COS 424: Interacting with Data

Lecturer: Dave Blei
Scribe: Sunghwan Ihm

Lecture #10
March 8, 2007

Homework #3 will be released soon. Some questions and answers given by comments: Computational efficiency of K-means clustering? It will be exponentially increased. There is a paper which shows that K-means takes $2^{\Omega\sqrt{n}}$ time in the worst case, where n is the number of data points. (http://www.stanford.edu/~darthur/kMeansLb.pdf) The relationship between K-means clustering and CS research? For example, similarity search.

1 Introduction

Probability is used to model uncertainty about data, and its role in CS is increasing now. (ex: randomized algorithm) Graphical models (or Baysien Network, in COS 402) are a marriage of probability theory and graph theory. Today, we will cover an overview and concrete example of Naive Bayes. This will be a building block for subsequent material.

1.1 Joint Probability Distribution

Let $\{X_1, X_2, \ldots, X_N\}$ be a set of random variables, for example, flips of coins, Gaussian random variables, DNA sequences, interactions between monkeys, and so on. These random variables are governed by a joint distribution. (Assume we have access to it.) Then some questions we can ask are:

marginal independence $X_A \perp\!\!\!\perp X_B$: Are X_A and X_B independent?

conditional independence $X_A \perp\!\!\!\perp X_B \mid X_C$: Given X_C, are X_A and X_B independent?

conditional probability $P(X_A \mid X_C)$: Given X_C, what is $P(X_A)$?

where X is a random variable, x is a realization of random variable, and A, B, C are sets of indices. All these can be answered by manipulating the joint distribution.

For example, we can compute $P(X_A, X_B)$ using *factorization*:

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

And how do we compute conditional probability $P(X_A \mid X_C)$? We can compute this using the *joint and marginal probability distributions*:

$$P(X_A \mid X_C) = \frac{P(X_A, X_C)}{P(X_C)}$$

Finally, how do we get $P(X_C)$? We can use *marginalization*:

$$P(X_C) = \sum_{\neg C} P(X_C, X_{\neg C})$$

1.2 Representation

Let's say we have 10 coin tosses. We need a (really big) table of size 2^N to represent the joint distribution, and it will be a problem. Here, graphical models solve this by taking advantages of the local relationships of conditional independence. Let's look at how the graphical models can solve this.

2 Discrete Graphical Model

Here we will cover a high-level overview of a graphical model. We will deal with discrete random variables only. Graphical models are *directed acyclic graphs* (DAG). Nodes are (individual) random variables and edges denote possible dependence. Thus, X_{Π_i} are the parents of X_i and X_i is dependent on X_{Π_i}. Figure 1 shows an example of a graphical model with 3 random variables. In the example, X_3 is dependent on values of X_1 and X_2.

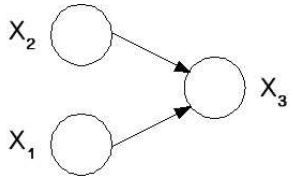

Figure 1: An example of a graphical model

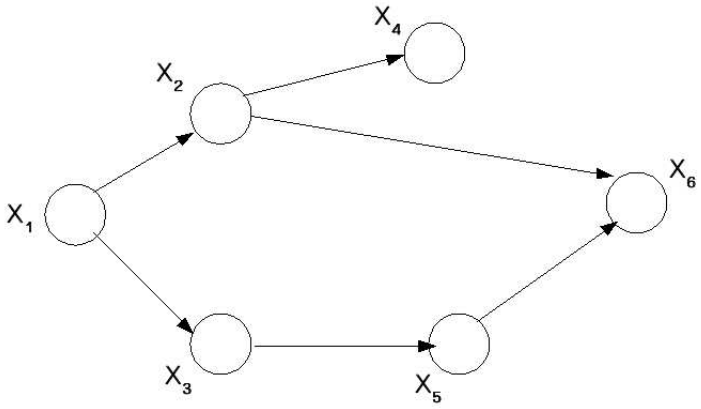

Figure 2: An example of a graphical model with 6 random variables

Now let's consider the graph with 6 random variables in Figure 2 and try to compute the joint distribution $P(X_1, X_2, \ldots, X_6)$. Without the graphical model, we can calculate this by definition (chain rule):

$$P(X_1, X_2, \ldots, X_6) = \prod_{i=1}^{6} P(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

But with the graphical model, the computation can be much simpler. We just multiply the conditional probabilities by exploiting the conditional independence relationships in the graphical model.

$$P(X_1, X_2, \ldots, X_6) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2)P(X_5 \mid X_3)P(X_6 \mid X_2, X_5)$$

On the right hand side, the term $P(X_1)$ is because $P(X_1)$ has no parent, and the term $P(X_2 \mid X_1)$ is because the parent of X_2 is X_1, and so on. In the graphical model, the joint distribution factorizes like this.

2.1 Local Probability Tables

The joint distribution is defined in terms of local probability tables. For binary random variables, the number of entries required is 2 for $P(X_1)$, and 4 for $P(X_2 \mid X_1)$. In general,

$$P(X_i \mid X_{\pi_i}) = \begin{cases} 4 & \text{if } |\pi_i| = 1 \\ 8 & \text{if } |\pi_i| = 2 \\ 2^{K+1} & \text{if } |\pi_i| = K \end{cases}$$

(Note that the joint distribution is not well defined in cycle. Also, in implementation, we can reduce the size of the table by only storing the $P(Head)$ and get $P(Tail)$ by $1 - P(Head)$. But for now, just be naive for illustration.)

So, how big is the full joint distribution with $N = 6$ in our example? It has $2^6 = 64$ entries. But with the graphical model, our representation has only $2 + 4 + 4 + 4 + 4 + 8 = 26$ entries. So why is the graphical model so important?

2.2 What the graphical model does (very important!)

2.2.1 We replaced exponential growth in N with exponential growth in $|\pi_h|$.

This is a huge saving in terms of both space and computation because many applications require a huge number of random variable, for example, sequence of DNAs, languages, or documents. (Assuming we made or know the graph.)

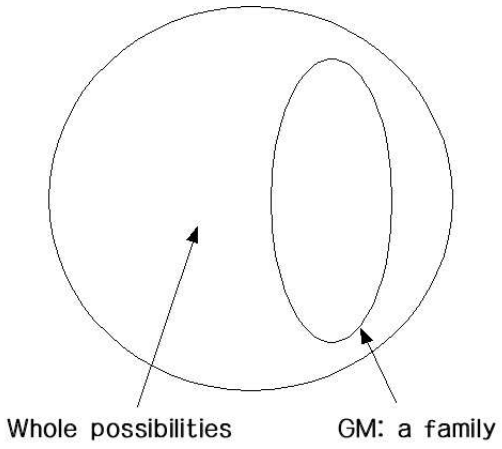

Figure 3: GM represents a family of distributions

2.2.2 The graphical model represents a *family* of distributions.

We've been discussing this without saying concrete probabilities like 0.6 or 0.4. Instead, we define a family and we can freely change local probabilities. So, the question is, are all the joint distributions in this family? (Suppose the graph is fixed.) With this graphical model, can I get every distribution? The answer is no. There are some joint distributions which is not in this family. Because not

every joint distribution factorizes like this! (Remind the number of local probability table entries: 64*vs.*26) For example, in Figure 2, X_5 can be dependent on X_2, but it cannot be represented in the graph. Thus, building a graphical model is imposing a factorization on joint distributions. Figure 3 depicts that a family is a subset of the whole possibilities. So, why do we need a family? There are 3 reasons.

1. We can do efficient inference. Complexity can be controlled by computing conditional and marginal independence probabilities. Generally the joint probability is given by *the chain rule*,

$$\begin{aligned} P(X_1, X_2, \ldots, X_6) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)P(X_4 \mid X_1, X_2, X_3) \\ &= P(X_5 \mid X_1, \ldots, X_4)P(X_6 \mid X_1, \ldots, X_5) \end{aligned}$$

But in our graphical model,

$$P(X_1, X_2, \ldots, X_6) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2)P(X_3 \mid X_1) = P(X_3 \mid X_1, X_2)$$

since X_3 is conditionally independent of X_2 given X_1 ($X_3 \perp\!\!\!\perp X_2 \mid X_1$) and this means if we know X_1, knowing X_2 gives no additional information. Also,

$$P(X_4 \mid X_1, X_2, X_3) = P(X_4 \mid X_2)$$

since X_4 is conditionally independent of X_1 given X_2 ($X_4 \perp\!\!\!\perp X_1 \mid X_2$) and X_4 is conditionally independent of X_3 given X_2. ($X_4 \perp\!\!\!\perp X_3 \mid X_2$) In the graphical model, when j is an ancestor of i, then $X_i \perp\!\!\!\perp X_j \mid X_{\pi_i}$ holds.

2. We can answer questions efficiently with *Bayes Ball* algorithm. "Bayes Ball" algorithm answers all probability questions efficiently. We will cover this later in the semester.

3. Graphical models are great for model building. It is an essential tool for complicated interacting data sets. We encode assumptions about data and reuse the structure model. Other than the graphical model, there are other similar algorithms. Figure 4 shows the hidden Markov model and Figure 5 depicts an example usage of Kalman Filter.

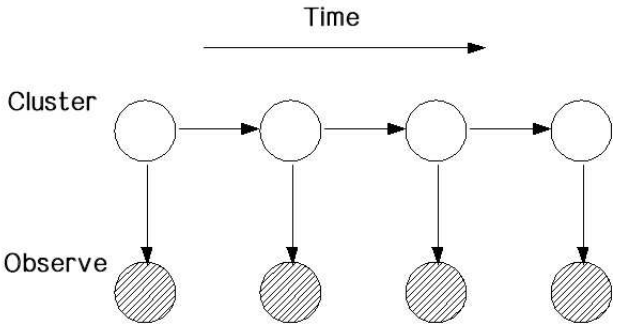

Figure 4: Hidden Markov Model

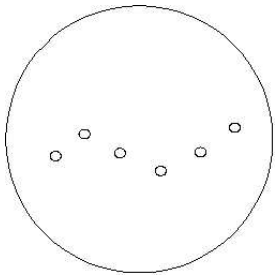

Figure 5: The positions of a plane

In Kalman Filter (in EE), positions of a plane predict the next position. Surprisingly, these two algorithms which are developed from totally different communities follow the same graphical model.

So, how to come up with the graphical model? It's a dark art. To build new creative graphical models, you should take a walk, and think. There is some research which tries to find the independence assumptions automatically. This is the end of high-level overview of the graphical model.

3 Gaussian i.i.d. Model

Let's assume one-dimension random variables from the same Gaussian distribution. Figure 6 shows the Gaussian i.i.d. model and Figure 7 represents plate notation.

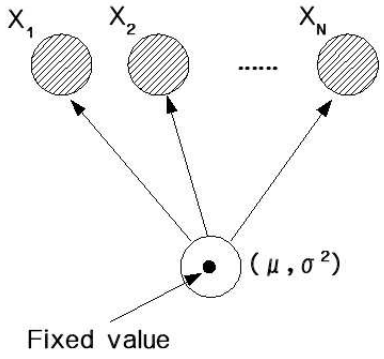

Figure 6: Gaussian i.i.d. Model

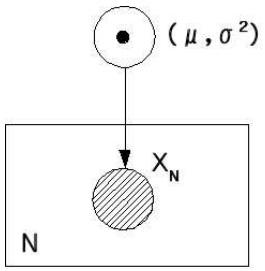

Figure 7: Plate Notation

4 Next Lecture

Next time we will cover Naive Bayes Classification of documents.