EENG 415: Data Science for Electrical Engineering

Dr. Salman Mohagheghi

HW1: Data Preparation

1. Create a random vector with 200 entries, consisting of positive values. One way to do this could be `abs(randn(200,1))`. Plot the histogram of the data and find its skewness. Then use the command `boxcox` in Matlab to transform it to a dataset with smaller skewness. Plot the new histogram.

Some Matlab functions that you may find useful in this problem:

- `skewness()`
- `hist()`

2. Generate a 2-dimensional dataset in which X_1 varies from –1 to +1 in 0.01 increments and X_2 is equal to $2 \cdot X_1$ + noise. For noise, you can use random white noise[1]. Find the scatterplot of the data (X_1 vs. X_2). Now, augment the dataset by adding 10 outliers to it. Make sure the outlier points are sufficiently different from the rest of the data pattern. Apply the spatial sign transformation to the new augmented dataset and compare the new scatterplot with the original one.

3. Load the database Pizza.csv into Matlab[2]. The dataset contains measurements of the contents of different Pizza samples. The variables in the data set are:

- brand: class label
- id: identifier for the sample analyzed
- mois: amount of water per 100 grams in the sample
- prot: amount of protein per 100 grams in the sample
- fat: amount of fat per 100 grams in the sample
- ash: amount of ash per 100 grams in the sample
- sodium: amount of sodium per 100 grams in the sample
- carb: amount of carbohydrates per 100 grams in the sample
- cal: amount of calories per 100 grams in the sample

Remove the first two columns and create a data table that consists of only the numeric values (columns 3 through 9). Normalize the matrix and find the principal components. Choose the number of components that explain at least 90% of the variations in data. Transform the data points into the new reduced-dimension domain.

Some Matlab functions that you may find useful in this problem:

- `normalize()`
- `cov()`
- `eig()`
- `sort()`

[1] Make sure you use a small magnitude so that the noise value does not overshadow X_2. A magnitude of 0.2 is reasonable.

[2] This database is taken from data.world.