Istanbul Technical University- Fall 2018
BLG527E Machine Learning
Homework 2

Purpose: Multivariate Classification, Dimensionality reduction, Clustering.
Total worth: 5% of your grade.
Handed out: Friday, October 19, 2018.
Due: Nov 14, 2018, 24:00. (through ninova!)
Instructor: Zehra Cataltepe (cataltepe@itu.edu.tr),
Responsible Assistant: Fulya Celebi Sarioglu (sarioglu16@itu.edu.tr),
Policy: Collaboration in the form of discussions is acceptable, but you should write your own answer/code by yourself. Cheating is highly discouraged for it could mean a zero or negative grade from the homework.
If a question is not clear, please let us know (via email, during office hour or in class).
Submission Instructions: Please submit through the class ninova site.
Please zip and upload all your files using filename studentID_HW2.zip. You must provide all functions you wrote with your zipped file. Functions you do not submit may cause you lose a portion of your grade. You must also include a .doc or pdf file with answers to the questions and how to call your python/matlab/R functions for each question so that we can run and check the results.

PLEASE FILL, COPY AND PASTE THIS TABLE ON THE FIRST PAGE OF ANSWERS FOR YOUR HOMEWORK.

Question	Q1	Q2a	Q2b	Q3	TOTAL
MaxGrade	1.5	1	1	1.5	5
ExpectedGrade					

QUESTIONS:
Dataset:
Optdigits data by Alpaydin and Kaynak, from UCI Machine Learning Repository:
ftp://ftp.ics.uci.edu/pub/ml-repos/machine-learning-databases/optdigits/
You need the files:

optdigits.names	explanation of data
optdigits.tra	training data
optdigits.tes	test data

Q1) [1.5 points] [Multivariate Analysis]
Assume that each class i's (i=0..9) inputs are distributed according to a normal with mean μ_i and a common covariance matrix Σ.

Implement g(x) discriminant function (i.e. the classifier) that determines the class of an input x. Write its formula in your report.
Report the training and test errors and confusion matrices (See Chapter 19) of your classifier, g(x).

Hint1: First of all, you should remove the features which have 0 variance.
Hint2 : You can use built-in functions for mean and variance calculations.

Q2) [2 points] [Dimensionality Reduction]
Reduce dimensionality of the optdigits dataset by using:
Q2a) PCA (Principle Component Analysis) on the training data.
Q2b) LDA (Linear Discriminant Analysis) on the training data.
For both PCA and LDA, plot the training and test data reduced to two dimensions in four separate plots, indicate each class with a different colo.
Measure the test error of dimensionality reduction as follows: for each instance, find the closest neighbor in the reduced space. If the label of the instance and the closest neighbor are different, then the instance is classified incorrectly.

Q3) [1.5 points] [Clustering]
Cluster the training dataset using k-means clustering with k=20 clusters and the Euclidean distance to measure distance between instances.
Measure the error of your clustering as follows; compute the majority label for each cluster, for each instance if the its label of the instance is not same as the cluster label, then it is classified wrong.
Repat clustering and error measurement using L1 norm (absolute value of the difference) as the distance measure.

Hint3: You can convert a distance measure dist(u,v) into a similarity measure as follows: sim(u,v)=1/(1+dist(u,v)) or sim(u,v)=exp(-dist(u,v))