ECE 485:
Assignment #2

Answer each of the following questions referring to the data sets Data1, Data2, and Data3 that are available on the ECE 485 web site under the Assignment's tab.

You are encouraged to use Matlab in doing the ECE 485 assignments (use 'help stats' or 'doc stats' at the Matlab command prompt to get the details on Matlab's available statistical functions).

1. Using the χ^2 goodness-of-fit test determine whether the data set contained in the files Data1 and Data2 can be reasonably modeled by a Gaussian distribution at an $\alpha = 0.05$ confidence level.

   (a) Plot the histogram for each data set.

   (b) Overlay the best fit Gaussian on this histogram plot.

   (c) Provide the results of the χ^2 goodness-of-fit test for each data sets, inclusive of its computed p-value.

   (d) For the data set(s) that fail the χ^2 goodness-of-fit test for Gaussian $p(x)$ determine which $p(x)$ distribution does provide a reasonable model for the data.

   - i.e., use the χ^2 goodness-of-fit test to determine statistically what other distribution could be used to model the data.
   - Hint: the shape of the data's histogram can provide a a good indication of which analytical $p(x)$'s you should test.

2. Write a function to generate N random data samples with mean $\boldsymbol{\mu} = [\mu_{x_1}\mu_{x_2}]^T$ and $\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1}\sigma_{x_2} \\ \sigma_{x_2}\sigma_{x_1} & \sigma_{x_2}^2 \end{bmatrix}$

   (a) Use this function to generate three sets of data all with $N = 1000$, $\boldsymbol{\mu} = [5, 5]^T$, with $\rho = -0.8$, $\rho = 0.2$ and $\rho = 0.9$, $\sigma_{x_1} = 2$ and $\sigma_{x_2} = 1$.

   (b) Produce a 2-D scatter plot for each generated data set.

   (c) On these scatter plots overlay the eigenvectors of Σ and draw the 1-, 2-, and 3-σ ellipses that are associated with the generated data.

3. The data in the file Data3 belongs to 3 classes, with the third column of the file denoting which class each data item belongs to.

   (a) For each class estimate its mean and covariance, assuming that the data within each class follows $p(x) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

   (b) For each of the following unclassified data point use the Mahalanobis distance to assigned each of the unclassified data points x_1, x_2, x_3 and x_4 to one of the three classes.

   $$x_1 = [10, 2], \quad x_2 = [-3, 4], \quad x_3 = [2, 2], \quad \text{and } x_4 = [5 - 7]$$

   (c) Provide scatter plots of each of the three clusters (all on the same plot but in different colours), place the x_1 through x_4 data points on this plot, and use your eigenvector and ellipse drawing routing from Question 2 to draw the principal axes and 1-, 2-, and 3-σ ellipses associated with each pattern class.

   (d) How would your estimations of the three classes' statistics change if you did not have *a priori* knowledge of the class labels?

   - i.e., If you were given the file Data3 with just its first 2 columns and not its third column.
   - This denotes the distinction between supervised learning (with column 3) and unsupervised learning (without column 3).

   (e) For each pair of classes plot (on the same plot as generated from 3(c) above) the 2-Class decision boundaries defined by where $p_{x_i}(x) = p_{x_j}(x)$, for $i \neq j$ and $i, j \in \{1, 2, 3\}$ and provide the formulas for these decisions boundaries.