# CSC 422/522: HW2

## [15 Points] Question 1

Write code in R or Matlab to perform each of the following tasks:

1. Generate a $3 \times 3$ matrix with input containing the sequence $1, 2, \ldots 9$.
2. Using the matrix from above, complete the following.
   (a) Access elements from the 2nd and 3rd columns only
   (b) Access elements of the 2nd and 3rd rows only
   (c) Access rows 1 and 3 only? (see rbind() function in R and vertcat() in matlab)
   (d) Calculate sum of the 2nd row, the diagonal and the 3rd column in the matrix.
   (e) Identify row and column dimensions of the matrix.
   (f) Transpose of a matrix.
   (g) Scalar multiplication of output matrix with itself.
   (h) Matrix multiplication of output matrix with itself.
   (i) Cross product of the output matrix from 1.
   (j) Check if a matrix is a square matrix.
   (k) Inverse of a matrix.
   (l) Identity of a matrix.
   (m) Sum of all elements in the matrix (use a for/while loop)

## [19 Points] Question 2

For this exercise, use the “values.txt” file provided. The file contains a list of 150 data instances. There are 2 columns representing the $x$ and $y$ coordinates. Complete the following tasks:

1. Load the file
2. Make a 2-D plot and label the axes
3. Find the correlation between the dimensions.
4. Now consider a point $(5, 1.5)$.
   (a) Compute the distance of this point from each of the 150 instances using Euclidean distance, Mahalanobis distance, City block metric, Minkowski metric, Chebychev distance and Cosine distance.

(b) For each distance measure, identify the 10 points from the dataset that are the closest to the point (5,1.5).

- Create plots, one for each distance measure. Place an 'X' for (5,1.5) and mark the 10 closest points. To mark them, you could place a circle or any other shape over the point.
- Verify if the set of points is the same across all the distance measures.

## [22 Points] Question 3

In this question you will be asked to summarize and explore data in file labelled `hw1q3.csv`, available on the course moodle site. This file contains 2 columns of 1000 data points each.

1. Load the data contained in the file using your platform of choice. Compute and report the mean, median, standard deviation, and the range (i.e. the minimum and maximum) for each variable x1, x2. As discussed in class, please be sure to use $s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ to compute the standard deviation. You don't have to do this by hand, but please verify any approach you use is based on this formulation. Report each estimate to 3 decimal places.

2. Compute the *quantiles* for each variable. The quantiles of data set are the 0,25,50,75,and 100 percentiles. Report each to 3 decimal places:

3. Create a histogram (see `hist()` in R and `hist()` in Matlab) for each variable using 10 bins. The scale of the y-axis should be in terms of density, not frequency. Also, overlay a hypothetical 'true' density as a line. For this problem investigate the claim that the data were generated from a normal density with mean (represented by $\mu$) 5 and and standard deviation (represented by $\sigma$) of 1.5. The normal *probability density function* (pdf) is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

   This function indicates the y-value for any x-value in the data set. You are asked to plot this line, $f(x)$, and overlay it on each histogram. This plot will be a good way to visually inspect if a variable comes from a normal distribution with the indicated parameters. Note: you may do this anyway you wish, so if there are built in functions that will help you accomplish this, you are free to use them.

4. Create a *quantile-quantile* plot (commonly called a QQ plot) for each variable. The purpose of a QQ plot is to visually match an observed distribution to a theoretical one. In the previous question, it was suggested that the data came from a normal distribution. If this is true, when we plot the quantiles of our data against the quantiles of a normal distribution, it should form a straight line. Construct your plot using `qqnorm()` in R and `qqplot()` in Matlab. Include in your plot a line indicating perfect agreement, i.e. y = x.

5. Comment briefly on what you have learned about each variable. Include comparisons between x1 and x2 using location measures (mean, median, etc.), spread measures (standard deviation), and the shape of the histograms. Qualitatively, does either variable appear to have come from a normal distribution? Why?

## [19 Points] Question 4

For this exercise, you will use the Auto MPG Data Set from the UCI Repository (`http://archive.ics.uci.edu/ml/datasets/Auto+MPG`). Download the "auto-mpg.data" dataset. DO NOT use "auto-mpg.data-original". Once you have downloaded the data, perform the following tasks.

1. Load the dataset

2. Display the first 10 rows of the dataset

3. Let us assume that the dataset is being used for a regression task to predict the MPG (miles per gallon) using $x = \{Displacement, Horsepower, Weight, Acceleration\}$.

   (a) The following linear model was determined empirically to estimate MPG:
   $MPGEst = w_0 + \sum_{i=1}^{n}(w_i x_i)$
   where, $w_0 = 45.5455$ and $w = \{-0.0177, -0.037, -0.0037, -0.2941\}$.
   Using the linear model, compute MPGEst for the entire dataset.

   (b) The dataset contains the true value of MPG for each car and the linear model above enabled you to estimate the MPG for each car. Is the MPGEst correct? If not, report the Root Mean Squared Error (RMSE).

   $$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{1}$$

   where, $y$ corresponds to MPG given in the dataset and $\hat{y}$ corresponds to MPGEst that you computed previously.

4. We would like to visualize how MPG varies with Acceleration and Displacement. Show the result of the following:

   (a) Display a 3-D scatter plot of Acceleration, Horsepower and MPG. Label the axes.

   (b) In the same 3-D scatter plot, visualize how both MPG and MPGEst vary with reference to Acceleration and Horsepower. Label the axes. Make sure that MPG and MPGEst are displayed in different colors. Show a legend.