COEN 281, Homework 1
Due: Thursday, October 18

Please turn in a paper copy in class. Email should only be used as a last resource. Work turned in d days late is graded and the grade is multiplied by (1 – d/10) if d ≤ 5, and 0 otherwise.

Work is to be done in groups of 2. Partner will be assigned randomly for each project. You must submit a confidential 1-to-5 (1=Poor; 5=Good) rating of your partner's contribution to the project. This rating will make 15% of the project's grade. Students with an average rating of 3 or below at the end of the quarter will have to submit himself/herself to a final exam. Please send an email to the instructor with the subject "HW1 – Group #," and the name of your partner and grade in the message body.

Data and R. The goal in this section is to get used to downloading data and working with R to inspect and explore it.

1. Exercise 8 from Chapter 2 of the class textbook. Show all your R code and output. For part (c)-vi, provide a brief summary of your answers to:

- What is the university with the most students in the top 10% of class?
- What university has the smallest acceptance rate?

2. The dataset "housetype.data" represents an extract from a commercial marketing database created from questionnaires filled out by shopping mall customers in the San Francisco Bay area. Report the commands you use to:

A. Read the data into a data frame. Be sure to keep the row and column names around. Show the dimensions of the data matrix and the upper 5x5 submatrix.

B. Write a function attributeHist that takes the name of an attribute, such as "age", finds the corresponding column in the table, and produces the histogram. By default, it should put the long attribute name into the title and on the horizontal axis. If the attribute contains "missing values" (represented as NA in the data), a message with the missing count should be printed. Specifically,
   a. attributeHist("age") should produce a histogram of the values for this attribute.
   b. attributeHist("hello") should print a message pointing out that here is no such attribute.
   c. attributeHist("eth") should display the histogram and print a message "61 missing values."

   Turn in the source code for your function as well as the output on the three calls above.

3. Relationship between classifier performance and number of observations and number of features. Exercise 1 from Chapter 2 of the class textbook.

4. Compound Bayesian Decision Theory. Suppose we have three categories with $P(\omega_1) = 1/2,\ P(\omega_2) = P(\omega_3) = 1/4$ and the following distributions

- $p(x \mid \omega_1) \sim \mathrm{N}(0,\ 1)$
- $p(x \mid \omega_2) \sim \mathrm{N}(0.5,\ 1)$
- $p(x \mid \omega_3) \sim \mathrm{N}(1,\ 1)$

Using the following, which assumes the independence of x_i and $\omega(i)$:

$$p(\mathbf{X} \mid \boldsymbol{\omega}) = \prod_{i=1}^{4} p(x_i \mid \omega(i)) \qquad ; \qquad P(\boldsymbol{\omega}) = \prod_{i=1}^{4} P(\omega(i))$$

and using the *dnorm*() R function, calculate explicitly the probability that the sequence $\mathbf{X} = \ < 0.6, 0.1, 0.9, 1.1 >$ came from $\boldsymbol{\omega} = < \omega_1, \omega_3, \omega_3, \omega_2 >$.

5. Discriminat Functions and Maximum-Likelihood (ML) Estimation. Consider the following data sets:

D1 = {<3, 4>, <4, 6>, <2, 6>, <3, 8>}
D2 = {<3, 0>, <1, -2>, <5, -2>, <3, -4>}

a. Using ML estimates for μ_1, μ_2 and Σ_1, Σ_2, write expressions for the discriminat functions $g_1(x)$ and $g_2(x)$
b. Assuming equal priors, find expression for decision boundary by setting $g_1(x) = g_2(x)$
c. Draw, by hand or otherwise, decision boundary, means, and data points.

6. K-Nearest Neighbor Classifier. Exercise 7 from Chapter 2 of the class textbook.