Spring 2022 Midterm Exam

PQHS 471 – Machine Learning & Data Mining

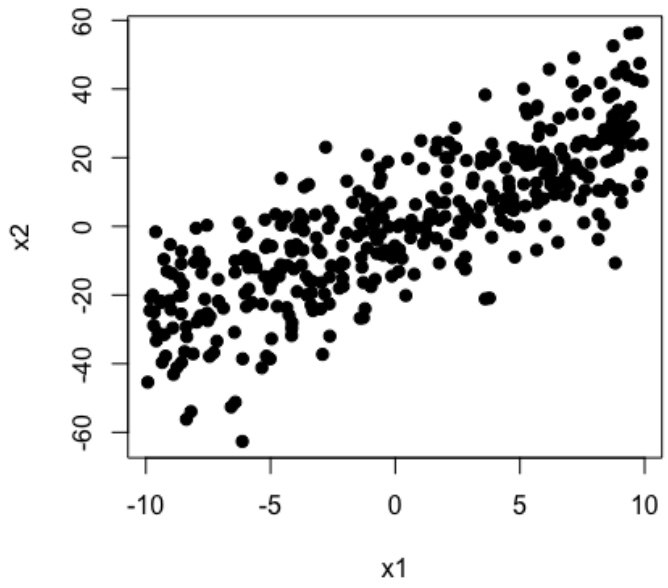
Your Name: _____

INSTRUCTIONS

1. This is a closed-book exam.
2. You may use a calculator.
3. Total points: 25.

Question 1. (3 points)

(i) Suppose you are working on a Principal Component Analysis(PCA) problem to get the PCs from a scaled, normalized dataset with variables x_1 and x_2. A scatterplot of x_1 and x_2 is given below. In the plot, roughly outline the directions of the 1st and 2nd PCs (with labels).



(ii) (*Fill in the blank*) Suppose the 1st PC can be represented as $a_1x_1+a_2x_2$. In PCA study, the elements in the coefficient vector (or eigenvector) of $\mathbf{a} = (a_1, a_2)^T$ is generally referred to as the _____.

(iii) What numerical constraint do we usually put on a_1 and a_2, to avoid the arbitrarily large elements in it?

Question 2. (2 points)

(i) In a PCA study(different from Question 1) with 5 variables, eigenvalues were obtained:

$$4, 2.5, 1.8, 1.2, 0.5$$

What's the proportion of total variance explained by the 1st PC? What's the proportion of total variance explained by the 2nd PC?

Question 3. (1 points)

(*True or False*)

Statement (i) is a stronger statement than statement (ii)

(i) Two random variables x_1 and x_2 have statistical independence.

(ii) Two random variables x_1 and x_2 have a 0 correlation.

Your T/F choice: _____.

Question 4. (3 points)

You are working on a research project to use non-negative matrix factorization (NMF) on a gene expression dataset. Human tumor samples were collected and measured for gene expression signals from a set of common 1,000 genes, for each sample. Samples from 60 individuals were obtained and measured. Your hypothesis is that for each individual, there are 5 underlying (unknown) cell types that made up of the observed expression signal. You want to use NMF to find out the underlying components, and the weights of each component, for each individual.

(i) Using the canonical form of $V = WH$, provide the dimensions of each of the three matrices in your project.

(ii) How would you interpret the column vectors in matrix W?

(iii) How would you interpret the 1st column vector in matrix H? How about the 3rd column vector in matrix H?

Question 5. (2 points)

(i) Name one major difference between LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) in their modeling assumptions.

(ii) From modeling flexibility perspective, how would you order (from most flexible to least) methods of logistic regression, kNN, LDA and QDA, in general?
Your answer:

_____ > _____ > _____ > _____

Question 6. (4 points)
Suppose there are 5 observations, for which we have already computed a similarity (distance) matrix as follows. Now, using this similarity matrix, sketch the dendogram that results from hierarchically clustering these 5 observations using the **complete linkage**. Show intermediate steps.

	x_1	x_2	x_3	x_4	x_5
x_1	0				
x_2	9	0			
x_3	3	7	0		
x_4	6	5	9	0	
x_5	11	10	2	8	0

Question 7. (5 points)
Below is a table showing the transaction IDs (TIDs) and the items bought from a store, in each transaction. Using the *Apriori* algorithm, with a **minimal absolute support of 2**, sketch the procedure to mine frequent itemsets. Show intermediate steps.

TID	**Items**
1	apple, candy, donut
2	banana, candy, egg
3	apple, banana, candy, egg
4	banana, egg

Question 8. (5 points)
Below is a table showing a dateset of biometrics predictors and the outcome of patients' prognosis after one month. Using Naive Bayes Classifier, make a prediction of the prognosis of this following patient:
(Protein = normal, HDL level = lack, BioE-Valve = wide, Diet = T)
Show intermediate steps.

Protein	HDL level	BioE-Valve	Diet	**Prognosis**
low	over	wide	F	poor
low	over	wide	T	poor
high	over	wide	F	good
normal	stable	wide	F	good
normal	lack	narrow	F	good
normal	lack	narrow	T	poor
high	lack	narrow	T	good
low	stable	wide	F	poor
low	lack	narrow	F	good
normal	stable	narrow	F	good
low	stable	narrow	T	good
high	stable	wide	T	good
high	over	narrow	F	good
normal	stable	wide	T	poor

END OF EXAM