HOMEWORK 5: PCA AND GRAPHICAL MODELS

10-701 Introduction to Machine Learning
(PhD) (Fall 2020)
Carnegie Mellon University
piazza.com/cmu/fall2020/10701/home
OUT: November, 4th, 2020*
DUE: November 16th, 2020 11:59 PM
TAs: Jie Jiao, Chandreyee Bhaumik

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the Academic Integrity Section on the course site for more information: https://www.cs.cmu.edu/~epxing/Class/10701-20/about.html#academic-integrity-policies
- **Late Submission Policy:** See the late submission policy here: https://www.cs.cmu.edu/~epxing/Class/10701-20/about.html#late-homework-policy
  NOTE: For this homework you are only allowed to use a maximum of 2 Grace Days
- **Submitting your work:**
  - **Gradescope:** There will be only one submission for this homework on Gradescope Written. The this homework only consists of problems such as short answer, multiple choice, derivations, proofs. Please use the provided template. The best way to format your homework is by using the Latex template released in the handout and writing your solutions in Latex. However submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Each derivation/proof should be completed in the boxes provided below the question, **you should not change the sizes of these boxes** as Gradescope is expecting your solved homework PDF to match the template on Gradescope. If you find you need more space than the box provides you should consider cutting your solution down to its relevant parts, if you see no way to do this it please add an additional page a the end of the homework and guide us there with a 'See page xx for the rest of the solution'.
    Regrade requests can be made after the homework grades are released, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted.

*Compiled on Thursday 5th November, 2020 at 01:07

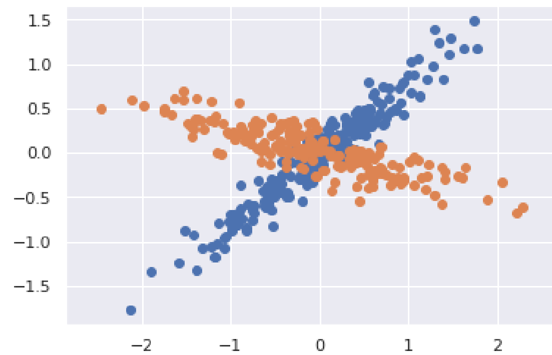
For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeXusers, use ■ and ●for shaded boxes and circles, and don't change anything else.

1 PCA [18 pt]

1.1 PCA Warm-Up [7 pt]

1. **[4 pt]** Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principle components on each plot.

(a)



2. **[2 pt]** Assume we are given a dataset for which the eigenvalues of the covariance matrix are: (2.1, 1.8, 1.3, 0.9, 0.4, 0.2, 0.15, 0.02, 0.001). What is the smallest value of K (dimension after reduction) we can use if we want to retain 75% of the variance (sum of all the variances in value) using the first principal components? Justify your answer.

3. **[1 pt] Select one:** Assume we apply PCA to a matrix $X \in \mathbb{R}^{N \times M}$ and obtain a set of PCA features, $X \in \mathbb{R}^{N \times M}$. We divide this set into two, Z1 and Z2. The first set, Z1, corresponds to the top principal components. The second set, Z2, corresponds to the remaining principal components.
Which is more common in the training data?
   - ◯ a point with large feature values in Z1 and small feature values in Z2
   - ◯ a point with small feature values in Z1 and small feature values in Z2
   - ◯ a point with large feature values in Z1 and large feature values in Z2
   - ◯ a point with small feature values in Z1 and large feature values in Z2

1.2 PCA and SVD [12 pt]

Given 6 data points in 5-d space, represented as rows in a 6 x 5 matrix X below:

$$X = \begin{bmatrix} -2 & -2 & -2 & 0 & 0 \\ -2 & -2 & -2 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

For all sub-questions, please show your derivation.

1. **[3 pt]** Write X in its SVD form.

(a) **[1 pt]** What's first Principal Component of the original data set?

(b) **[2 pt]** If we project the original data set onto 1-D space using the first Principal Component, what's the variance of the projected data?

(c) **[2 pt]** For the projected data in b, what is the reconstruction error?

2. **[4 pt]** In linear PCA, the covariance matrix of the data $C_x = X^T X$ is obtained by the weighted sum of its eigenvalues(λ) and eigenvectors($\mathbf{p}$):

$$C_x = \sum_i \lambda_i \mathbf{p}_i \mathbf{p}_i^T$$

Prove mathematically that the largest eigenvalue of C_x, λ_1 is equal to the variance of projecting the original data set onto the corresponding eigenvector $\mathbf{p}_1$.

2 Graphical Models [14 pts]

2.1 Conditional Independence Between Sets [5 pts]

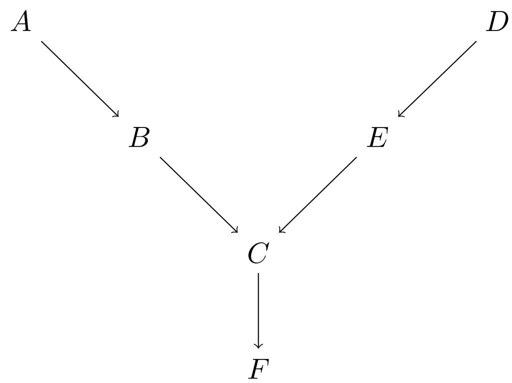


Figure 2.1: Graphical Model

1. **[1 point]** Specify two distinct (non-overlapping) sets Z such that $B \perp E|Z$ (in other words, B is independent of E given Z).

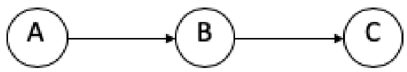
2. **[1 point]** Can you find another distinct set for Z (i.e. a set that does not intersect with any of the sets listed in 1)?

3. **[1 point]** What is the maximum number of distinct (non-overlapping) sets Z that you can find such that $A \perp E|Z$? What are they?

4. **[2 points]** If $W \perp X|Z$ and $X \perp Y|Z$ for some distinct variables W, X, Y, Z, can you say $W \perp Y|Z$? If so, show why. If not, find a counter example from the graph above.

2.2 Graphical Models and EM [9 pts]

1. Consider the graphical model below, consisting of three Boolean variables, and the associated training data:



5 training examples

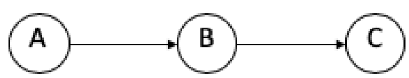
A	B	C
1	1	0
1	0	1
1	1	1
0	1	0
1	1	1

Figure 2.2: A Network, Plus training data.

   (a) **[3 points]** What probability will your trained network assign to $P(B = 1|A = 1, C = 1)$? Show the steps to your answer.

(b) **[2 points]** Is your Bayes Net estimate of $P(B = 1|A = 1, C = 1)$ the same or different from what you would obtain if you calculated the MLE of $P(B = 1|A = 1, C = 1)$ directly from the training data?

2. For this section, consider using EM to train the Bayes Net, using the data provided. **Note the value of B in training example 6 is not observed.**



6 training examples

A	B	C
1	1	0
1	0	1
1	1	1
0	1	0
1	1	1
1	?	1

Figure 2.3: A Network, Plus partially observed training data.

(a) **[2 points]** Consider using EM to train this network with this data. Furthermore, assume the network parameters are initialized not with random values, but instead using maximum likelihood estimate derived from the first five data points

During the first E step, precisely what quantities must be calculated? What are their values?

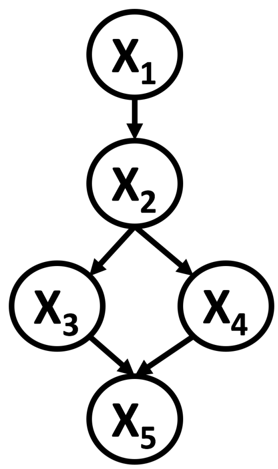
(b) **[2 points]** After completing this E step, the M step will re-estimate the entire set of network parameters. What will be the updated estimate for the parameter $P(B = 1|A = 1)$? Give your answer to 2 decimal places.

3 Graphical Models [12pts]

In the Kingdom of Westeros, Summer has come. Jon Snow, the King in the North, has taken the responsibility to defeat the Giants and protect the realm.

If Jon Snow can get Queen Cersei and Daenerys Queen of the Dragons to help him Jon is likely to beat the giants. Cersei and Daenerys are powerful women who are skeptical of the existence of Giants and will most likely only consider joining Jon if they are shown evidence of an imminent Giant attack. They can only be shown of an attack if Jon captures a live Giant.

The Bayesian network that represents the relationship between the events described above is shown below. Use the following notation for your variables: Jon Snow captures a live Giant (X_1), Jon shows Censei and Daenerys a live Giant (X_2), Cersei agrees to help (X_3), Daenerys agrees to help (X_4) and Giants defeated (X_5).



1. For the following questions fill in the blank with the smallest set $\mathcal{S}$ of random variables needed to be conditioned on in order for the independence assumption to hold. For example $X_i \perp X_j \mid \mathcal{S}$. What is the smallest set $\mathcal{S}$ that makes this statement true? The empty set $\emptyset$ is a valid answer, additionally if the independence assumption cannot be satisfied no matter what we condition on then your answer should be 'Not possible'.

   (a) **[1pt]** $X_1 \perp X_3 \mid$ $\boxed{}$

   (b) **[1pt]** $X_1 \perp X_5 \mid$ $\boxed{}$

   (c) **[1pt]** $X_2 \perp X_4 \mid$ $\boxed{}$

   (d) **[1pt]** $X_3 \perp X_4 \mid$ $\boxed{}$

(e) **[1pt]** $X_2 \perp X_5 \mid \boxed{}$

2. Jon gets his friend Sam to calculate some estimates of his chances. Sam returns to Jon with the following conditional probabilities tables:

$X_1 = 0$	0.3
$X_1 = 1$	0.7

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	0.8	0.25
$X_2 = 1$	0.2	0.75

	$X_2 = 0$	$X_2 = 1$
$X_3 = 0$	0.5	0.6
$X_3 = 1$	0.5	0.4

	$X_2 = 0$	$X_2 = 1$
$X_4 = 0$	0.3	0.2
$X_4 = 1$	0.7	0.8

	$X_3 = 0, X_4 = 0$	$X_3 = 0, X_4 = 1$	$X_3 = 1, X_4 = 0$	$X_4 = 1, X_3 = 1$
$X_5 = 0$	0.4	0.7	0.8	0.5
$X_5 = 1$	0.6	0.3	0.2	0.5

Table 3.1: Sam's Conditional Probability tables

Using the conditional probabilities for our graphical model, compute the following (Your answers should be given to 5 decimal places):

(a) **[2pts]** $P(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 1, X_5 = 0)$.

(b) **[5pts]** $P(X_1 = 1 | X_3 = 1)$

4 Markov Equivalence [18 pts]

Markov equivalence of two graphs is defined as follows: Let G1 and G2 be two Directed Acyclic Graphs (DAGs) on the same set of nodes U. We say that G1 and G2 are Markov equivalent iff for any three disjoint sets of nodes A, B and C from U we have $I_1(A, B|C)$ iff $I_2(A, B|C)$ where I_1 is conditional independence in G1 and I_2 is for G2.

Markov equivalence is an important concept since it can be used to show if two graphs represents the same conditional probability distributions. Please answer the following questions about Markov equivalent graphs:

1. **[5 pts]** Begin by proving the following lemma which will be useful for the other parts of this question: Prove that for two nodes X and Y in a DAG G, X and Y are adjacent (connected by an edge) if and only if they are not d-separated by some set in G (which could be the empty set)

2. **[8 pts]** For two non-adjacent variables X and Y in G, assume we have a chain in the graph such that $X - Z - Y$. Prove that the following are equivalent:

a. The structure of the chain is $X \rightarrow Z \leftarrow Y$

b. There exists a set not containing Z that d-separates X and Y.

c. All sets containing Z do not d-separate X and Y

3. **[3 pts]** Prove that if G1 and G2 are Markov equivalent, then X and Y are adjacent in G1 if and only if they are adjacent in G2.

4. **[2 pts]** Prove that if two DAGs G1 and G2 with the same set of nodes are Markov equivalent then they must have the same edges (though edges may have different directions in the two graphs) and for all nodes X and Y not connected by an edge, if we have a chain $X \to Z \leftarrow Y$ in G1 we must have a similar chain in G2 and vice versa.

5 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment? **Solution** Yes / No.

   (b) If you answered 'yes', give full details (e.g. "Jane Doe explained to me what is asked in Question 3.4")

   Solution

2. (a) Did you give any help whatsoever to anyone in solving this assignment? **Solution** Yes / No.

   (b) If you answered 'yes', give full details (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")

   Solution

3. (a) Did you find or come across code that implements any part of this assignment? **Solution** Yes / No.

   (b) If you answered 'yes', give full details (book & page, URL & location within the page, etc.).

   Solution