Homework 8
PCA, SVMs, Graphical Models, Baum Welch, Ensemble Methods/Recommender Systems [1]

CMU 10-301/10-601: Machine Learning (Fall 2019)

https://piazza.com/cmu/fall2019/1030110601/
OUT: Sunday, Nov 24th, 2019

DUE: Wednesday, Dec 4th, 2019, 11:59pm
TAs: Brynn Edmunds, Yujia Chen, Lisa Hou, Ayushi Sood

START HERE: Instructions

Homework 8 covers topics on PCA, SVMs, Graphical Models, Baum Welch/Forward-Backward Algorithm, Ensemble Methods and Recommender Systems. 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#7-academic-integrity-policies
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/about.html#6-general-policies
- **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

[1] Compiled on Wednesday 4th December, 2019 at 23:36

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 Ensemble Methods [22pt]

1. **[3pts]** In the AdaBoost algorithm, if the final hypothesis makes no mistakes on the training data, which of the following is correct?

   Select all that apply:

   ■ Additional rounds of training can help reduce the errors made on unseen data.

   □ Additional rounds of training have no impact on unseen data.

   □ The individual weak learners also make zero error on the training data.

   □ Additional rounds of training always leads to worse performance on unseen data.

2. **[3pts]** Which of the following is true about ensemble method?

   Select all that apply:

   ■ Ensemble methods combine together many simple, poorly performing classifiers in order to produce a single, high quality classifier.

   ■ Neural networks can be used in the ensemble methods.

   □ For the weighted majority algorithm, the weak classifiers are learned along the way.

   ■ For the weighted majority algorithm, we want to give higher weights to better performing models.

3. **[2pt] True or False:** In AdaBoost weights of the misclassified examples go up by the same multiplicative factor.

   ● True

   ○ False

4. **[2pt] True or False:** AdaBoost will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

   ○ True

   ● False

5. **[12pts]** In the last semester, someone used AdaBoost to train some data and recorded all the weights throughout iterations but some entries in the table are not recognizable. Clever as you are, you decide to employ your knowledge of Adaboost to determine some of the missing information.

   Below, you can see part of table that was used in the problem set. There are columns for the Round # and for the weights of the six training points (A, B, C, D, E, and F)

Round	$D_t(A)$	$D_t(B)$	$D_t(C)$	$D_t(D)$	$D_t(E)$	$D_t(F)$
1	?	?	$\frac{1}{6}$	?	?	?
2	?	?	?	?	?	?
...						
219	?	?	?	?	?	?
220	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{7}{14}$	$\frac{1}{14}$	$\frac{2}{14}$	$\frac{2}{14}$
221	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{7}{20}$	$\frac{1}{20}$	$\frac{1}{4}$	$\frac{1}{10}$
...						
3017	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	0
...						
8888	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

at the start of each round. Some of the entries, marked with "?", are impossible for you to read.

In the following problems, you may assume that non-consecutive rows are independent of each other, and that a classifier with error less than $\frac{1}{2}$ was chosen at each step.

(a) **[3pts]** The weak classifier chosen in Round 1 correctly classified training points A, B, C, and E but misclassified training points D and F. What should the updated weights have been in the following round, Round 2? Please complete the form below.

Round	$D_2(A)$	$D_2(B)$	$D_2(C)$	$D_2(D)$	$D_2(E)$	$D_2(F)$
2	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$

(b) **[3pts]** During Round 219, which of the training points (A, B, C, D, E, F) must have been misclassified, in order to produce the updated weights shown at the start of Round 220? List all the points that were misclassified. If none were misclassified, write 'None'. If it can't be decided, write 'Not Sure' instead.

Not sure

(c) **[3pts]** During Round 220, which of the training points (A, B, C, D, E, F) must have been misclassified in order to produce the updated weights shown at the start of Round 221? List all the points that were misclassified. If none were misclassified, write 'None'. If it can't be decided, write 'Not Sure' instead.

A, B, E

(d) **[3pts]** You observes that the weights in round 3017 or 8888 (or both) cannot possibly be right. Which one is incorrect? Why? Please explain in one or two short sentences.

○ Round 3017 is incorrect.

○ Round 8888 is incorrect.

● Both rounds 3017 and 8888 are incorrect.

NOTE: Please do not change the size of the following text box, and keep your answer in it. Thank you!

Round 3017 is wrong because $D_{3017}(F) = 0$. This cannot happen since according to the update rule of $D_t(i)$, we have $D_{t+1}(i) = \frac{D_t(i)}{z_t} e^{+-a_t}$. Based on that, initializing $D_t(i) = \frac{1}{6} > 0 \ \forall i$, and $e^{+-a_t} > 0 \ \forall a_t$. Consequently, D_t can never be 0. Round 8888 is wrong because the weights should sum up to 1, and they don't.

2 Recommender Systems [10pt]

1. [**4pts**] In which of the following situations will a collaborative filtering system be the most appropriate learning algorithm compared to linear or logistic regression?

   Select all that apply:

   ■ You manage an online bookstore and you have the book ratings from many users. For each user, you want to recommend other books she will enjoy, based on her own ratings and the ratings of other users.

   ■ You run an online news aggregator, and for every user, you know some subset of articles that the user likes and some different subset that the user dislikes. You'd want to use this to find other articles that the user likes.

   □ You've written a piece of software that has downloaded news articles from many news websites. In your system, you also keep track of which articles you personally like vs. dislike, and the system also stores away features of these articles (e.g., word counts, name of author). Using this information, you want to build a system to try to find additional new articles that you personally will like.

   □ You manage an online bookstore and you have the book ratings from many users. You want to learn to predict the expected sales volume (number of books sold) as a function of the average rating of a book.

2. [**3pts**] What is the basic intuition behind matrix factorization?

   Select all that apply:

   □ That content filtering and collaborative filtering are just two different factorizations of the same rating matrix.

   ■ That factoring user and item matrices can partition the users and items into clusters that can be treated identically, reducing the complexity of making recommendations.

   □ The user-user and item-item correlations are more efficiently computed by factoring matrices.

   ■ That user-item relations can be well described in a low dimensional space that can be computed from the rating matrices.

3. **[3pts]** When building a recommender system using matrix factorization, the regularized objective function we wish to minimize is:

$$J(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \sum_{u,i \in \mathcal{Z}} (v_{ui} - \mathbf{w}_u^T \mathbf{h}_i)^2 + \lambda (\sum_u ||\mathbf{w}_u||^2 + \sum_i ||\mathbf{h}_i||^2)$$

where $\mathbf{w}_u$ is the uth row of $\mathbf{W}$ and the vector representing user u; $\mathbf{h}_i$ is the ith row of $\mathbf{H}$ and the vector representing item i; $\mathcal{Z}$ is the index set of observed user/item ratings in the training set; and λ is the weight of the L2 regularizer. One method of solving this optimization problem is to apply Block Coordinate Descent. The algorithms proceeds as shown below:

- while not converged:
  - for u in $\{1, \ldots, N_u\}$:
    - $\mathbf{w}_{u'} \leftarrow \text{argmin}_{\mathbf{w}_{u'}} J(\mathbf{W}, \mathbf{H})$
  - for i in $\{1, \ldots N_i\}$
    - $\mathbf{h}_{i'} \leftarrow \text{argmin}_{\mathbf{h}_{i'}} J(\mathbf{W}, \mathbf{H})$

Doing so yields an algorithm called Alternating Least Squares (ALS) for matrix factorization. Which of the following is equal to the *transpose* of $\text{argmin}_{\mathbf{w}_{u'}} J(\mathbf{W}, \mathbf{H})$?
Select one:

○ $v_u H (H^T H + \lambda I)^{-1}$

● $(H^T H + \lambda I)^{-T} v_u H$

○ $v_u H (H^T H + \lambda I)^{-T}$

○ $v_u H (H^T H)^{-1}$

3 Baum Welch [19pt]

In the following HMM questions, suppose the hidden states corresponding to a sequence of T observations $O = o_1, ..., o_T$ are $x_1, ..., x_T$, where each hidden states has N possible values. π_i represents the initial probablity (i.e. $P(x_0 = i)$). A represents transition probabilities where each component a_{ij} is the probability of transitioning from state i to state j (i.e., $= P(x_{t+1} = j|x_t = i)$). B represents the emission probabilities where each component $b_i(o_t)$ is the likelihood of o_t generated from a state i (i.e., $= P(o_t|x_t = i)$). $\lambda = (A, B)$ uniquely defines the HMM problem.
HMM is composed of three fundamental problems: **likelihood**, **decoding** and **learning**. The likelihood problem is to determine $P(O|\lambda) = P(o_1, ..., o_T|\lambda)$. The decoding problem is to discover the best hidden state sequence $X = x_1, ..., x_T$. The learning problem is to learn HMM parameters, and the following questions will guide you to solve each of the problems described above.

1. **[3pts]** Which of the following quantities are equivalent to **likelihood** $P(O|\lambda)$? Recall that we defined $\alpha_t(j) = P(o_1, o_2, ..., o_t, x_t = j|\lambda)$ and $\beta_t(i) = P(o_{t+1}, ..., o_T|x_t = i, \lambda)$.

   Select all that apply:

   ☐ $\sum_{i=1}^{N} \beta_0(i)\pi_i$

   ■ $\sum_{i=1}^{N} \beta_0(i)$

   ■ $\sum_{i=1}^{N} \alpha_T(i)$

   ☐ $\sum_{i=1}^{N} \alpha_T(i)\pi_i$

   ☐ $\sum_{i=1}^{N} \beta_T(i)$

   ☐ None of the above

2. **[4pts]** In this question, we explore the problems of **decoding** and **learning**. Which of the following statements are correct?

   Select all that apply:

   ■ The Viterbi algorithm is used for the decoding problem and has to know the hidden states in the training set in order to find the best sequence of hidden states for a sequence of testing observations.

   ■ The goal of the Baum Welch algorithm iteratively learn α_i's and β_i's.

   ☐ Baum Welch algorithm is guaranteed to learn the best set of parameters since we use maximum likelihood estimators.

   ■ Unlike SGD, the likelihood that the Baum Welch algorithm tries to maximize is never going to decrease throughout iterations.

   ☐ None of the above

3. **[12pts]** In this question, we will explore the **learning** problem by discussing the Baum Welch algorithm. The algorithm uses two intermediate variables ξ and γ.

$$\xi_t(i,j) = P(x_t = i, x_{t+1} = j|O,\lambda) = \frac{P(x_t = i, x_{t+1} = j, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

$$\gamma_t(i) = P(x_t = i|O,\lambda) = \frac{P(x_t = i, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$$

And the MLE of A (i.e., a_{ij}'s) and B (i.e., $b_j(o_t)$'s) can be found using ξ and γ.

(a) **[4pts]** The initialization of A and B will not affect the eventual HMM that we will learn. Justify your answer for no more than two sentences.
Select one:

○ True

● False

NOTE: Please do not change the size of the following text box, and keep your answer in it. Thank you!

The likelihood space is very non-convex. This means that for different values of A and B we will end up at different local maxima, hence the initial values matter.

(b) **[2pts]** $\sum_{t=0}^{T-1} \xi_t(i,j)$ is the expected number of times that a hidden state i is followed by a hidden state j in the observation O.
Select one:

○ True

● False

(c) **[2pts]** We can use $\xi^l(i,j)$ (i.e., the value of $\xi(i,j)$ in the l^{th} iteration) alone to find a_{ij}^{l+1} (i.e., the a_{ij} in the $l+1^{\text{th}}$ iteration).
Select one:

○ True

● False

(d) **[2pts]** In one iteration of Baum Welch, we first solve the **encoding** problem (i.e., finding the most probable sequence of hidden states) using the Viterbi algorithm in order to find the MLE for A and B.
Select one:

○ True

● False

(e) **[2pts]** Baum Welch is an unsupervised learning method while Viterbi decoding is a supervised learning method.
Select one:

● True

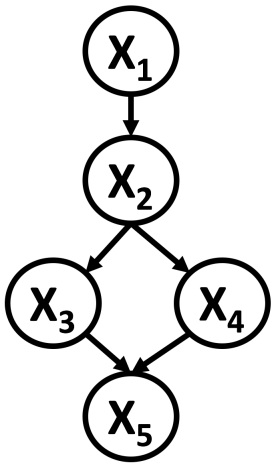
○ False

4 Graphical Models [24pts]

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. [**3pt**] Write down the factorization of the above directed graphical model.

$$P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_3,x_4)$$

2. [**3pt**] 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.

11

3. [**3pt**] 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?

31

4. **[10pts]** 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) **[2pt]** $X_1 \perp X_3 \mid$ x_2

   (b) **[2pt]** $X_1 \perp X_5 \mid$ x_2

   (c) **[2pt]** $X_2 \perp X_4 \mid$ Not possible

   (d) **[2pt]** $X_3 \perp X_4 \mid$ x_2

   (e) **[2pt]** $X_2 \perp X_5 \mid$ x_3, x_4

5. **[5pts]** 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]** $P(X_1 = 0, X_2 = 1, X_3 = 0, X_4 = 1, X_5 = 0)$.

0.02016

(b) **[3pts]** $P(X_1 = 1 | X_3 = 1)$

0.66111

5 Support Vector Machines [14 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

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

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 any $\mathbf{w}$ with finite entries.

   ☐ 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. **[4pts]** Which of the following are true when $C \to \infty$? Assume that the data is **not** linearly separable, unless otherwise 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 with a non-infinite objective value.

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

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

3. [**4pts**] Which of the following are true when $C \to 0$? Assume that the data is **not** linearly separable, unless otherwise 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 with a non-infinite objective value.

   ■ 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

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

   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.

6 PCA [11pts]

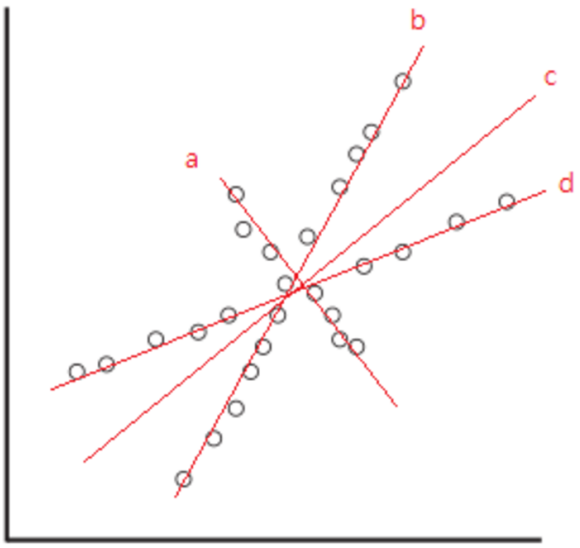
1. **[3pts]** Assume we are given a dataset X for which the eigenvalues of the covariance matrix are: (2.2, 1.7, 1.4, 0.8, 0.4, 0.2, 0.15, 0.02, 0.001). What is the smallest value of k we can use if we want to retain 75% of the variance (sum of all the variances in value) using the first k principal components?

3

2. **[2pts]** 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 $Z1$

Use the figure below to answer the following questions.



3. [**2pts**] 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

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

1. No
2. No
3. No