Homework 9
SVMs, K-Means, PCA, Graphical Models

CMU 10-601: Machine Learning (Spring 2019)
https://piazza.com/cmu/spring2019/1030110601
OUT: Wednesday, 24th April, 2019
DUE: Wednesday, 1st May, 2019, 11:59pm EDT
TAs: Chenxi, Loki, Eric, Gauri, Daniel

START HERE: Instructions

Homework 9 covers topics on SVMs, K-Means, PCA and Graphical Models. The homework includes multiple choice, True/False, and short answer questions.

- **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: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html
- **Submitting your work:**
  - **Gradescope:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using Gradescope (https://gradescope.com/). Please use the provided template. 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. Alternatively, submissions can be written in LaTeX. Regrade requests can be made, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed on a separate page. For short answer questions, you **should not** include your work in your solution. If you include your work in your solutions, your assignment may not be graded correctly by our AI assisted grader. In addition, please tag the problems to the corresponding pages when submitting your work.

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.

Instructions for Specific Problem Types

For “Select One” questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Matt Gormley

○ Marie Curie

~~●~~ Noam Chomsky

For “Select all that apply” questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

□ I don’t know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

~~■~~ I don’t know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601

10-~~7~~601

1 Support Vector Machines [19 pts]

In class, we discussed the properties and formulation of hard-margin SVMs, where we assume the decision boundary to be linear and attempt to find the hyperplane with the largest margin. Here, we introduce a new class of SVM called soft margin SVM, where we introduce the slack variables e_i to the optimization problem and relax the assumptions. The formulation of soft margin SVM with no Kernel is

$$\begin{aligned} \underset{\mathbf{w},b,e}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\left(\sum_{i=1}^{N} e_i\right) \\ \text{subject to} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - e_i,\ \forall\, i = 1, \dots, N \\ & e_i \geq 0,\ \forall\, i = 1, \dots, N \end{aligned}$$

1. **[3pts]** Consider the ith training example $(\mathbf{x}^{(i)}, y^{(i)})$ and its corresponding slack variable e_i. Assuming $C > 0$ and is fixed, what would happen as $e_i \to \infty$?

   Select all that apply:

   □ the constraint $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - e_i$ would hold for almost all $\mathbf{w}$.

   □ there would be no vector that satisfies the constraint $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - e_i$

   □ the objective function would approach infinity.

   With this in mind, we hope that you can see why soft margin SVM can be applied even when the data is not linearly separable.

2. **[5pts]** What **could** happen as $C \to \infty$? Do **not** assume that the data is linearly separable unless specified.

   Select all that apply:

   □ When the data is linearly separable, the solution to the soft margin SVM would converge to the solution of hard margin SVM.

   □ There is no solution $\mathbf{w}, b$ satisfying all the constraints in the optimization problem.

   □ Any arbitrary vector $\mathbf{w}$ and scalar b can satisfy the constraints in the optimization problem.

   □ The optimal weight vector would converge to the zero vector $\mathbf{0}$.

   □ When C approaches to infinity, it could help reduce overfitting.

3. **[5pts]** What **could** happen as $C \to 0$? Do **not** assume that the data is linearly separable unless specified.

   Select all that apply:

   □ When the data is linearly separable, the solution to the soft margin SVM would converge to the solution of hard margin SVM.

   □ There is no solution $\mathbf{w}, b$ satisfying all the constraints in the optimization problem.

   □ Any arbitrary vector $\mathbf{w}$ and scalar b can satisfy the constraints in the optimization problem.

   □ The optimal weight vector would converge to be the zero vector $\mathbf{0}$.

   □ When C approaches to 0, doing so could help reduce overfitting.

4. **[3pts]** An extension to soft margin SVM (or, an extension to the hard margin SVM we talked in class) is the 2-norm SVM with the following primal formulation

$$\begin{aligned} \underset{\mathbf{w},b,e}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\left(\sum_{i=1}^{N} e_i^2\right) \\ \text{subject to} \quad & y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - e_i, \ \forall\, i = 1, \dots, N \\ & e_i \geq 0, \ \forall\, i = 1, \dots N \end{aligned}$$

   Which of the following is true about the 2-norm SVM? (Hint: think about ℓ_1-regularization versus ℓ_2 regularization!)

   Select one:

   ○ If a particular pair of parameters $\mathbf{w}^*, b^*$ minimizes the objective function in soft margin SVM, then this pair of parameters is guaranteed to minimize the objective function in 2-norm SVM.

   ○ 2-norm SVM penalizes large e_i's more heavily than soft margin SVM.

   ○ One drawback of 2-norm SVM is that it cannot utilize the kernel trick.

   ○ None of the above.

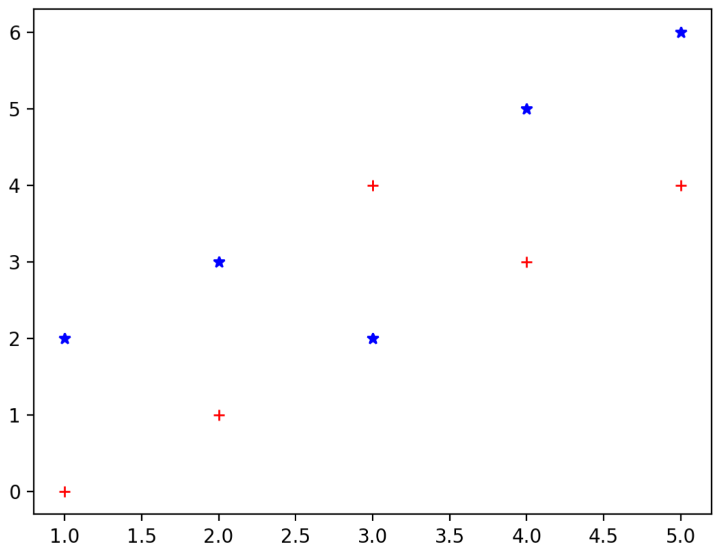


Figure 1: SVM dataset

5. **[3pts]** Consider the dataset shown in Figure 1. Which of the following models, when properly tuned, could correctly classify **ALL** the data points?

Select all that apply:

□ Logistic Regression without any kernel

□ Hard margin SVM without any kernel

□ Soft margin SVM without any kernel

□ Hard margin SVM with RBF Kernel

□ Soft margin SVM with RBF Kernel

2 Kernels [19pts]

1. **[2pt]** Consider the following kernel function:

$$K(x, x') = \begin{cases} 1, \text{ if } x = x' \\ 0, \text{ otherwise} \end{cases}$$

True or False: In this kernel space, any labeling of points from any training data X will be linearly separable.

○ True

○ False

2. **[3pts]** Suppose that input-space is two-dimensional, $x = (x_1, x_2)^T$. The feature mapping is defined as -

$$\phi(x) = (x_1^2, x_2^2, 1, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2)^T$$

What is the corresponding kernel function, i.e. $K(x, z)$? **Select one.**

○ $(x_1z_1)^2 + (x_2z_2)^2 + 1$

○ $(1 + x^Tz)^2$

○ $(x^Tz)^2$

○ x^Tz

3. **[3pts]** Suppose that input-space is three-dimensional, $x = (x_1, x_2, x_3)^T$. The feature mapping is defined as -

$$\phi(x) = (x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)^T$$

Suppose we want to compute the value of kernel function $K(x, z)$ on two vectors $x, z \in \mathbb{R}^3$. We want to check how many additions and multiplications are needed if you map the input vector to the feature space and then perform dot product on the mapped features. Report $\alpha + \beta$, where α is the number of multiplications and β is the number of additions.

Note: Multiplication/Addition with constants should also be included in the counts.

□

4. **[3pts]** Suppose that input-space is three-dimensional, $x = (x_1, x_2, x_3)^T$. The feature mapping is defined as -

$$\phi(x) = (x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)^T$$

Suppose we want to compute the value of kernel function $K(x, z)$ on two vectors $x, z \in \mathbb{R}^3$.We want to check how many additions and multiplications are needed if you do the computation through the kernel function you derived above. Report $\alpha + \beta$, where α is the number of multiplications and β is the number of additions.

Note: Multiplication/Addition with constants should also be included in the counts.

5. **[3pts]** Suppose one dataset contains four data points in $\mathbb{R}^1$ space, as shown in Figure 2

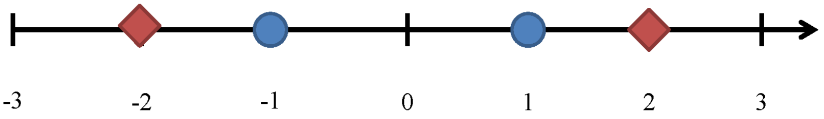


Figure 2: Data in $\mathbb{R}^1$

Different shapes of the points indicate different labels. If we train a linear classifier on the dataset, what is the lowest training error for a linear classifier on $\mathbb{R}^1$?

6. **[3pts]** Following the above question, which of the feature mappings below can we use to project the dataset to higher dimensional space such that training a linear classifier on the projected dataset would yield zero training error?

○ $\phi(x) = (x, 1)$

○ $\phi(x) = (x, x^3)$

○ $\phi(x) = (x, x^2)$

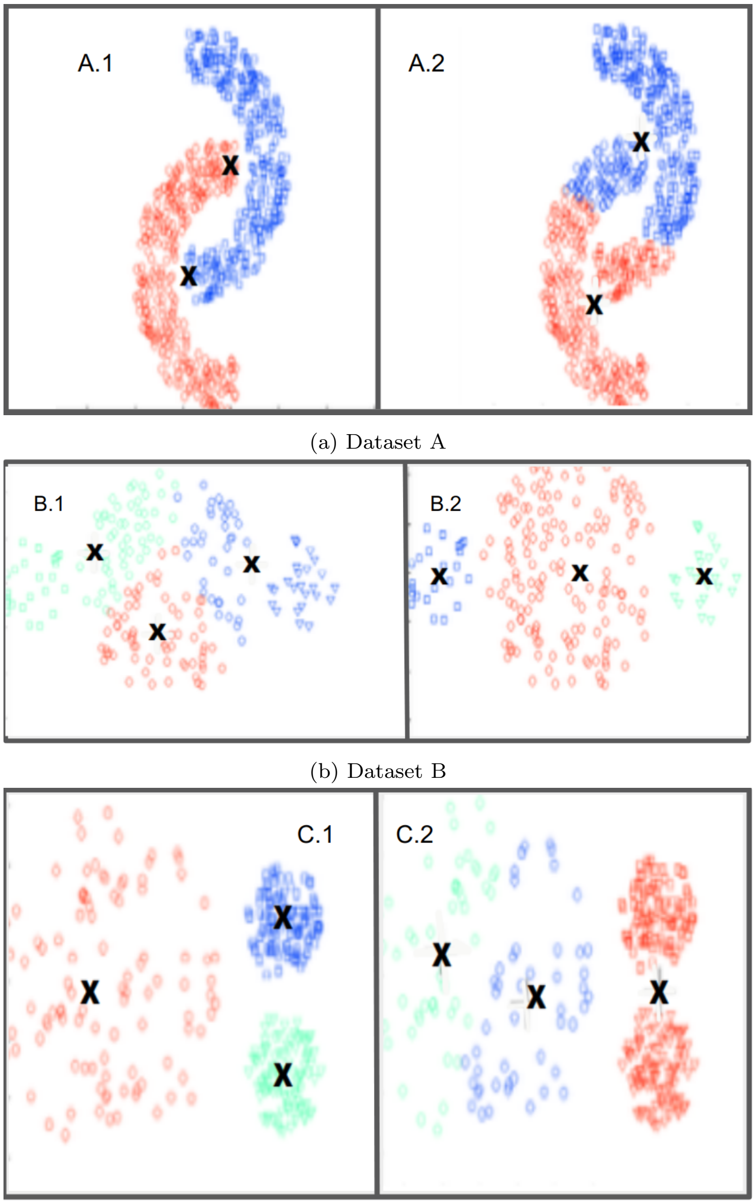
○ $\phi(x) = (x, (x + 1)^2)$

7. **[2pt]** **True or False:** Given the same training data, in which the points are linearly separable, the margin of the decision boundary produced by SVM will always be greater than or equal to the margin of the decision boundary produced by Perceptron.

○ True

○ False

3 K Means [19pts]



(a) Dataset A

(b) Dataset B

(c) Dataset C

Figure 3: Datasets

1. **[3pts]** Consider the 3 datasets A, B and C as shown in Figure 3. Each dataset is classified into k clusters as represented by different colors in the figure. For each dataset, determine which image with cluster centers (denoted by X) is generated by K-means method.The distance measure used here is the Euclidean distance.

   1.1. **[1pt]** Dataset A **(Select one)**

      ○ A.1

      ○ A.2

   1.2. **[1pt]** Dataset B **(Select one)**

      ○ B.1

      ○ B.2

   1.3. **[1pt]** Dataset C **(Select one)**

      ○ C.1

      ○ C.2

2. **[10pts]** Consider a Dataset **D** with 5 points as shown below. Perform a k-means clustering on this dataset with k as 2 using the Euclidean distance as the distance function. Remember that in the K-means algorithm, an iteration consists of performing following tasks: Assigning each data point to it's nearest cluster center followed by recomputation of those centers based on all the data points assigned to it. Initially, the 3 cluster centers are chosen randomly as $\mu 0 = (5.3, 3.5)$ (0), $\mu 1 = (5.1, 4.2)$ (1).

$$D = \begin{bmatrix} 5.5 & 3.1 \\ 5.1 & 4.8 \\ 6.6 & 3.0 \\ 5.5 & 4.6 \\ 6.8 & 3.8 \end{bmatrix}$$

   2.1. **[3pts]** Which of the following points will be the center for cluster 0 after the first iteration? **Select one:**

      ○ (5.7 , 4.1)

      ○ (5.6 , 4.8)

      ○ (6.3 , 3.3)

      ○ (6.7 , 3.4)

2.2. [**3pts**] Which of the following points will be the center for cluster 1 after the first iteration? **Select one:**

○ (6.1 , 3.8)

○ (5.5 , 4.6)

○ (5.4 , 4.7)

○ (5.3 , 4.7)

2.3. [**2pt**] How many points will belong to cluster 1 after the first iteration?

□

2.4. [**2pt**] How many points will belong to cluster 2 after the first iteration?

□

3. [**6pts**] Recall that in k-means clustering we attempt to find k cluster centers $c_j \in \mathbb{R}^d, j \in \{1, \ldots, k\}$ such that the total distance between each datapoint and the nearest cluster center is minimized. Then the objective function is,

$$\sum_{i=1}^{n} \min_{j \in \{1,\ldots,k\}} ||x_i - c_j||^2 \tag{1}$$

In other words, we attempt to find $c_1, \ldots, c_k$ that minimizes Eq. (1), where n is the number of data points. To do so, we iterate between assigning x_i to the nearest cluster center and updating each cluster center c_j to the average of all points assigned to the j th cluster. Instead of holding the number of clusters k fixed, your friend John tries to minimize Eq. (1) over k. Yet, you found this idea to be a bad one.

Specifically, you convinced John by providing two values α, the minimum possible value of Eq. (1), and β, the value of k when Eq. (1) is minimized.

3.1. [**3pts**] What is the value of $\alpha + \beta$ when $n = 100$?

□

3.2. [**3pts**] We want to see how k-means clustering works on a single dimension. Consider the case in which k = 3 and we have 4 data points $x_1 = 1, x_2 = 2, x_3 = 5, x_4 = 7$. What is the optimal value of the objective Eq. (1)?

□

4 PCA [13pts]

1. **[4pts]** Assume we are given a dataset X for which the eigenvalues of the covariance matrix are: (2.2, 1.7, 1.4, 0.8, 0.4, 0.2, 0.15, 0.02, 0.001). What is the smallest value of k we can use if we want to retain 75% of the variance (sum of all the variances in value) using the first k principal components?

□

2. **[3pts]** Assume we apply PCA to a matrix $X \in R^{n \times m}$ and obtain a set of PCA features, $Z \in R^{n \times m}$ .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: **Select one:**

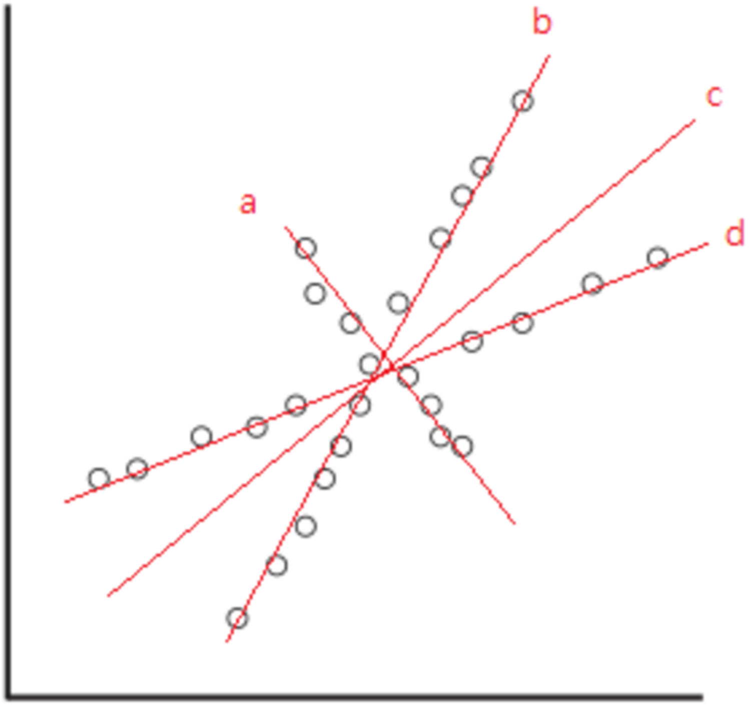
   ○ a point with large feature values in $Z1$ and small feature values in $Z2$

   ○ a point with large feature values in $Z2$ and small feature values in $Z1$

   ○ a point with large feature values in $Z2$ and large feature values in $Z1$

   ○ a point with small feature values in $Z2$ and small feature values in $Z1$

3. **[2pts]** For the data set shown below, what will be its first principal component?



Select one:

○ d

○ b

○ c

○ a

4. [**2pts**] **NOTE : This is continued from the previous question.** What is the second principal component in the figure from the previous question? **Select one:**

   ○ d

   ○ b

   ○ c

   ○ a

5. [**2pts**] **NOTE : This is continued from the previous question.** What is the third principal component in the figure from the previous question? **Select one:**

   ○ (a)

   ○ (b)

   ○ (c)

   ○ (d)

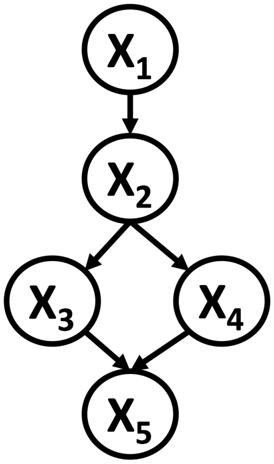
   ○ None of the above

5 Graphical Models [15pts]

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 the 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. **[1pt]** Write down the factorization of the above directed graphical model.

2. **[1pt]** Each random variable represented in the above Bayesian network is binary valued (i.e. either the event happens or it does not). State the minimum number of parameters you need to fully specify this Bayesian network.

3. **[1pt]** If we didn't use these conditional independence assumptions above, what would be the minimum number of parameters we would need to model any joint distribution over the same set of random variables?

4. **[5pts]** 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]** $X_1 \perp X_3 \mid$ □

   (b) **[1pt]** $X_1 \perp X_5 \mid$ □

   (c) **[1pt]** $X_2 \perp X_4 \mid$ □

   (d) **[1pt]** $X_3 \perp X_4 \mid$ □

   (e) **[1pt]** $X_2 \perp X_5 \mid$ □

5. **[7pts]** 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 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]**$P(X_1 = 1|X_3 = 1)$

Collaboration Questions Please answer the following:

After you have completed all other components of this assignment, report your answers to the collaboration policy questions detailed in the Academic Integrity Policies found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? Is so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? Is so, include full details.
3. Did you find or come across code that implements any part of this assignment ? If so, include full details.

Solution