Homework 1
CS228: Probabilistic Graphical Models

Instructor: Stefano Ermon
ermon@stanford.edu

Available: Jan. 7, 2020
Due date: 11:59pm January 21, 2020, via GradeScope. Total points: 100

Problem 1: Probability theory (4 points)

The doctor has bad news and good news for X. The bad news is that X tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, and the probability of testing negative given that you don't have the disease is also 0.99). The good news is that this is a rare disease, striking only one in 10,000 people. Why is it good news that the disease is rare? What are the chances that X actually has the disease?

Problem 2: Review of dynamic programming (7 points)

Suppose you have a probability distribution P over random variables $X_1, X_2, \ldots, X_n$ which all take values in the set $\mathcal{S} = \{v_1, \ldots, v_m\}$, where the v_j are some distinct values (e.g., integers or letters).

Suppose that P satisfies the *Markov assumption*: for all $i \geq 2$ we have

$$P(x_i|x_{i-1}, \ldots, x_1) = P(x_i|x_{i-1}).$$

In other words, P factorizes as

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)\cdots P(x_n|x_{n-1}).$$

For each factor $P(x_i|x_{i-1})$ for $i \geq 2$ you are given the probability $P(X_i = u|X_{i-1} = v)$ for each $u, v \in \mathcal{S}$ in the form of a $m \times m$ table. You are also given $P(X_1 = v)$ for each $v \in \mathcal{S}$.

- (7 points) Find an algorithm to solve the optimization problem

$$\max_{x_1, x_2, \ldots, x_n \in \mathcal{S}^n} P(x_1, x_2, \ldots, x_n).$$

  State the complexity of your algorithm using Big O notation. Your algorithm should run in time polynomial in m and n. (Hint: use dynamic programming, decompose the problem into a sequence of optimization problems, each over a single variable.)

Problem 3: Bayesian networks (6 points)

Let us try to relax the definition of Bayesian networks by removing the assumption that the directed graph is acyclic. Suppose we have a directed graph $G = (V, E)$ and discrete random variables $X_1, \ldots, X_n$, and define

$$f(x_1, \ldots, x_n) = \prod_{v \in V} f_v(x_v|x_{pa}(v))$$

where $X_{pa}(v)$ refers to the parents of variable X_v in G and $f_v(x_v|x_{pa}(v))$ specifies a distribution over X_v for every assignment to the parents of X_v, i.e. $0 \leq f_v(x_v|x_{pa}(v)) \leq 1$ for all $x_v \in Val(X_v)$, and

for all $x_{pa}(v) \in Val(X_{pa}(v))$ we have $\sum_{x_v \in Val(X_v)} f_v(x_v|x_{pa}(v)) = 1$. Recall that this is precisely the definition of the joint probability distribution associated with the Bayesian network G, where the f_v are the conditional probability distributions. Show that if G has a directed cycle, f may no longer define a valid probability distribution.

In particular, give an example of a cyclic graph G and distributions f_v that leads to an improper probability distribution. Remember, a valid probability distribution must be non-negative and sum to one. This is why Bayesian networks must be defined on acyclic graphs.

Problem 4: Conditional Independence (12 points)

The question investigates the way in which conditional independence relationships affect the amount of information needed for probabilistic calculations. Let α, β, and γ be three random variables.

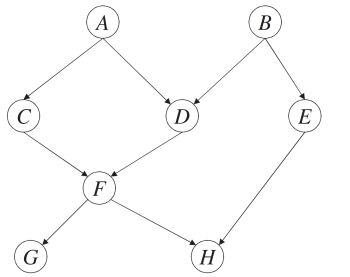
- (6 points) Suppose we wish to calculate $\Pr(\alpha|\beta,\gamma)$ and we have no conditional independence information. Which of the following quantities is sufficient for the calculation? (Assuming that all the parameters about the provided distributions are given.)

  1. $\Pr(\beta,\gamma)$, $\Pr(\alpha)$, $\Pr(\beta|\alpha)$ and $\Pr(\gamma|\alpha)$.
  2. $\Pr(\beta,\gamma)$, $\Pr(\alpha)$ and $\Pr(\beta,\gamma|\alpha)$
  3. $\Pr(\beta|\alpha)$, $\Pr(\gamma|\alpha)$ and $\Pr(\alpha)$.

  For each case, justify your response either by showing how to calculate the desired answer or by explaining why this is not possible.

- (6 points) Suppose we know that β and γ are conditionally independent given α. Now which of the preceding three sets is sufficient? Justify your response as before.

Problem 5: Bayesian networks (AD Exercise 4.1) (5 points)



A	Θ_A
1	.2
0	.8

B	Θ_B
1	.7
0	.3

B	E	$\Theta_{E\|B}$
1	1	.1
1	0	.9
0	1	.9
0	0	.1

A	B	D	$\Theta_{D\|AB}$
1	1	1	.5
1	1	0	.5
1	0	1	.6
1	0	0	.4
0	1	1	.1
0	1	0	.9
0	0	1	.8
0	0	0	.2

Consider the Bayesian network $\mathcal{B}$ given above.

1. (2 points) Compute $\Pr(A = 0, B = 0)$ and $\Pr(E = 1|A = 1)$. Justify your answers.

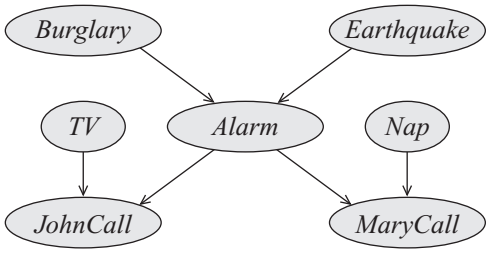
2. (3 points) True or false? Why?
   (a) d-sep$_{\mathcal{B}}(A; E|\{B, H\})$
   (b) d-sep$_{\mathcal{B}}(G; E|D)$
   (c) d-sep$_{\mathcal{B}}(\{A, B\}; \{G, H\}|F)$, (Notation: $\{A, B\}$ and $\{G, H\}$ are pairwise independent conditioned on F).

Problem 6: Bayesian Networks and explaining away (7 points) You want to model the admission process of Farm University. Students are admitted based on their Creativity (C) and Intelligence (I). You decide to model them as continuous random variables, and your data suggests that both are uniformly distributed in $[0, 1]$, and are independent of each other. Formally $I \sim \text{Uniform}[0, 1]$, $C \sim \text{Uniform}[0, 1]$, $C \perp I$. Being very prestigious, the school only admits students such that $C + I \geq 1.5$.

1. (1 points) What's the expected Creativity score of a student?
2. (2 points) What's the expected Creativity score of an admitted student?
3. (2 points) What's the expected Creativity score of a student with $I = 0.95$ (a highly intelligent student)?
4. (2 points) What's the expected Creativity score of an admitted student with $I = 0.95$? How does it compare to the expected Creativity score of an admitted student (computed in 2)?

Hint: it might be helpful to think about the correlation between Creativity and Intelligence in the general student population and among admitted students.

Problem 7: Bayesian networks (Exercise 3.11 from Koller-Friedman) (16 points)



1. (8 points) Consider the Burglary Alarm network given above. Construct a Bayesian network, over all the nodes **except** 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.

2. (8 points) Generalize the procedure you used above to an arbitrary network. More precisely, assume we are given a network BN, an ordering $X_1, \ldots, X_n$ 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 the marginal distribution $P_{BN}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Your answer must be an explicit specification of the set of parents for each variable in BN'.

Problem 8: Towards inference in Bayesian networks (8 points)

1. (4 points) Suppose you have a Bayes' net over variables $X_1, \ldots, X_n$ and all variables except X_i are observed. Using the chain rule and Bayes' rule, find an efficient algorithm to compute $P(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n)$, in terms of local conditional distributions. In particular, your algorithm should not require evaluation of the full joint distribution.

2. (4 points) Find an efficient algorithm to generate random samples from the probability distribution defined by a Bayesian network. You can assume access to a routine that generates random samples from any given categorical distribution. Hint: it is possible to sample from any joint distribution $P(X, Y)$ by first drawing a sample $x \sim P(X)$ and then drawing a sample $y \sim P(Y|X = x)$. Hint: You may want to check out topological sorting.

Problem 9: Programming assignment (35 points)

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.

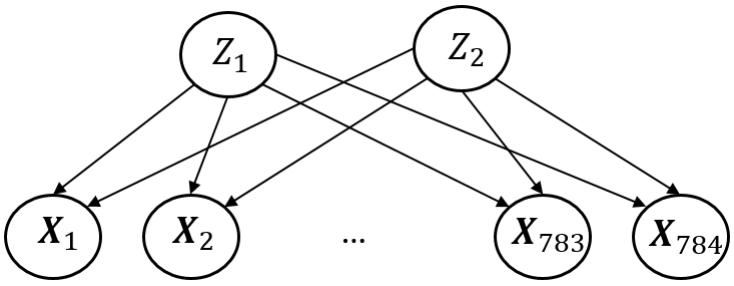


Figure 1: Bayesian network for the MNIST dataset. $X_{1:784}$ variables correspond to pixels in an image. Z_1 and Z_2 variables are latent.

The 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 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. We consider a model parameterized using neural networks, trained using stochastic gradient descent. Bayesian networks specified as such are popularly referred to as variational autoencoders and represent one of the most powerful existing deep generative models in current use. We will return to the exact details of learning such models later in the course.

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 $Val(Z_1) = Val(Z_2) = \{-3, -2.75, \ldots, 2.75, 3\}$ and $Val(X_j) = \{0, 1\}$, i.e., the image is binary.

Note: the index for X starts with 0 and ends with 783 in the Python starter code (which corresponds to $X_{1:784}$ in the problem description)

Note: You must include all plots in your writeup to be considered for full credit.

1. (2 points) How many values can the random vector $X_{1:784}$ take, i.e., how many different 28×28 binary images are there?

2. (2 points) How many parameters would you need to specify an arbitrary probability distribution over all possible 28×28 binary images?

3. (4 points) How many parameters do you need to specify the Bayesian network in Figure 1?

For parts 4-7 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.

4. (5 points) 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!

5. (5 points) For each possible value of

$$(\bar{z}_1, \bar{z}_2) \in \{-3, -2.75, \ldots, 2.75, 3\} \times \{-3, -2.75, \ldots, 2.75, 3\},$$

   compute the conditional expectation $\mathbb{E}[X_{1:784}|Z_1, Z_2 = (\bar{z}_1, \bar{z}_2)]$. 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?

6. (10 points) In `q6.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. Take extra caution: the variables are labeled from 1 to 784, not from 0 to 783, please read the comments in the code to avoid indexing errors.

7. (7 points) In `q7.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 $p((Z_1, Z_2) = (\bar{z}_1, \bar{z}_2) \mid X_{1:784} = I^k)$. Use these probabilities to compute the conditional expectation

$$\mathbb{E}[(Z_1, Z_2) \mid X_{1:784} = I^k)].$$

   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 5?