CS 67800, Spring 2017/18

Problem Set 1: Bayesian Networks

Submission Date : Sunday 15/4/18, 23:59

---

1. Consider the following distribution over 3 binary variables X, Y, Z:

$$P(x,y,z) \quad = \quad \begin{cases} 1/12 & x \oplus y \oplus z = 0 \\ 1/6 & x \oplus y \oplus z = 1 \end{cases}$$

   (where $\oplus$ denotes a XOR function).
   Show that there is no DAG structure such that $I_{d-sep}(G) = I(P)$.
   (Hint: show that $(X \perp Z) \in I(P)$ and $(X \perp Y) \in I(P)$).

2. Let X, Y, Z be binary random variables with joint distribution that factorizes over the directed graph $X \rightarrow Z \leftarrow Y$ (v-structure). We define the following quantities:

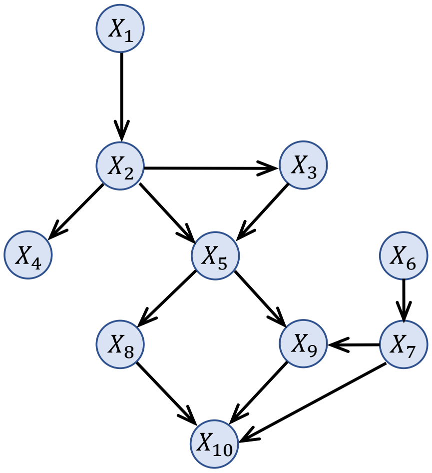
$$\begin{aligned} a =& P(X=1) \\ b =& P(X=1|Z=1) \\ c =& P(X=1|Z=1, Y=1) \end{aligned}$$

   (a) For all the following cases, provide examples of conditional probability tables (table CPDs), and compute a, b, c, such that:
      - $a > c$
      - $a < c < b$
      - $b < a < c$

   (b) Think of X, Y as causes of Z, and for all the above cases summarize (in a sentence or two) why the statements are true for your examples.

   (Hint: think about positive and negative correlations along edges).

3. **Markov blanket**
   Let $\mathcal{X} = \{X_1, ..., X_n\}$ be a set of random variables with distribution P given by the following graph.

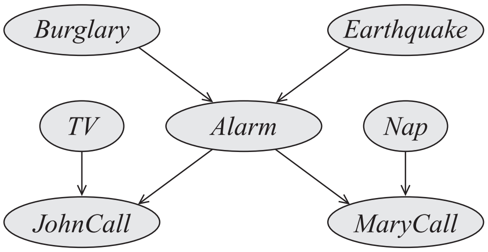


(a) Consider the variable X_5. What is the minimal subset of the variables, $A \subseteq \mathcal{X} - \{X_5\}$, such that $(X_5 \perp \mathcal{X} - A - \{X_5\} | A)$? Justify your answer.

(b) Now, generalize this to any BN defined by $(\mathcal{G}, P)$. Specifically, consider variable X_i. What is the *Markov blanket* of X_i? Namely, the minimal subset of variables $A \subseteq \mathcal{X} - \{X_i\}$ such that $(X_i \perp \mathcal{X} - A - \{X_i\} | A)$? Prove that this subset is necessary and sufficient.

(Hint: Think about the variables that X_i cannot possibly be conditionally independent of, and then think some more).

4. **Bayesian networks (Exercise 3.11 from Koller-Friedman)**



(a) Consider the Burglary Alarm network given above. Construct a Bayesian network over all the node **except** the Alarm that is a minimal I-map for the marginal distribution over the remaining variables (namely, over B, E, N, T, J, M). Be sure to get all the dependencies from the original network.

(b) Generalize the procedure you used above to an arbitrary network. More precisely, assume we are given a network BN, an ordering $X_1, \cdots, X_n$ that is consistent with the ordering of the variables in BN, and a node X_i to be removed. Specify a network BN' such that BN' is consistent with this ordering, and such that BN' is a minimal I-map of $P_{BN}(X_1, \cdots, X_i, X_{i+1} \cdots X_n)$. Your answer must be an explicit specification of the set of parents for each variable in BN'.

5. **Towards inference in Bayesian networks**

Suppose you have a Bayes' net over variables $X_1, \cdots, X_n$ and all variables except X_i are observed. Using the chain rule and Bayes' rule, find an efficient algorithm to compute $P(x_i | x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$. In particular, your algorithm should not require evaluation of the full joint distribution.

6. **Programming Task**

In this programming assignment, we will investigate the structure of the binarized MNIST dataset of handwritten digits using Bayesian networks. The dataset contains images of handwritten digits with dimensions 28 × 28 (784) pixels. Consider the Bayesian network in Figure 1. The network contains two layers of variables. The variables in the bottom layer, $X_{1:784}$ denote the pixel values of the flattened image and are referred to as *manifest variables*. The variables in the top layer, Z_1 and Z_2, are referred to as *latent variables*, because the value of these variables will not be explicitly provided by the data and will have to be inferred.

The Bayesian network specifies a joint probability distribution over binary images and latent variables $p(Z_1, Z_2, X_{1:784})$. The model is trained so that the marginal probability of

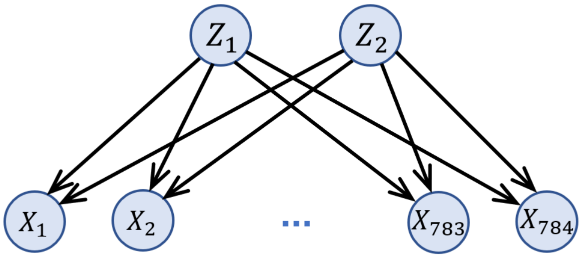


Figure 1: Bayesian network for the MNIST dataset. $X_{1:784}$ variables correspond to pixels in an image. Z_1 and Z_2 variables are latent.

the manifest variables, $p(x_{1:784}) = \sum_{z_1,z_2} p(z_1, z_2, x_{1:784})$ is high on images that look like digits, and low for other images.

For this programming assignment, we provide a pretrained model `trained_mnist_model`. The starter code `pa1.py` loads this model and provides functions to directly access the conditional probability tables. Further, we simplify the problem by discretizing the latent and manifest variables such that $Val(Z_1) = Val(Z_2) = \{-3, -2.75, \dots, 2.75, 3\}$ and $Val(X_j) = \{0, 1\}$, i.e., the image is binary.

(a) How many values can the random vector $X_{1:784}$ take, i.e., how many different 28×28 binary images are there?

(b) How many parameters would you need to specify an arbitrary probability distribution over all possible 28×28 binary images?

(c) How many parameters do you need to specify the Bayesian network in Figure 1?

For parts 6d-6g below, refer to `pa1.py`. The starter code contains some helper functions for solving these questions. It is not compulsory to use them and you are allowed to use your own implementations. Also, feel free to introduce your own additional helper functions when useful.

(d) Produce 5 samples from the joint probability distribution $(z_1, z_2, x_{1:784}) \sim p(Z_1, Z_2, X_{1:784})$, and plot the corresponding images (values of the pixel variables).

Hint: they should look like (binarized) handwritten digits. Imagine we could build such a model not for handwritten digits, but for Renaissance paintings. Each sample from the model would produce a new piece of art!

(e) For each possible value of

$$(\overline{z}_1, \overline{z}_2) \in \{-3, -2.75, \dots, 2.75, 3\} \times \{-3, -2.75, \dots, 2.75, 3\},$$

compute the conditional expectation $E[X_{1:784} | Z_1, Z_2 = (\overline{z}_1, \overline{z}_2)]$. This is the expected image corresponding to each possible value of the latent variables Z_1, Z_2. Plot the images on on a 2D grid where the grid axes correspond to Z_1 and Z_2 respectively. What is the intuitive role of the Z_1, Z_2 variables in this model?

(f) In `q_6f.mat`, you are given a *validation* and a *test* dataset. In the test dataset, some images are "real" handwritten digits, and some are anomalous (corrupted images). We would like to use our Bayesian network to distinguish real images from the anomalous ones. Intuitively, our Bayesian network should assign low probability to corrupted images and high probability to the real ones, and we can use this for classification. To do this, we first compute the average marginal log-likelihood,

$$\log p(x_{1:784}) = \log \sum_{z_1} \sum_{z_2} p(z_1, z_2, x_{1:784})$$

on the validation dataset, and the standard deviation (again, standard deviation over the validation set). Consider a simple prediction rule where images with marginal log-likelihood, $\log p(x_{1:784})$, outside three standard deviations of the average marginal log-likelihood are classified as corrupted. Classify images in the test set as corrupted or real using this rule. Then plot a histogram of the marginal log-likelihood for the images classified as "real". Plot a separate histogram of the marginal log-likelihood for the images classified as "corrupted".

Hint: If you run into any flow issues, search for the "log-sum-exp trick" online for help.

(g) In `q_6g.mat`, you are given a labeled dataset of images of handwritten digits (the label corresponds to the digit identity). For each image I^k, compute the conditional probabilities $p((Z_1, Z_2) = (\overline{z}_1, \overline{z}_2) | X_{1:784} = I^k)$. Use these probabilities to compute the conditional expectation

$$E[(Z_1, Z_2) | X_{1:784} = I^k)]$$

Plot all the conditional expectations in a single plot, color coding each point as per their label. What is the relationship with the figure you produced for part 6e?