IE 529
Stats of Big Data and Clustering
Computational Assignment: 1

I. Regression analysis:

1. The data is given to you in *Comp1_IE529* contains two vectors, 'lift_kg' and 'putt_m', where 'lift_kg(i)' corresponds to a maximum weight lifted by athlete i in kilograms, and 'putt_m(i)' corresponds to a longest shot-put by athlete i in meters. Your assignment is to use regression methods to determine a model that describes the relationship between the two variables. That is, suppose x_1 = 'lift' and x_2 = 'putt'; you should find a mathematical model relating x_1 and x_2, such as
$$x_2 = 10 + 2x_1 + x_1^2 - 0.1x_1^3.$$
The relationship may be linear/affine, polynomial or logistic.
2. Write a simple program to compute a least squares solution for the case of linear or polynomial regression, and determine the lowest order model that fits the data reasonably well (order 1 is linear, and higher orders are polynomial).
3. Call an existing logistic regression function in Matlab or Python to determine if a logistic model will fit the data much better or not.
4. State which of your candidate models best describes the data, taking into account that simplicity is preferred. **Provide plots** of a linear fit, one polynomial fit (i.e., second order or higher) and one logistic fit. Compute the sum-of-square of residuals, i.e., give the actual cost $\|\epsilon_i\|_2^2$, for each of the models plotted.
5. Turn in a report including (a) clearly labeled plots with associated explicit models and costs, (b) a brief discussion explaining your model choice, and (c) your code/function calls.

II. Principal Component Analysis: real data

1. The data for this portion of the computational assignment consists of 4 vectors of data, each with 150 entries (NOTE: each row is a sample with 4 entries; the number of rows is the number of samples). Each column represents a measurement, in centimeters, of one specific feature, taken from a sample of 150 flowers. The four features are sepal length and width, and petal length and width. We would like to distill this data down to a lower dimension (2 or 3), and try to determine how many species might be represented by the data.
2. For the given data, de-mean each entry, using column sample means. Perform a PCA on the de-meaned data. Clearly explain how you performed the PCA (submit pseudocode for your own PCA code, or reference and describe any function you called, i.e, from what library, how it works, what it takes as inputs and produces as outputs).
3. State what portion of the variance in the data is contained in each of the 4 principal components. From these values, state how many *true* components are needed to represent the data.

4. On a 2D graph with the horizontal axis given by the first PC, and the vertical plot given by the second PC, plot the 2D representation of all the data points (i.e., find the 2D projection of each row). Discuss: (a.) are there any visually apparent clusters, if so, how many, and (b.) from this do you think you can conjecture anything about the number of species represented by the data?
5. Repeat parts (2.)-(4.) for *standardized* data: in addition to de-meaning the entries by column means, you should also scale by the inverse of the sample standard deviation, i.e., for each element x_{ij} in the original matrix X, normalize the element to $\tilde{x}_{ij}$ where

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}.$$

   Note whether or not this scaling changes the outcome of your analysis.
6. Turn in a report including (a) a matrix with the 4 components for the data and their associated variances, and (b) plots for part 4. for both the de-meaned and the standardized data sets.