In-class Assignment # 2

Due: Thursday, February 15, 2018, 11:59 p.m.
Total Points: 50

This assignment is designed to help you better understand concepts that were presented during class. You must complete this assignment on your own, but feel free to ask questions to your classmates and the instructor. **Each student is responsible for submitting their own solutions**. Be sure to include your name in your work. Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct.

Your assignment must be submitted as a single pdf document on Canvas. For each question, your submission should include text for the question, your MATLAB results/answers, and your MATLAB code. The questions and your answers must be typed; for example, in Latex or Microsoft Word. Images may be scanned and inserted into the document if it is too complicated to draw them properly. Be sure to show all work. **All assignments must be submitted on time to receive credit.** No late work will be accepted, unless you have a prior arrangement with the instructor.

Question 1. [25 POINTS]

There is an Excel file called in_class2_data.csv that contains two columns of data. The data in the first two columns are random samples of data from two different distributions, respectively. The objective of this problem is to use parameter estimation to determine the unknown parameters.

Create a script file called inclass2.m. Perform the following steps within this script file.

(a) Use the csvread.m function to read the data from the Excel file into a matrix. Be sure that the columns of data in the matrix, match the columns of data in the Excel file.

(b) The distribution for the data in the first two columns is unknown. Generate separate histogram plots, one for the data in the first column and the other for the data in the second column. Based on the histograms, what family of distribution(s) best describes the two data sets? Be sure to include this answer in your writeup or just display it to the screen.

(c) Using the distribution from the previous part, estimate the expected values (e.g. E[X]) for the data in the first two columns using maximum likelihood (ML) estimation. Look at the slides on parameter estimation for details. Note that there should be an expected value for the first column of data, and a separate expected value for the second column of data. Display the expected values to the screen.

(d) In a similar fashion, estimate the unknown variances (e.g. $Var[X]$) for the data in the first two columns using maximum likelihood (ML) estimation. Look at the slides on parameter estimation for details. Use the estimated expected values from above to estimate variance. Note that there should be a variance for the first column of data, and a separate variance for the second column of data. Display the variances to the screen.

Question 2. [25 POINTS]

We will continue to use the Excel file named in_class2_data.csv and the prior results for this problem. The first column represents data from hypothesis H_0, whereas the second column represents data from H_1. The parameters (mean and variance) that were computed in the previous problem represent parameters for the likelihood distribution of x given H_0 (e.g. $P(x|H_0)$) and of x given H_1 (e.g. $P(x|H_1)$), respectively. The objective of this problem is to perform Bayesian classification.

In the same script file that was created for the previous problem, perform the following steps:

(a) For each data sample from the first column, compute the likelihood that it was generated from H_0 and from H_1. Hence you should have two probability values for each sample from column one. Hint: write a function that computes the probability from the given data sample, mean, and variance. Use the distribution family that was indicated in question 1(b).

(b) Repeat the above step, but use data from the second column of the Excel file.

(c) Assuming uniform costs and equal priors, compute the likelihood ratio and compare it to the likelihood ratio test, in order to classify each data sample from each column. Store your decision (0 or 1) for each data sample of column one in a variable, and likewise store your decision for the column two data samples in a separate variable. Hint: Use the '>' operator to output 1 when the left-hand side is greater than the right-hand side, and 0 otherwise.

(d) Compute the decision error (e.g. number of incorrectly classified points divided by the total number of points) for each column, considering that the first column of data should be classified as H_0 (e.g. 0) and the second column of data should be classified as H_1 (e.g. 1). Display these scores to the screen. Hint: It may help to use logic operators (e.g. '==' or '$\sim$=').

(e) Use the decision errors from each column to compute the probability of error (e.g. P(E)). Display this error to the screen. Hint: assume equal priors.