COS513: NOTES FROM September 22, 2010

JOHN ASMUTH

0.1. Probability Review.

0.1.1. *Joint Distributions.* A joint distribution is the probability distribution over some set of random variables (RVs), denoted, for example, $P(X_1, X_2, X_3)$.

Since this is a distribution, we know that it necessarily sums to 1:

$$\sum_{x_1} \sum_{x_2} \sum_{x_3} p(x_1, x_2, x_3) = 1.$$

This particular distribution, assuming the RVs are binary, can be represented with a table of size 2^3, with one entry for each unique combination of the values for X_1, X_2 and X_3.

In general, a joint distribution over N binary RVs requires a table of size 2^N.

0.1.2. *Marginal Probabilities.* A marginal probability is found by summing out all other RVs. For example, if we know the joint $P(X_1, X_2)$ and we want to find the marginal $P(X_2)$, it can be found with the following computation:

$$P(X_2) = \sum_{x_1} P(X_1 = x_1, X_2).$$

0.1.3. *Conditional Probabilities.* A conditional probability is the distribution over values for one RV or set of RVs, assuming that the values for some other set of RVs are held constant. For example, the distribution over X_1 for some known value of X_2 is denoted $P(X_1|X_2)$, and read "the probability of X_1 given X_2". The '|' is called the "conditioning bar".

The conditional can be calculated by dividing the joint by the marginal. Figure 1 shows an example.

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)}$$

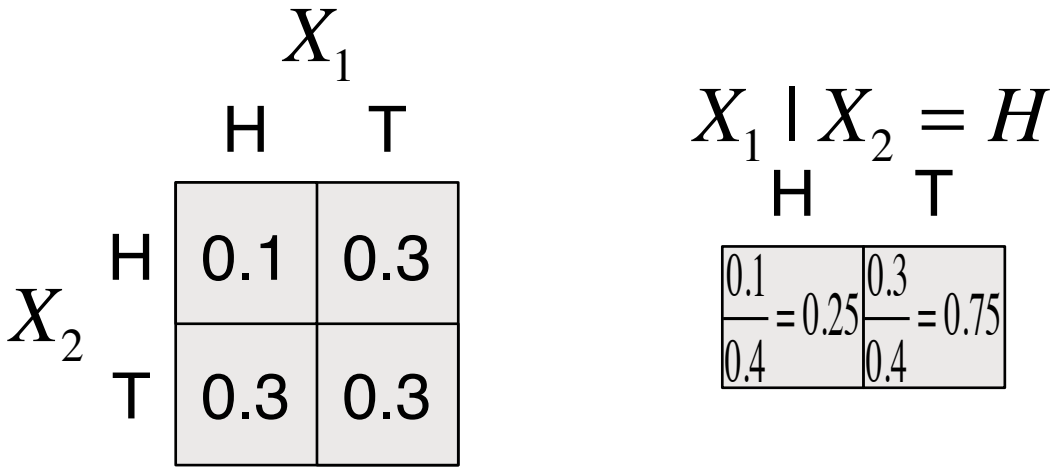

FIGURE 1. **Left:** A table representing the joint distribution over two binary RVs. **Right:** A table representing the conditional probability distribution $P(X_1|X_2 = H)$, calculated by dividing the joint $P(X_1, X_2)$ by the marginal $P(X_2)$.

0.1.4. *Independence.* The independence relation between two RVs X_1 and X_2 is written $X_1 \perp\!\!\!\perp X_2$.

$$\begin{aligned} X_1 \perp\!\!\!\perp X_2 \quad &\text{iff} \quad P(X_1, X_2) = P(X_1)P(X_2) \\ &\text{or} \quad P(X_1|X_2) = P(X_1) \end{aligned}$$

This relation is symmetric: $X_1 \perp\!\!\!\perp X_2 \leftrightarrow X_2 \perp\!\!\!\perp X_1$.

1. Graphical Models

Suppose we have a set of RVs $\{X_1, ..., X_n\}$, and we are interested in questions of independent and conditional probability. Both of these kinds of questions can be answered using the joint probability.

- Independence questions: factorization of the joint.
- Conditional questions: normalization and marginalization of the joint.

For now, we shall consider only discrete RVs.

If each RV can have r different values, then the table used to represent $P(X_1, ..., X_n)$ has r^n entries.

Graphical Models (GMs) provide a more economic representation of the joint by taking advantage of local relationships between RVs.

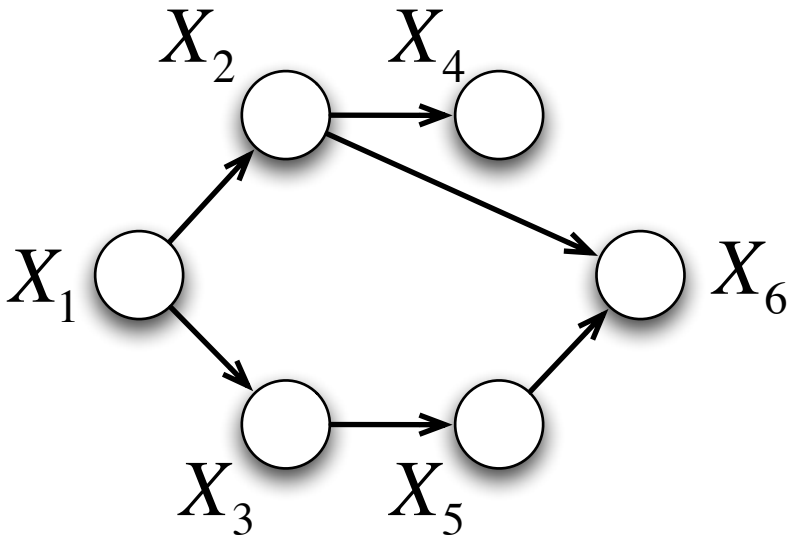

FIGURE 2. A DGM describing a family of joint distributions on the RVs $\{X_1, ..., X_6\}$. The DGM also indicates the "parent of" relationship for each RV. For example, $\pi_6 = \{X_2, X_5\}$.

1.1. **Directed GMs.** A Directed Graphical Model (DGM) is a Directed Acyclic Graph (DAG) $G = \{V, E\}$.

- The nodes correspond to RVs
- The edges denote a "parent of" relationship. $\pi_i \equiv$ parents of X_i

The joint probability defined by the DGM in Figure 2 is

$$P(X_1, ..., X_6) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_6|X_3)P(X_4|X_2)P(X_5|X_3)P(X_6|X_2, X_5).$$

In general, $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_{\pi_i})$.

As a consequence, the joint probability is defined in terms of many local probability tables.

The sizes of these tables grow exponentially with the number of parents, where before the size of the (one) table grew exponentially with the number of RVs.

The sizes of the tables for the DGM in Figure 2 are

$$\begin{aligned} & |P(X_1)| + |P(X_2|X_1)| + |P(X_3|X_1)| + |P(X_6|X_3)| + |P(X_4|X_2)| + |P(X_5|X_3)| + |P(X_6|X_2, X_5)| \\ = \; & 2 + 4 + 4 + 4 + 4 + 8 \\ = \; & 26. \end{aligned}$$

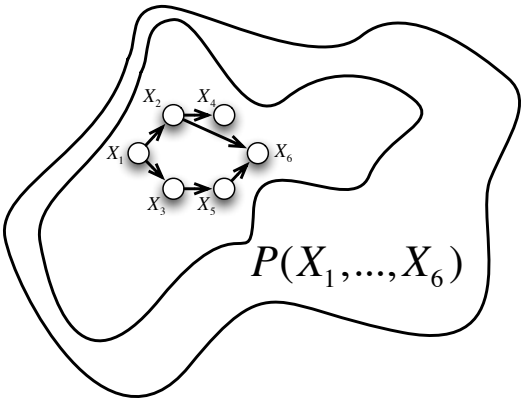

FIGURE 3. The DGM from Figure 2 describes only a subset of all possible distributions on 6 RVs.

In all, 26 numbers are required to exactly represent distributions for the DGM in Figure 2, when the RVs are all binary. A full description would require $2^6 = 64$ numbers.

By choosing specific tables for all $P(X_i|X_{\pi_i})$, we produce a joint.

N.B. The possible joints are not all possible $P(X_1, ..., X_6)$. That is, a set of distributions indicated by a DGM is a subset of all possible joint distributions for the RVs.

1.1.1. *Conditional Independence.* Recall that for some sets of RVs X_A and X_B, that

$$X_A \perp\!\!\!\perp X_B \Leftrightarrow P(X_A, X_B) = P(X_A)P(X_B).$$

We can also denote conditional independence on another set of RVs X_C with

$$\begin{aligned} X_A \perp\!\!\!\perp X_B|X_C \quad &\Leftrightarrow \quad P(X_A, X_B|X_C) = P(X_A|X_C)P(X_B|X_C) \\ &\text{or} \quad P(X_A|X_B, X_C) = P(X_A|X_C). \end{aligned}$$

Questions of independence are questions about how the marginals factorize. These questions can be answered by examining the GM structure.

Basic conditional independence statements:

- Chain rule:

$$\begin{aligned} P(X_1, ..., X_6) &= P(X_1) \\ &\quad P(X_2|X_1) \\ &\quad P(X_3|X_1, X_2) \\ &\quad P(X_4|X_1, X_2, X_3) \\ &\quad P(X_5|X_1, X_2, X_3, X_4) \\ &\quad P(X_6|X_1, X_2, X_3, X_4, X_5). \end{aligned}$$

  In general,

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1}).$$

  It is easy to show this for a joint of 3 RVs:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \\ &= P(X_1)\frac{P(X_1, X_2)}{P(X_1)}\frac{P(X_1, X_2, X_3)}{P(X_1, X_2)} \\ &= P(X_1, X_2, X_3). \end{aligned}$$

  The chain rule suggests:

$$P(X_6|X_1, ..., X_5) = P(X_6|X_2, X_5) \Leftrightarrow X_6 \perp\!\!\!\perp \{X_1, X_3, X_4\}|\{X_2, X_5\}.$$

  The statement $X_6 \perp\!\!\!\perp \{X_1, X_3, X_4\}|\{X_2, X_5\}$ is one of the basic conditional independence statements about our example DGM in Figure 2.

- Let I be a topological ordering of the RVs. That is, if $j \in \pi_i$, then j precedes i in the order. Let ν_i be the set of indices appearing before i, not including those in π_i. The basic conditional independence statements are

$$\{X_i \perp\!\!\!\perp X_{\nu_i}|X_{\pi_i}\}.$$

  For example, if $I = \{1, 2, 3, 4, 5, 6\}$ is the topological ordering for the DGM in Figure 2, we have the following basic independence statements:

$$\begin{array}{lcc|c} X_1 & \perp\!\!\!\perp & \emptyset & \emptyset, \\ X_2 & \perp\!\!\!\perp & \emptyset & X_1, \\ X_3 & \perp\!\!\!\perp & X_2 & X_1, \\ X_4 & \perp\!\!\!\perp & \{X_1, X_3\} & X_2, \\ X_5 & \perp\!\!\!\perp & \{X_1, X_2, X_4\} & X_3, \\ X_6 & \perp\!\!\!\perp & \{X_1, X_3, X_4\} & \{X_2, X_5\}. \end{array}$$

  Question: are these the only independence assumptions we can make from our DGM? (no)

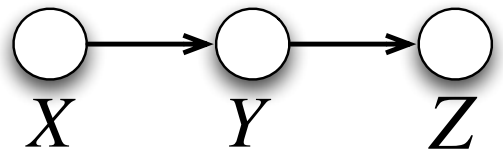

FIGURE 4. "A little sequence." In this DGM, $P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$.

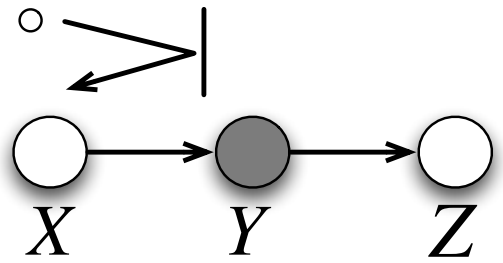

FIGURE 5. Same DGM as in Figure 4, except the node for Y is shaded to show that we are conditioning on its value.

1.1.2. *Bayes Ball Algorithm.* The metaphor is that of bouncing a ball around your GM. If the ball can "bounce" from one RV node to another, then we cannot make an independence assumption about those two nodes.

To illustrate the rules which determine how the ball may bounce, we use examples on 3-node DGMs. One of which appears in this lecture.

(1) Figure 4 shows a DGM that implies exactly one independence statement:

$$X \perp\!\!\!\perp Z|Y.$$

No other independence statements necessarily hold for all joints in the family described in Figure 4. Figure 5 shows us how conditioning on Y's value will block a ball from bouncing from X to Z.

Presumably we will hear more about this algorithm in the future.