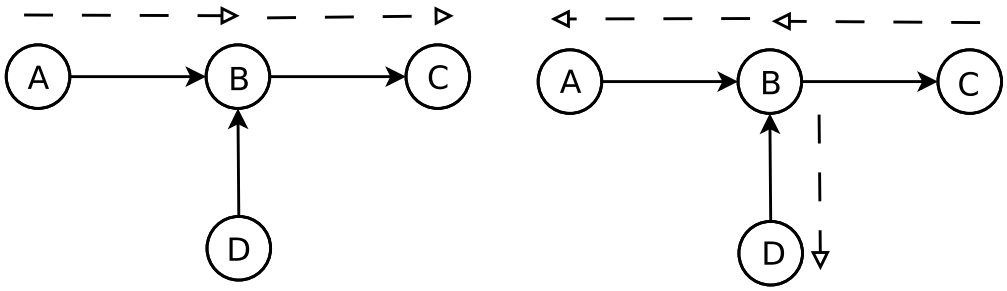

Figure 8: The ball from A can reach B, C (and A itself) and, from C, will reach A, B, D, by rule 1 above.

Another example is given by Figure 9 to illustrate the use of rule 2. In the left figure, the ball that starts from A can reach B and D, from which we learn that $A \perp C|B$; in the right figure, the ball that starts from C is stuck in B, from which we can conclude $A \perp C|B$ and $C \perp D|B$.

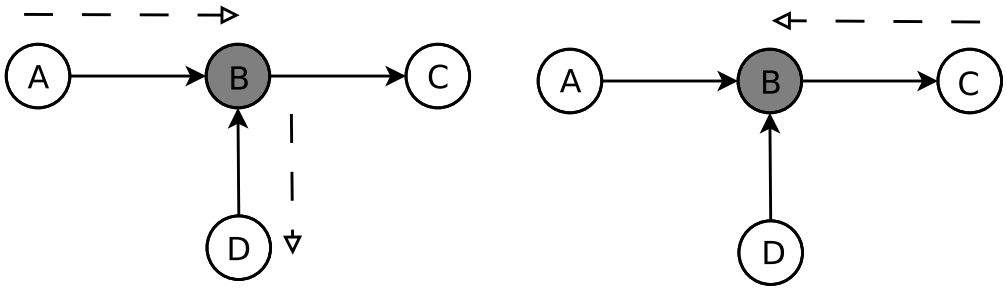

Figure 9: By rule 2, the ball released at A cannot reach C, while the ball released at C cannot exit from B.