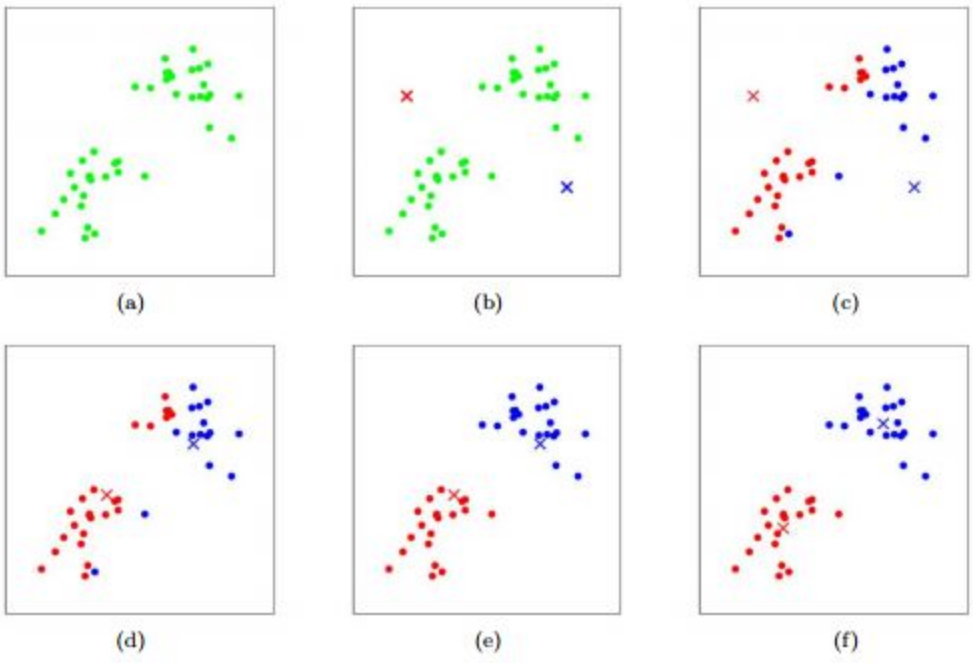
Josh Morris
OLA 4 Report - K-Means
CSCI 4350
Dr. Phillips

Unsupervised learning methods attempt to find patterns in data based on their features instead of their associated labels. The advantage of this approach is that labelled data is unnecessary for training. While supervised learning can often be more effective in learning a desired trend, labelled data is often scarce or expensive to acquire. Thus, unsupervised learning can be a good approach to find trends in large amounts of data. For this lab, a k-means method was used to classify both iris data and breast cancer data.

K-means clustering is a parametric method for performing unsupervised clustering. K-means clustering works through placing N centers randomly in the decision space. Think of these centers as being representative of some significant relationship in the data. On each iteration of the algorithm, the euclidean distance is calculated from each data point to each center. Each data point is then paired with its closest center. Once all data points have been paired with a center, the points associated with each center are averaged to calculate a new position for each center. This process is repeated until the centers no longer move. K-means is considered parametric because it divides the decision space linearly.




(Piech)

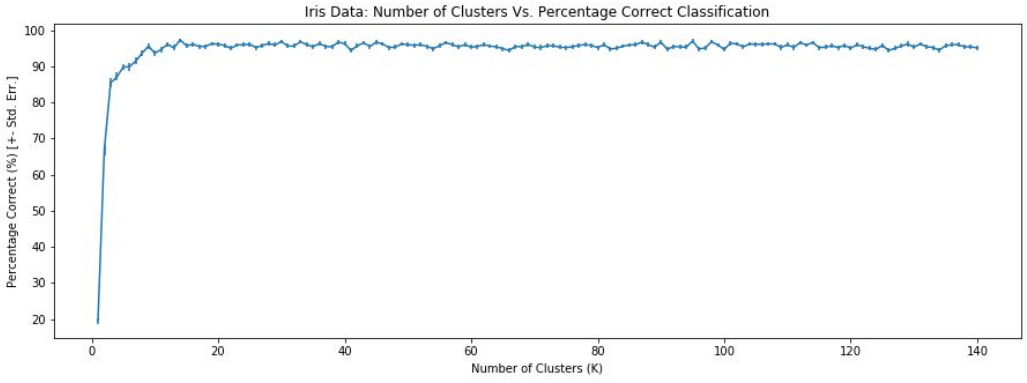
The figure above is a good visual representation of the k-means algorithm. In b the centers are randomly initialized. Then in c through f, the centers are pulled toward clusters in the data finally stopping in the middle of the clusters.

The k-means algorithm for this lab was implemented using C++. The implementation takes 5 command line arguments: random seed, number of clusters, the number of real-valued features, training data path, and test data path. The program loads data from the training file that is formatted such that there is a data point per each line with each attribute being white space delimited and the last item on each line being the data point's associated label. Following loading data, the number of centers specified by the user are initialized to be a random point in the training data. This initialization serves to keep centers from having no data associated with them during iterations of the algorithm. Following initialization of the centers, training is started. Training is carried out in the method described above and does not stop until the centers are the same for two iterations.

Once training iterations are complete, the centers are then labelled with the majority classification of the points associated with each center. Note that this is the first time the data labels are used, until this point in the program only the features of each point were used for classification. This is what makes the approach unsupervised.

Following center labelling, the model is tested using a user supplied testing set. This test set helps to gauge the effectiveness of the trained model in generalizing to unseen data. This testing set should not contain training examples, but should follow the same format. Each training point is then associated with its closest center via euclidean distance, the label of that center is then applied to the data point. This label is then compared with the data point's actual classification. A count of correct labels is kept and output after all testing examples have been evaluated.

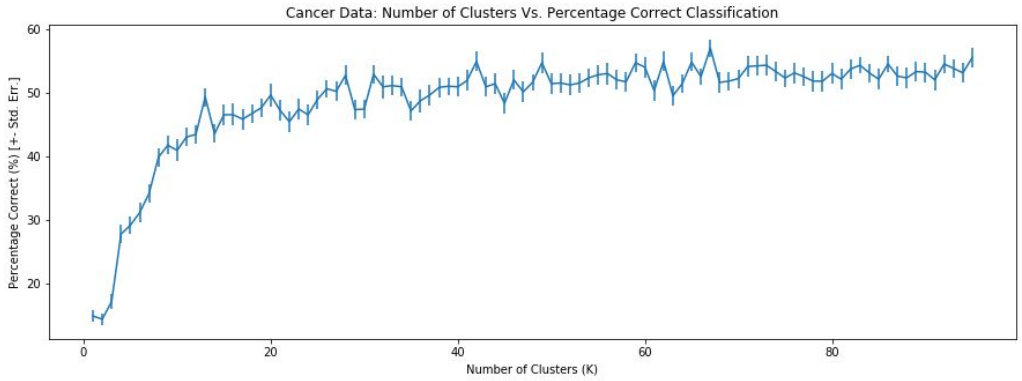
The first data set evaluated using k-means clustering was iris data. This data is originally from UCI and has specific measurements of various types of iris flowers. For the purposes of this lab, these flowers were divided into three distinct classes. For the purposes of this lab, we were interested in testing the effectiveness of the number of clusters. Thus, 1 center to N-10 centers, N being the number of training examples, were tested using 10-fold random subsampling validation. 10-fold random subsampling validations means 10 random examples were held out from training for testing purposes on each trial of the algorithm. For each number of clusters, 100 trials were collected for random seeds 1 to 100. The average result with standard error can viewed in the graph below:




It can be seen from the graph that k-means clustering reaches around 90% accuracy with 7 clusters. This means that the training data can be represented by roughly 7 center vectors. This is

a compression of data of about 22:1 (150 data points / 7 feature vectors). This really serves to demonstrate the potential effectiveness of k-means clustering.

The second dataset evaluated for this lab was a breast cancer dataset from UCI. The same methods were used as for the iris dataset from UCI. The results are as follows:




It can be seen by the graph above, that k-means clustering is not as effective for breast cancer data as it is for iris data. Even at 95 clusters, accuracy remains under 60%. This suggest that the attribute features supplied to train do not contain enough information for accurate classification. To get more accurate result further testing may be required or more information may be necessary from the previous test. It could also mean that the relationship of the features of the data are nonparametric in their relationship and a different unsupervised approach might be more effective. At under 60% accuracy, it is hard to say results from the classifier are all that helpful or useful.

When I initially implemented k means, I read the specification wrong and used the average of the number of clusters of randomly selected data points to initialize each center. This actually proved to not be as effective as it might appear. Often centers were left with no points assigned to them, thus dividing the problem space in weird ways. It lead to a 20 point reduction in effectiveness on the breast cancer data. When I changed the implementation to use a single randomly selected data point to initialize each center, the algorithm performed much more stably.

In all, k-means can be an effective way of discovering trends in data in an unsupervised fashion. For the iris data it was over 90% effective in classification with only 7 center vectors, leading to a 22:1 reduction in size from the input data. However, these results are not the rule. K-means clustering and effectiveness is dependent on the input data. This is demonstrated by the breast cancer data where classification was below 60% accurate after training. Therefore, it should be noted that k-means clustering can be effective in discovering trends, but any model should be rigorously evaluated and not used without such testing.

## Works Cited

Piech, Chris . “K Means.” CS221, Stanford, stanford.edu/~cpiech/cs221/handouts/kmeans.html.