+

Homework 2 of CS 165A (Winter 2019)

University of California, Santa Barbara

Assigned on January 31, 2019 (Thursday)
Due at 12:30 pm on February 12, 2019 (Tuesday)

Notes:

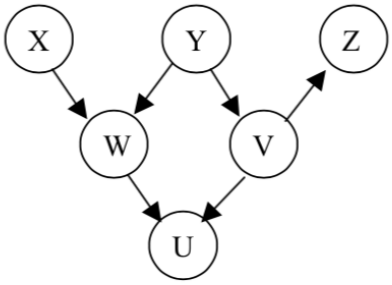
- Be sure to read "Policy on Academic Integrity" on the course syllabus.
- Any updates or correction will be posted on the course Announcements page and piazza, so check there occasionally.
- You must do your own work independently.
- Please typeset your answers and you must turn in a hard copy to the CS 165A homework box in the copy room of Harold Frank Hall before the due time or turn in at the beginning of due date's class.
- We also encourage you to submit a digital copy on the GauchoSpace for record purpose, we won't grade this.
- Keep your answers concise. In many cases, a few sentences are enough for each part of your answer.

Did you receive any help whatsover from anyone in solving this assignment?

Did you give any help whatsover to anyone in solving this assignment ?

Problem 1 (16')

Consider the Bayes Net below:



(a) Is it true that $P(X|Y,W) = P(X|W)$? Explain. (3')

(b) Write down the expression for computing $P(X|Y)$ using the above Bayes Net. (2')

(c) Are variables X,W conditionally independent of variables V,Z, given Y? Explain. (2')

(d) Are variables X,W conditionally independent of variables V,Z, given U? Explain.(2')

(e) Are variables W and Z independent? Explain.(2')

(f) Write down the Markov Blanket of variable W and variable Y. (2')

(g) Assume all the variables are binary, either take value 0 or 1. Write down the expression to compute $P(U = 1, V = 1, W = 1, X = 0, Y = 0, Z = 1)$ using notation like $P(X = 1|W = 0)$.(3')

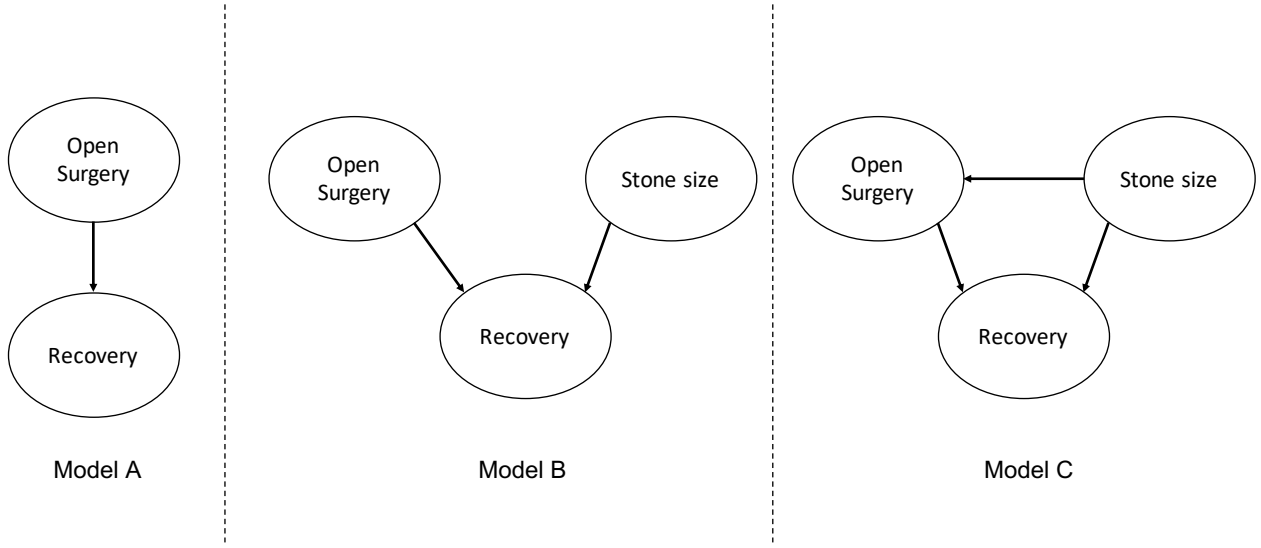
Problem 2 (19')

Consider the following data set about kidney stone treatments. There are two possible treatments: Open surgery and small puncture. A total of 350 patients are treated with open surgery and another 350 patients are treated with small punctures. The number of patients that recovered are given in the following table. Each patient can either have a small kidney stone or a large kidney stone.

Table 1: Kidney stone treatments

Context	Open surgery	Small puncture
Small Stone	**93% (81/87)**	87% (234/270)
Large Stone	**73% (192/263)**	69% (55/80)
All	78% (273/350)	**83% (289/350)**

Consider the following three different graphical models:



(a) Write down the CPTs for each of these three models and provide the estimated values of the CPTs using the provided data. (6')

(b) Using the data and each of the model, apply probabilistic inference to answer the question: Is Open Surgery better than Small Puncture in terms of the rate of recovery? Are the answers consistent across models? If not, why is it the case? (6')

(c) Enumerate all the marginal and conditional independences implied by Model B and Model C. (4')

(d) Which model between Model B and Model C, do you think, fits the data best? Explain the reasoning behind your answer. (Hint: using the definition of independences and the observed data.) (3')

Problem 3 (16')

Assume we have n data samples $x_1, x_2.....x_n \in \mathbb{R}^d$ and their corresponding label $y_1, ..., y_n \in \{1, 2, 3, ..., k\}$. Suppose we want to train Gaussian Naive Bayes classifer on this data set, which assumes that $(x_1, y_1), ..., (x_n, y_n)$ are drawn independently from the following generative process

$$\begin{aligned} y &\sim \text{Categorical}(\tau) \\ x[j] &\sim N(\mu_j, \sigma_j^2) \quad \text{independently} \quad \forall j = 1, 2, ..., d. \end{aligned}$$

$\tau, \mu_1, ..., \mu_d, \sigma_1, ..., \sigma_d$ are parameters of this model.

Write down your work step by step.

(a) Write down the likelihood function. (4')

(b) Rewrite it into log-likelihood format. (2')

(c) Derive the Maximum Likelihood Estimator (MLE) for τ, μ_j and σ_j (denote as $\hat{\tau}$, $\hat{\mu}_j$, $\hat{\sigma}_j$). (Hint: take the derivative and assign it to 0.) (6')

(d) Now we know that the prior distribution of mean μ_j itself follows another Gaussian Distribution with mean v and variance β^2. Derive the Maximum A Posteriori estimator $\hat{\mu}_j^{MAP}$ for μ_j. (Hint: Use logarithm.) (4')

Problem 4 (15')

Consider the following training data.

Text	Label
"A great game"	Sports
"The election was over"	Non-Sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Non-Sports

We want to use the Multinomial Naive Bayes algorithm to perform a simple text classification on the text "A very close game".

1. Build the vocabulary from training data. (2')

2. Denote the size of your "vocabulary" by d, the number of data points by b, and the number of classes by k. Write down the generative process (similar to the one in Problem 3) of the Multinomial Naive Bayes model with respect to the bag-of-word features. Derive the maximum likelihood estimator of these model parameters. (3')

3. Compute the likelihood of each word in the sentence to be classified, using Laplace Smoothing. Perform the necessary calculation and classify the text. (4')

4. Read about Dirichlet distribution on Wikipedia. Assign a Dirichlet prior distribution with parameter vector $[\alpha_1, \alpha_2, ..., \alpha_d]$ to the parameter of the Multinomial distribution, derive the associated MAP estimator. (4')

5. What is the connection between Laplace Smoothing and MAP estimator? (2')

Problem 5 (10')

Consider a set of 2D points: (1,3), (2,3), (1,4), (1,5), (1,6), (2,6), (3,3), (4,1), (4,2), (5,1), (5,2). We wish to cluster these points using k-means algorithm(k=2). Cluster means are initialized at (5,2) and (2,4). Show the partition result. And what will the new means be after first iteration? After second iteration?

Problem 6 (10')

Assume that the data point (x_t, y_t) comes in one at a time for $t = 1, 2, 3, ...$ We are limited by a memory constraint and cannot store (x_t, y_t) beyond time t.

Derive the Stochastic Gradient Descent algorithm for optimizing the log-likelihood function that you derived for the Gaussian Naive Bayes classifier (Problem 3) and Multinomial Naive Bayes (Problem 4).

Problem 7 (14')

Let the training data be $(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \{1, 2, ..., k\}$ and the test data $(x'_1, y'_1), ..., (x'_n, y'_n) \in \mathbb{R}^d \times \{1, 2, ..., k\}$ be drawn i.i.d. from the same distribution. Let $\{f_1, ..., f_H\}$ be a set of classifiers such that $f_h : \mathbb{R}^d \to \{1, 2, ..., k\}$

Apply **Hoeffding's inequality** and the **union bound** that we learned in the class to

derive a high probability upper bound of the generalization error:

$$\max_h \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f_h(x_i)==y_i\}} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f_h(x_i')==y_i'\}} \right|.$$

Write down a bound that holds with probability at least $1 - n^{-20}$. Make sure you explain your work step by step.