# Abstract

Unsupervised learning aims to provide insights into unlabeled data where there isn’t any classification of the data to be learned. Unsupervised learning has usuage in itself for data visualization or can act as a preprocessing of data for eventual supervised learning. This paper explores several unsupervised learning algorithm like K-Means, Principal Component Analysis, Independent Component Analysis, Randomized Projection and Decision Tree based Feature Selection. The analysis is executed on breast cancer data, and car buying quality data. Furthermore we use the mentioned unsupervised learning tactics mentioned above to create new or reduced feature that are then used for Neural Network Analysis.

# Dataset

## Data Set 1: Wisconsin Breast Cancer Data

The Wisconsin breast cancer data obtained from UCI Machine Learning Repository contains 699 samples with 9 input attributes. The samples are classified as either benign or malignant. In addition there were 16 missing values within the Bare Nuclei attribute.

The data is interesting for machine learning because cancer is a terrifying disease, which can be controlled if identified earlier stage of the disease. Machine Learning techniques can help provide insights into various tests results and their predictability of someone having a malignant cancer. Furthermore, since the treatment are expensive and also has various side effects by identifying people who do not have the decision can save hardship for the individual patients and save on healthcare cost.

In preparing the data the sample code number was removed, as it is just an ID number that doesn’t have any relationship with the cancer data. Next the 16 missing values of Bare Nuclei were populated using a mean strategy, where the missing values were filled with the mean of all the other data in that column, which turned out to be 4. Finally the data was scaled to have zero mean and an unit variance.

## Car Quality Data

The car quality check data set also obtained from UCI Machine Learning Database had a sample size of 1728 with 6 attributes. The samples are classified as unacceptable, acceptable, good, or very good.

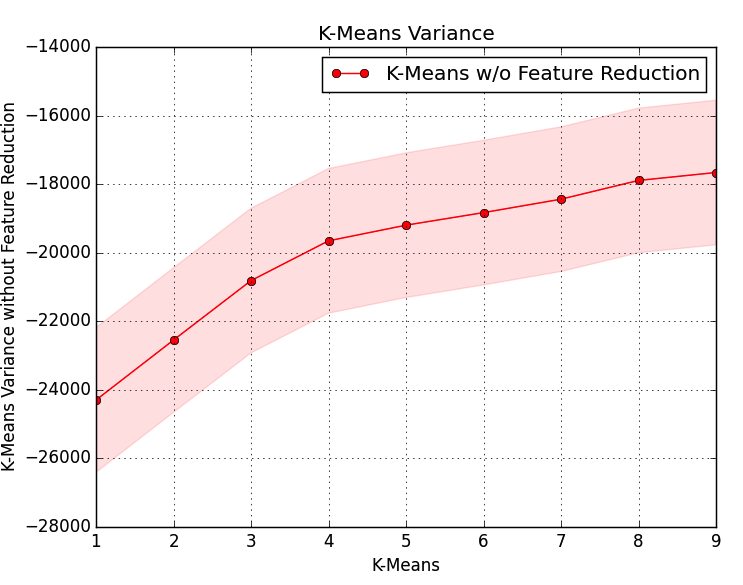
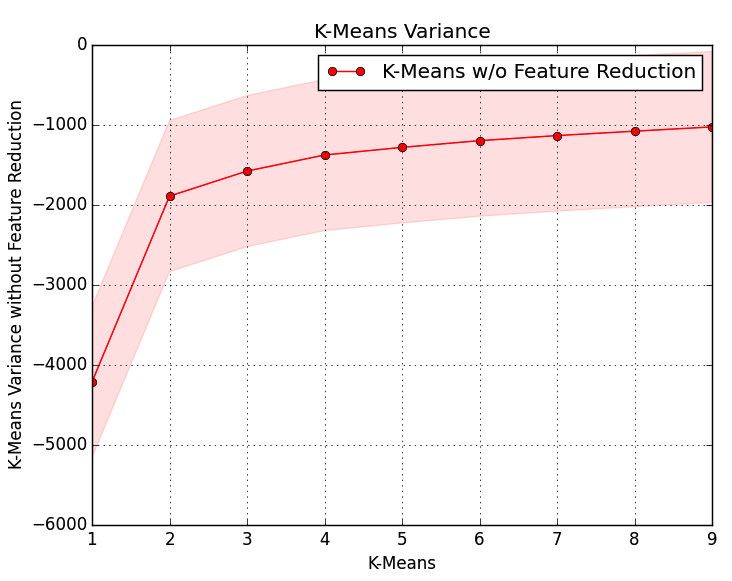
This is an interesting topic for machine learning from a business and consumer view point. Every days many consumers are in the market for a new car. Having smart algorithms that can guide the buyers through their journey can turn into actual business software. Furthermore, such analysis also will help car dealers price the cars within he right buying range.

The feature sets of the this data captures the car’s buying prices, maintenance price, number of doors, number of people it can fit, truck size, and safety rating. All these features are categorical data. As such the data preparation ensured that the features were converted into vector fields with one hot encoding (e.g.[1, 0, 0 , 0])

# Part 1: K-Means Cluster Analysis

K-Means clustering works by grouping data points to k number of centers (random points in the data space). The data points are grouped to the closet center and the centers are recomputed as the means of the data points in that cluster. This is repeated for several iterations or until convergence. The ultimate goal of K-Means is to cluster the data such that there is a valance between having low variance between the data with their cluster centers and avoiding the trivial solution of having one cluster per data point. K-Means analysis was done on both of the data sets to understand the makeup of the data as they relate to each other. Since there is no definite number of clusters we can identify the analysis looked at two methods for finding the optimal K (number of clusters): first by evaluating the aggregated distance from closest centroid, and secondly by evaluating the how closely the count of data points cluster labels matched the actual target class counts.

Below graph shows the performance of k clusters against the aggregated variance of the data to their cluster centers. Note the variance from the center is calculated as a negative distance in this analysis. The first graph below shows the results for the breast cancer data and the second graph shows the results for the car quality data. As you can see for breast cancer data the subsequent improvement of variance reduces drastically after k = 2 and for car quality data we can see the same effect after k = 4. This makes sense intuitively as the data breast cancer data and car data sets have 2 and 4 target classes, respectively. The number of target classes seems to naturally fit the optimal number of classes.



Additional analysis was done to see how the count of data points in each target class match to the count of the data points on each clusters. Since clusters are not created in any particular order effort is taken to give the algorithm the best-fit chance. That means highest data count of data points in a cluster is matched to the highest count of data in target class.

# Bibliography

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
2. Stephan Chalup and Frederic Maire: “Design Issues In Hill Climbing For Neural Network Training”.
3. Jeremy S. De Bonet, Charles L. Isbell, Jr. and Paul Viola: “MIMIC: Finding Optima by Estimating Probability Densities”.