DI501 INTRODUCTION TO DATA INFORMATICS
FINAL EXAM
PART 2
31 JANUARY 2022 (Duration: 90 minutes)

IMPORTANT:

In the final exam, you are not allowed to communicate with your classroom mates for any reason. You cannot discuss any matter with your friends during the exam. It is also strictly forbidden that you seek advice or feedback from any professionals in the domain, previous year students or any other. You are required to answer the questions alone. You are not allowed to give or receive any form of exam aid to any other student in this course during the exam. Any questions should be directed to the exam proctor via e-mail or Microsoft Teams course section.

If you need to scan your results on the paper, please use a professional application such as Office Lens with a good resolution. The image should only show your paper (not your desk, arm, fingers, pencils etc.) without any distortion. If your image does not satisfy these criteria and the document is not readable, you will get zero points. You need to crop the image if it's necessary.

Question 1 (25 pts):

Figure (a)

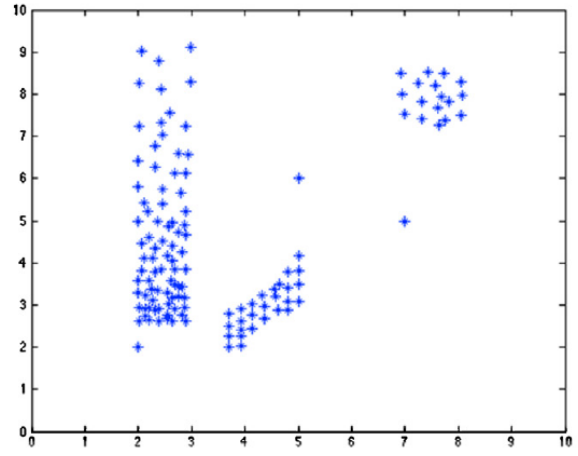
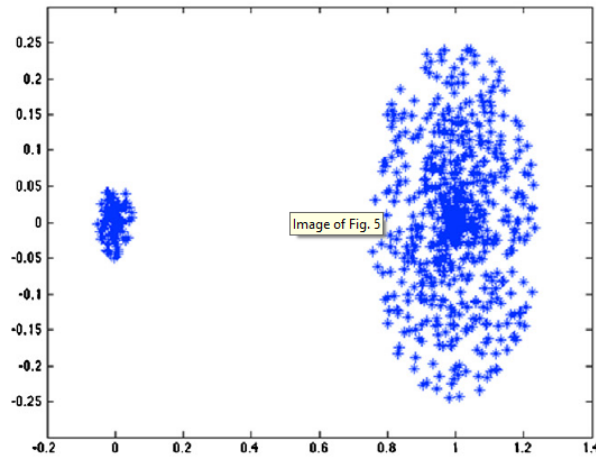


Figure (b)



Consider the two datasets shown in Figures while answering the following two (a and b) questions:

a) Compare approximately the average silhouette scores for k=3 and k=6 respectively using k-means algorithm (which one will most likely to be less than the other, you don't need to compute them) in Figure (a). Give your reasons. Show how the clusters might be formed on the images. Ensure to have two Figure (a) (copy and paste it) images to show clusters separately.

Note: Think whether outliers should be included in the clusters with the standard (vanilla) k-means algorithm. Moreover, think whether a cluster with one point cluster can be formed with this algorithm.

b) How many clusters approximately do you anticipate to have when eps= 0.2 and minpts=3 are chosen using density-based clustering in Figure (b)? What might happen when eps = 0.05 and minpts=3 are chosen?
c) Does silhouette analysis work successfully for all types of cluster shapes? Justify your answer. Give an example for demonstration.

Question 2 (20 pts):

a) Explain whether overfitting or underfitting is possible with cross-validation technique. Give your reasons.
b) Assume that you would like to predict the popularity of users given that you have a mixed type variable in your dataset. The target variable is highly skewed continuous variable.
    i. Discuss the consequences of model fitting using 90% uniformly random selected training dataset and the remaining 10% for your testing dataset. Give at least two consequences as examples.
    j. Discuss the consequences of model fitting using nested cross validation based on this dataset. Give at least one negative consequence as an example. Propose a solution for this problem.

Question 3 (10 pts):

Consider that our dataset includes the following input variables: V1 (binary), V2 (categorical with 3 possible values), V3 (binary). We have our Target (binary) variable as well.
We would like to run a naive Bayes classifier on this dataset.

a) How many parameters do we need to estimate to train our classifier?
b) How many parameters do we need if we don't have the naive Bayes conditional independence assumption?

Question 4 (10 pts):

Consider the following dataset where we classify the symptoms of Covid as (low (L), medium (M) or high (H)) when people received their first dose of vaccination and the second (True (T) or False (F)).

Symptom	First Dose V	Second Dose V
L	T	T
M	T	F
L	T	F
H	F	F
H	T	F
H	F	F
M	T	F
M	T	T
L	T	T

a) Draw a decision tree according to ID3 algorithm using the above dataset. You are not supposed to show the calculations.
b) Discuss whether strongly calculated variables and redundant attributes affect the model performance (such as accuracy) of decision trees. Give your reasons.