Foundations of Graphical Models: Homework 1

Out: W 2015-09-30
Due: M 2015-10-12

Problem 1

Consider binary random variables $X_1, X_2, \cdots, X_6$.

a) How large is the probability table of the joint distribution?

b) How large is the probability table of the conditional distribution $p(X_1|X_2, X_3, X_4)$?

c) How large is the probability table of the conditional distribution $p(X_1, X_2|X_3, X_4, X_5)$?

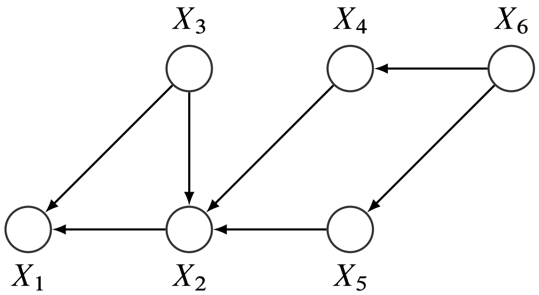

Figure 1: Graphical model.

d) Figure 1 shows a graphical model for these nodes. What is the factorized joint distribution?

e) How many entries are required for the corresponding factorized distribution?

f) How many entries are involved in $p(X_1|X_2, X_3, X_4)$ assuming it is consistent with the graphical model?

g) How many entries are involved in $p(X_1, X_2|X_3, X_4, X_5)$ assuming it is consistent with the graphical model?

Problem 2

D-separation Use the Bayes-Ball algorithm to determine if the following (conditional) independence assumptions hold in Figure 1:

a) $X_1 \perp\!\!\!\perp X_4$

b) $X_1 \perp\!\!\!\perp X_4 | X_3$

c) $X_1 \perp\!\!\!\perp X_4 | X_2$

d) $X_1 \perp\!\!\!\perp X_4 | X_2, X_3$

e) $X_3 \perp\!\!\!\perp X_4$

f) $X_3 \perp\!\!\!\perp X_4 | X_1$

g) $X_3 \perp\!\!\!\perp X_4 | X_1, X_2$

Problem 3

Consider a sequence of T observed random variables $x_1, \cdots, x_T$. In this problem we will explore how to encode different independence assumptions in graphical models.

a) Draw a directed graphical model where all variables are independent.

b) Draw a directed graphical model where the variables obey the Markov assumption.

c) Draw a directed graphical model where (i) no two random variables x_i, x_j can be assumed independent given all other x and (ii) there can be additional nodes, but the number of edges does not exceed the number of nodes.

d) What are some qualitative differences between the three models? What are different types of data in which you would use one model or the other?

Problem 4

A dishonest soccer referee uses two different coins, one is fair and one is loaded. For the fair coin F, the probability of tossing heads equals the probability of tossing tails, i.e. $p(\text{heads}) = p(\text{tails}) = 0.5$. For the loaded coin L, $p(\text{heads}) = 0.8$ and $p(\text{tails}) = 0.2$.

A soccer tournament has three games. Between games, the referee probabilistically switches between coins. The varaible $z_t \in \{F, L\}$ is a (hidden) variable indicating whether the fair coin F or the loaded coin L was used in game $t \in \{1, 2, 3\}$. The conditional probabilities are as follows,

$$\begin{aligned} p(z_t = F|z_{t-1} = F) &= 0.8 \\ p(z_t = L|z_{t-1} = F) &= 0.2 \\ p(z_t = F|z_{t-1} = L) &= 0.1 \\ p(z_t = L|z_{t-1} = L) &= 0.9. \end{aligned}$$

The initial probability of choosing the fair coin is $p(z_1 = F) = 0.75$.

At the beginning of each game, the referee chooses a coin and flips it. (This is to determine which team gets to pick a side.) We observe that the outcomes in the three games are $\{\text{heads}, \text{tails}, \text{heads}\}$. Note we do not observe which coin was flipped for each game.

We will use an HMM to try to infer which coin was used in each game.

a) Draw the graphical model for the corresponding HMM. The three hidden states z_1, z_2, z_3 correspond to indicators of which coin was used; the three observations x_1, x_2, x_3 correspond to the outcomes of the coin tosses.

b) We can use message passing to compute the marginal probabilities of each hidden state given the observations. Some messages depend on other messages. What is a good ordering in which to compute the messages? Why?

c) Use the message passing schedule you specified above to compute $p(z_2)$, i.e., the marginal probability of the hidden state of the second coin toss.

Problem 5

There are two goals of this exercise. (i) The first goal is for you to familiarize yourself with the tools for loading, pre-processing and visualizing data. (We recommend using Python and/or R.) (ii) The second goal is for you to start thinking about the kinds of problems you will solve in your final project.

a) Choose a data set to work with. We encourage you to find a data set you might be interested in analyzing for your final project. Describe the data. How was it collected? Why is it interesting? What are some problems that involve this data?

b) For this problem you will load the data and form visualizations about it. (These can be plots, graphs, tables, or other things.) For example, you might make a scatter plot to compare two of the dimensions and notice something about their relationship. Or your data may involve a time series, and a plot reveals an interesting seasonal pattern. Be creative, and experiment with different kinds of visualizations.

   Form three different visualizations. For each, briefly write about why it is interesting. What does it tell you about the data?