# 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.

## Data Set 2: 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 viewpoint. 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])

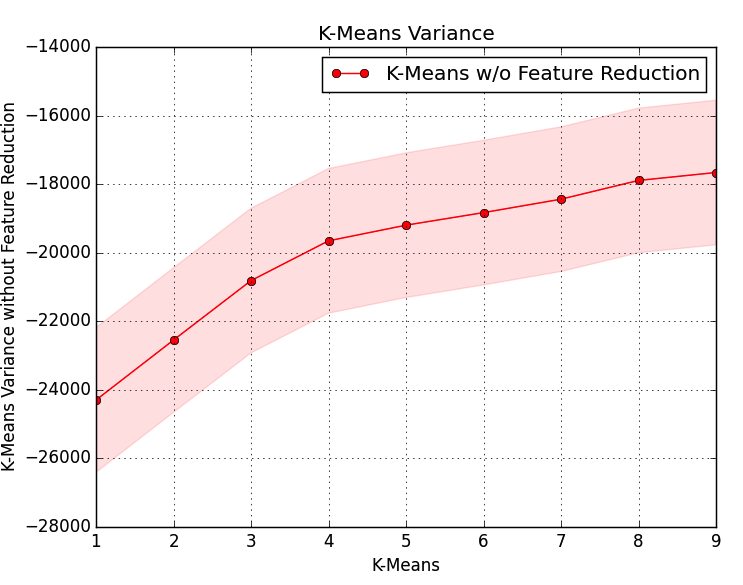
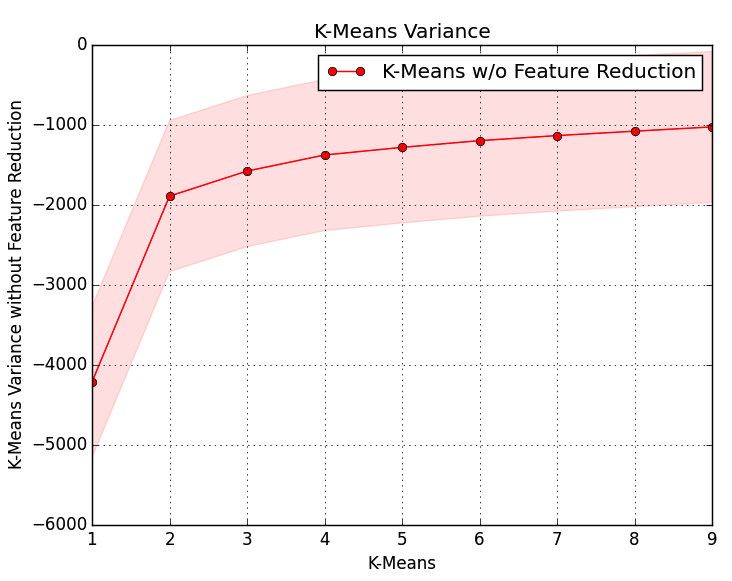
# Part 1: Clustering

Clustering is the unsupervised technique on finding groupings of unlabeled datasets. In this section we discuss the K-Means clustering algorithm and the Expectation Maximization clustering algorithm.

## K-Means

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.

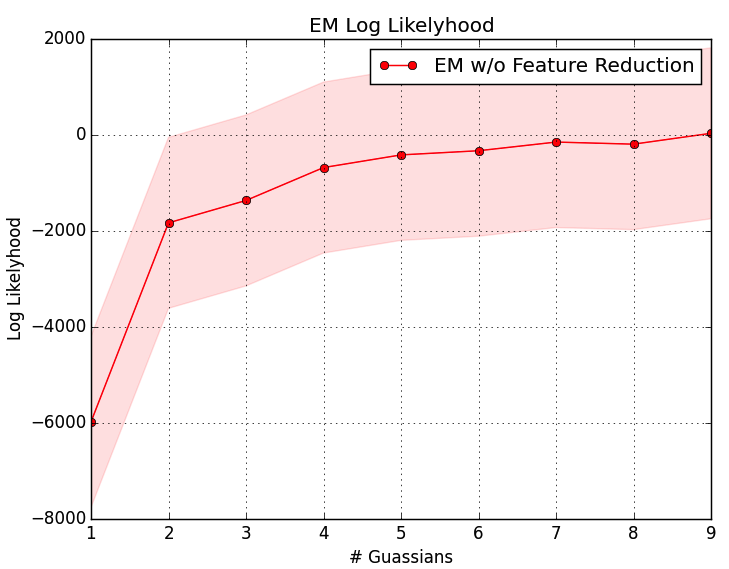
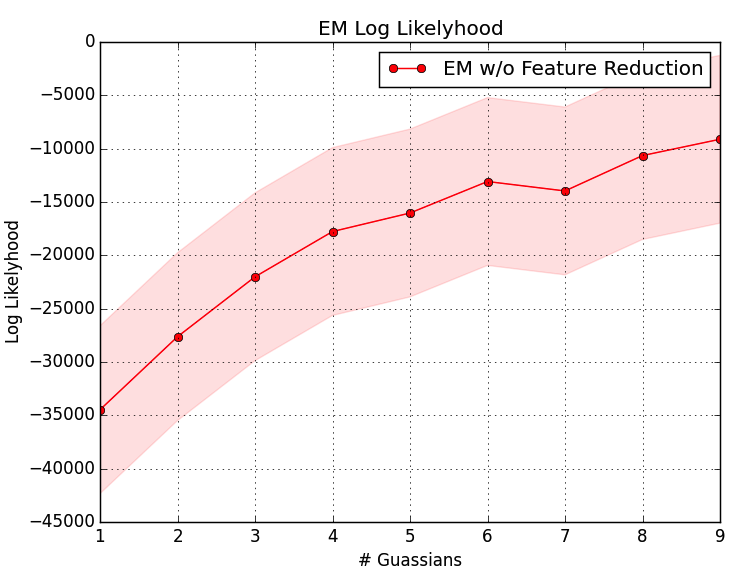
Below graph shows the performance of k clusters against the aggregated variance of the data to their cluster centers. Note the variance from the center is calculated as a negative distance in this analysis. The first graph below shows the results for the breast cancer data and the second graph shows 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.



## Expectation Maximization

Often time it is possible for a data point to have similarities with two data datasets and do not necessarily belong in one particular cluster. Expectation Maximization (EM) solves for this scenario by soft assignment of data points to a cluster. Where K-Means attempts to label data into hard clusters EM uses a probabilistic approach to calculate the log likelihood of a particular data point belonging to a particular cluster. EM expects that the all data points are generated from Gaussian distributions. Since Gaussian expends infinitely there is always a small probability that a certain data point belongs to more than one cluster. Similar to K-Means finding the optimal number of clusters depends on the analysis of the resulting data of multiple EM runs with various values of K (number of clusters). However, instead of looking at the aggregated variance we look at the sum of the log likelihood to identify the optimal number of clusters.

Below graphs show the aggregated log likelihood score of EM with various values of cluster size (Gaussians). The first graph is for the breast cancer data and the second graph is for the car quality data. As we saw with K-Means variance analysis below it appears that the log likelihood slows down in its improvement as we past the number of target class for each of the datasets, namely 2 for breast cancer data and 4 for car quality data. This again tends to fit into the natural clusters we would expect the datasets to have based on their target classes. However, the data for car quality data is not as clear in identifying the optimal number of clusters as the breast cancer data. One can argue that the optimal number of clusters for the car quality data based on the log likelihood analysis is 5 or even 6. This makes sense because the quality of the data depends on multiple factors and is more of a qualitative identification. What may be good quality to one person can be an acceptable quality to another person.

# Part 2: Dimensionality Reduction

One of the key challenges faced by machine learning community is the high dimensionality of the data. Any practical data set of interest to machine learning community has high dimensions. As we go past our 2D or 3D space it becomes harder and harder to conceptually visualize or grasp the data we are analyzing. Additionally, often times our data set will contain features that may be redundant or may not have any relevance to our problem at hand. Dimensionality reduction attempts to solve this problem by extracting features are of most relevant to the learning problem. This section reviews four of these techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projection (RP), and Decision Tree Feature Extraction.

## PCA

# 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”.