PATTERN RECOGNITION -- Spring 2019

Assignment 1: Discriminant Functions

DUE: Before 12midnight on 27 Feb 2019 (Wednesday)

INSTRUCTIONS:

i. Please do the assignment in Python.

ii. You need to submit pdf files to the TAs. One file should contain your answers, results and analysis. A separate file should contain code you have written and its sample output.

iii. At the top-right of the first page of your submission, include the assignment number, your name and roll number.

iv. *IMPORTANT:* Make sure that the assignment that you submit is your own work. *Do not directly copy any part from any source* including your friends, seniors or the internet.

v. Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

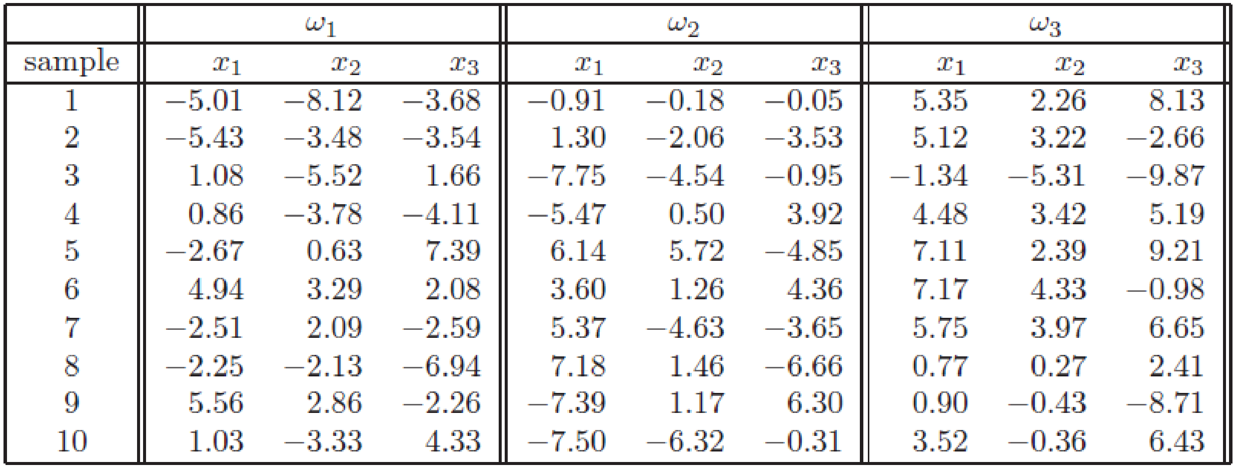
Preamble:

The idea of this assignment is to explore the material presented on Discriminant Functions (Linear and Quadratic).

Practical Exercises:

Use the following two data sets for the given exercises:

i) Synthetic Data (from DHS, Chapter 2):



	ω_1			ω_2			ω_3		
sample	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	−5.01	−8.12	−3.68	−0.91	−0.18	−0.05	5.35	2.26	8.13
2	−5.43	−3.48	−3.54	1.30	−2.06	−3.53	5.12	3.22	−2.66
3	1.08	−5.52	1.66	−7.75	−4.54	−0.95	−1.34	−5.31	−9.87
4	0.86	−3.78	−4.11	−5.47	0.50	3.92	4.48	3.42	5.19
5	−2.67	0.63	7.39	6.14	5.72	−4.85	7.11	2.39	9.21
6	4.94	3.29	2.08	3.60	1.26	4.36	7.17	4.33	−0.98
7	−2.51	2.09	−2.59	5.37	−4.63	−3.65	5.75	3.97	6.65
8	−2.25	−2.13	−6.94	7.18	1.46	−6.66	0.77	0.27	2.41
9	5.56	2.86	−2.26	−7.39	1.17	6.30	0.90	−0.43	−8.71
10	1.03	−3.33	4.33	−7.50	−6.32	−0.31	3.52	−0.36	6.43

ii) IRIS dataset (from the UCI ML benchmark dataset repository)

1. Implement the following general procedures that would be useful for several exercises. Demonstrate that they work with an example of your own.
   (a) Write a procedure to generate random samples according to a normal distribution N($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$) in *d* dimensions.
   (b) Write a procedure to calculate the discriminant function of the form given below (see equation in Section 2.6 DHS) for a given normal distribution and prior probability *P(ω_i)*:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

   (c) Write a procedure to calculate the *Euclidean distance* between two arbitrary points.
   (d) Write a procedure to calculate the *Mahalanobis distance* between the mean $\boldsymbol{\mu}$ and an arbitrary point **x**, given the covariance matrix $\boldsymbol{\Sigma}$.
2. Using the procedure implemented in 1(b) above, let us attempt to classify the samples in the synthetic dataset table given above. Assume that the synthetic data samples are generated from a normal distribution.
   a. Univariate exercise (with two-category dichotomizer): Assume that the prior probabilities for the first two categories are equal (*P(ω_1) = P(ω_2) = 0.5*). Design a dichotomizer for classes 1 and 2 using only one feature, namely x_1 (the first feature).
   b. Determine the empirical training error on the 10 samples (defined here as the percentage of points misclassified).
   c. Multivariate exercise (with two-category dichotomizer):Repeat (a) and (b) with two features, x_1 and x_2.
   d. Multivariate exercise (with two-category dichotomizer):Repeat (a) and (b) with three features, x_1, x_2 and x_3.
   e. Consider now data from all the three categories and assume uniform prior probability (*P(ω_i) = 1/3*). Now classify the following four test points using Mahalanobis distance between the test point and each of the category means: $(1, 2, 1)^t$; $(5, 3, 2)^t$; $(0, 0, 0)^t$; $(1, 0, 0)^t$
   f. Repeat (e) with unequal priors: *P(ω_1) = 0.8; P(ω_2) = P(ω_3) = 0.1. What are the class labels for the test points now? Comment on the change in class labels now, if any, as compared to those found in (e).*
3. Construct linear (LDF) and quadratic discriminant functions (QDF) for the IRIS dataset. Remember that linear discriminant function corresponds to Cases I (take average variance as the common variance) and II (take the average covariance matrix as the common covariance matrix) and QDF corresponds to Case III.

As part of the submission include the code for each of the algorithms along with a small report that explains the algorithms, implementation details, the results and their analysis.