CIS 520, Machine Learning, Fall 2018: Assignment 7
Due: Wednesday, November 28th, 11:59pm
[100 points]

Instructions. Please write up your responses to the following problems clearly and concisely. We encourage you to write up your responses using LaTeX; we have provided a LaTeX template, available on Canvas, to make this easier. **Submit your answers in PDF form to GradeScope. We will not accept paper copies of the homework.**

Collaboration. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups up to size **two students.** However, each student must write down the solution **independently**, and without referring to written notes from the joint session. **In addition, each student must write on the problem set the names of the people with whom you collaborated.** You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

1 Hidden Markov Models [20 points]

On any given day, Alice is in one of two states: happy or sad. You do not know her internal state, but get to observe her activities in the evening. Each evening, she either sings, goes for a walk, or watches TV.

Alice's state on any day is random. Her state Z_1 on day 1 is equally likely to be happy or sad:

$$P(Z_1 = \mathsf{happy}) = 1/2\,.$$

Given her state Z_t on day t, her state Z_{t+1} on the next day is governed by the following probabilities (and is conditionally independent of her previous states and activities):

$$P(Z_{t+1} = \mathsf{happy} \,|\, Z_t = \mathsf{happy}) = 4/5 \qquad P(Z_{t+1} = \mathsf{happy} \,|\, Z_t = \mathsf{sad}) = 1/2\,.$$

Alice's activities are also random. Her activities vary based on her state; given her state Z_t on day t, her activity X_t on that day is governed by the following probabilities (and is conditionally independent of everything else):

$$\begin{array}{ll} P(X_t = \mathsf{sing} \,|\, Z_t = \mathsf{happy}) = 5/10 & P(X_t = \mathsf{sing} \,|\, Z_t = \mathsf{sad}) = 1/10 \\ P(X_t = \mathsf{walk} \,|\, Z_t = \mathsf{happy}) = 3/10 & P(X_t = \mathsf{walk} \,|\, Z_t = \mathsf{sad}) = 2/10 \\ P(X_t = \mathsf{TV} \,|\, Z_t = \mathsf{happy}) = 2/10 & P(X_t = \mathsf{TV} \,|\, Z_t = \mathsf{sad}) = 7/10\,. \end{array}$$

1. **[20 points]** Suppose you observe Alice singing on day 1 and watching TV on day 2, i.e. you observe $x_{1:2} = (\mathsf{sing}, \mathsf{TV})$. Find the joint probability of this observation sequence together with each possible

hidden state sequence that could be associated with it, i.e. find the four probabilities below. Show your calculations.

$$P\big(X_{1:2} = (\mathsf{sing}, \mathsf{TV}), Z_{1:2} = (\mathsf{happy}, \mathsf{happy})\big)$$
$$P\big(X_{1:2} = (\mathsf{sing}, \mathsf{TV}), Z_{1:2} = (\mathsf{happy}, \mathsf{sad})\big)$$
$$P\big(X_{1:2} = (\mathsf{sing}, \mathsf{TV}), Z_{1:2} = (\mathsf{sad}, \mathsf{happy})\big)$$
$$P\big(X_{1:2} = (\mathsf{sing}, \mathsf{TV}), Z_{1:2} = (\mathsf{sad}, \mathsf{sad})\big)$$

Based on these probabilities, what is the most likely hidden state sequence $z_{1:2}$? What is the individually most likely hidden state on day 2 ?

2 Missing Data [12 points]

In this question, you shall get first hand experience on working with missing data. The data being used is the same **breast cancer** data used in previous homework. There are three variations of the data set in the `data.mat` file:

1. The entire, complete data without missing values. (`Xtrain_full, Xtest_full`)
2. Data with 20% fields missing at random, the missing values are substituted with "NaN"s. (`Xtrain_random, Xtest_random`)
3. Data with 20% fields missing, but **not** at random, the missing values are the smallest 20% values and are also substituted with "NaN"s. (`Xtrain_nrandom, Xtest_nrandom`)

In this question, we will explore 2 common ways to deal with missing data.

1. Impute missing data with mean (of training data).
2. Add an additional column to specify whether a feature is missing or not.

Please use Matlab's built-in decision tree classifier (fitctree) with default setting for the predictions.

Your task is the following:

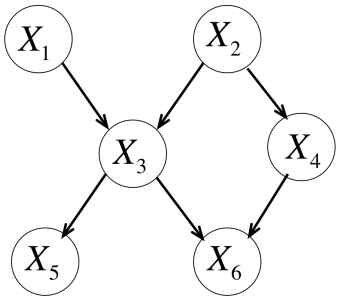
1. **[3 points]** Report test accuracy on the complete data set using decision trees.
2. **[3 points]** Report test accuracy on the data set with random NaNs using the two methods mentioned above.
3. **[3 points]** Report test accuracy on the data set with non-random NaNs using the two methods mentioned above.
4. **[3 points]** In one or two sentences explain the relative accuracies you found.

Do not include code snippet in your final pdf.

3 Bayesian Networks [20 points]

Consider the Bayesian network over 6 random variables $X_1, X_2, X_3, X_4, X_5, X_6$ shown below (assume for simplicity that each random variable takes 2 possible values):

1. **[2 points]** Write an expression for the joint probability mass function $p(x_1, x_2, x_3, x_4, x_5, x_6)$ that makes the same (conditional) independence assumptions as the Bayesian network above.



2. **[3 points]** Consider a joint probability distribution satisfying the following factorization:

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5 \mid x_3)p(x_6 \mid x_3)\,.$$

   Is this distribution included in the class of joint probability distributions that can be represented by the Bayesian network above? Briefly explain your answer.

3. **[3 points]** If the edge from X_3 to X_6 is removed from the above network, will the class of joint probability distributions that can be represented by the resulting Bayesian network be smaller or larger than that associated with the original network? Briefly explain your answer.

4. **[12 points]** Given the above figure, determine whether each of the following is true or false. Briefly justify your answer.

   (a) $p(x_3, x_4) = p(x_3)p(x_4)$

   (b) $p(x_1, x_4) = p(x_1)p(x_4)$

   (c) $p(x_1, x_2 \mid x_6) = p(x_1 \mid x_6)p(x_2 \mid x_6)$

   (d) $p(x_1, x_6 \mid x_3) = p(x_1 \mid x_3)p(x_6 \mid x_3)$

4 Belief Net Construction [20 points]

Given the following observed counts for all different combinations of the binary random variables A, B, C and D (each variable can be true (T) or false (F)), construct a belief net using the algorithm described in class, where variables are added sequentially to the network. Note that there are 4400 observations in total. *Consider the variables in the order A, B, C, D*, and make sure to give *both* the graph and the conditional probability tables. *Hint:*

A	B	C	D	Count
T	T	T	T	600
T	T	T	F	200
T	T	F	T	0
T	T	F	F	800
T	F	T	T	400
T	F	T	F	800
T	F	F	T	0
T	F	F	F	200
F	T	T	T	200
F	T	T	F	400
F	T	F	T	0
F	T	F	F	200
F	F	T	T	200
F	F	T	F	200
F	F	F	T	0
F	F	F	F	200

In this problem, you will be comparing many conditional probabilities to assess whether some variables are independent of others. In modeling situations like this we need some criteria for deciding whether one variable depends on the other based on the observed conditional probabilities. For this problem, let's suppose if the range of conditional probabilities for a set of variables is larger than 0.05, then the variables are dependent. (In reality, we might evaluate dependence by either i) doing a statistical significance test, or ii) regularizing the Bayes net by adding a cost for each new connection, like a BIC / MDL regularization).

Remember to check for both single and joint dependencies between variables. The key question is always to ask for each possible link that one might add to the graph: can it be removed?

5 EM [28 points]

We want to estimate the model parameters in the belief net

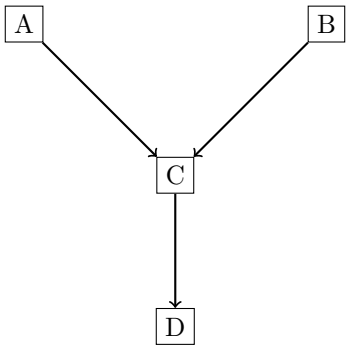


Figure 1: Belief Net

where A,B,C, and D are all Boolean but only A,B and D are observed. C is hidden.

1. **[4 points]** What are the model parameters to be learned to fully specify the model?
2. **[8 points]** What is the E step? Complete the equation.

$p(C_i = z \mid x_i,\ \theta^{t-1}) = \ldots$

3. **[8 points]** What is the M step? Complete the equation.

   $\theta^t = \arg\max_\theta \ldots$

4. **[8 points]** Now imagine the general case in which we have missing data we want to estimate. Instead of feature C being missing from every example, it might be that some features are missing for some examples and other features are missing for other examples. Describe how would you modify your approach from above to estimate the model parameters with missing data in the general case. Now specifically, suppose that C is missing the first $\frac{1}{2}$ the examples, D is missing from the next $\frac{1}{4}$ the examples, and both C and D are missing from the final $\frac{1}{4}$ of the examples. Give the equations for the E and M steps for this case.