Istanbul Technical University- Spring 2018
BLG527E Machine Learning
Homework 1

Purpose: Better understanding of PR/ML basics.
Total worth: 5% of your grade.
Handed out: Thu, March 2, 2018.
Due: Sun, Mar 18, 2017 10:00pm. (through ninova!)
Instructor: Zehra Cataltepe (cataltepe@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, subject: blg527e hw1).

Submission Instructions: Please submit through the class ninova site.
Please zip and upload all your files using filename studentID_HW1.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** functions for each question so that we can run and check the results.
Output format:
Results for each subquestion (i.e. Q1a, Q1b,…) must be on a separate page

QUESTIONS:
Q1: [2 points] [Bayesian Decision Theory]
Assume a discriminant function of the form: $g_i(x) = \ln p(x \mid w_i) + \ln P(w_i)$
Assume that $x \in R$, for class 1: $x \sim N(\mu_1, \sigma)$ and for class 2: $x \sim N(\mu_2, \sigma)$. Also assume that $P(w_1) = P(w_2)$.

a) [0.5 points] For $\mu_1 = -5$, $\mu_2 = -10$, and $\sigma = 5$, plot the pdf's of the two classes inputs and also the separating surface.
b) [0.5 points] If $P(w_1)$ increases to 0.8, where would the separating surface be?
c) [1 points] Generate random datasets for a) and b), compute the estimates for the mean and plot histograms and the computed separating points on the histograms.

Dataset for Q2 and Q3:
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

Q2) [1 points] [Multivariate Analysis]
Assume that each class i's (i=0..9) inputs are distributed according to a normal with mean μ_i and covariance matrix Σ_i. Also assume that $\Sigma_i = \Sigma$ (common covariance matrix) and Σ is diagonal.
Implement g(x) discriminant function clearly (add comments) in your code and also write its formula into the report.
Compute and report the confusion matrix, error for each class and overall error on both training and test datasets.
Hint1: *Remove the features with 0 variance.*
Hint2:*You can use built-in functions for mean and covariance/variance calculations.*

Q3) [2 points] [Dimensionality Reduction]
Q3a) [1 points] Project the instances in optdigits.tra dataset using the LDA algorithm into a two dimensional space as shown in the figure below.
Q3b) [0.5 points] Using the parameters of the LDA transformation you computed for the optdigits.tra dataset, show the two dimensional projection of optdigits.tes dataset in another plot.
Q3c) [0.5 points] Assume that for each class i, $\Sigma_i = \Sigma$ (common covariance matrix) and Σ is diagonal. Recompute and replot the LDA projection. Compare to the results in Q3a) and Q3b).
Hint1: *do not re-compute the LDA parameters for the optdigits.tes set, just reuse the ones you computed for the optdigits.tra set.*
Hint2: *Do not use a built-in LDA function, implement it yourself.*