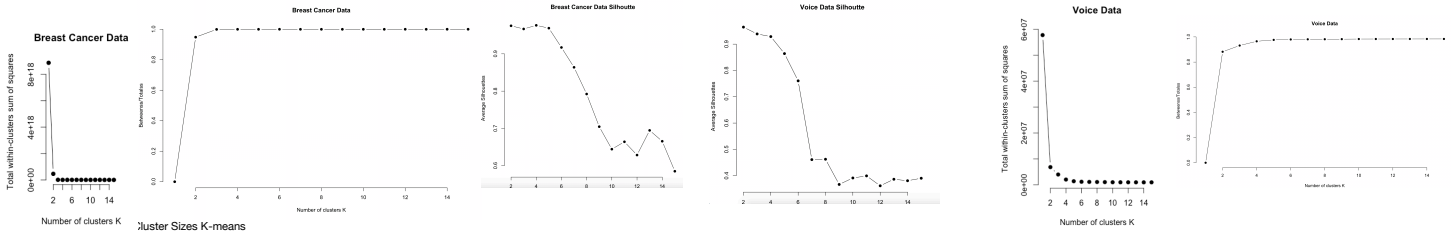
Please refer last page for description of the datasets. Please zoom in to see the plots, I had to reduce their size due to space crunch. Sorry

K-Means Clustering

K-means clustering algorithm was implemented using R. The number of clusters k were varied from 1 to 15. The parameters used for assessing the clusters formed are 'Total within-cluster sum of squares' , 'Ratio of Betweenss/Totalss'.K-means clustering algorithm starts by choosing random points equal to the number of clusters defined as the centers of the clusters and then assigns all the data points to the cluster belonging to one of the centers such as there is minimum distance between the data points and the cluster centers. In other words, the "Total within cluster sum of squares should be minimum". On the other hand the clusters should be as distinct as possible so the distance between the centers of two clusters should be as maximum as possible . So, the ratio of "Betweenss/Totalss" should be as high as possible.The algorithm chosen to implement k-means was 'Hartigan and Wong' with number of random restarts as 25 and the maximum number of iterations for a single run chosen were 100 to ensure that the algorithm does not get stuck in local minima while optimizing the cluster centers and the corresponding clusters to the centers using random hill-climbing approach. Foe all the graphs in this assignment the silhouette method is plotted for total within cluster sum of squares v/s number of clusters

Voice Data- While running k-means clustering algorithm, I encountered convergence warning in the Quick transfer stage which reflects that the data points are very close to each other . However, I resolved this warning by increasing the number of iterations from 50 to 100. The fact that the data points are very close to each other can also be visualized by the Cluster sizes bar plots where it is clear that the weight of one of the clusters was very high and the other clusters had very less weight. So, for all the k-values defined we can see that all the data points are clustered to a single cluster and very few data points are assigned to the other clusters , this suggests that there is not much variance among the data points and they are very close to each other in the dimensional space



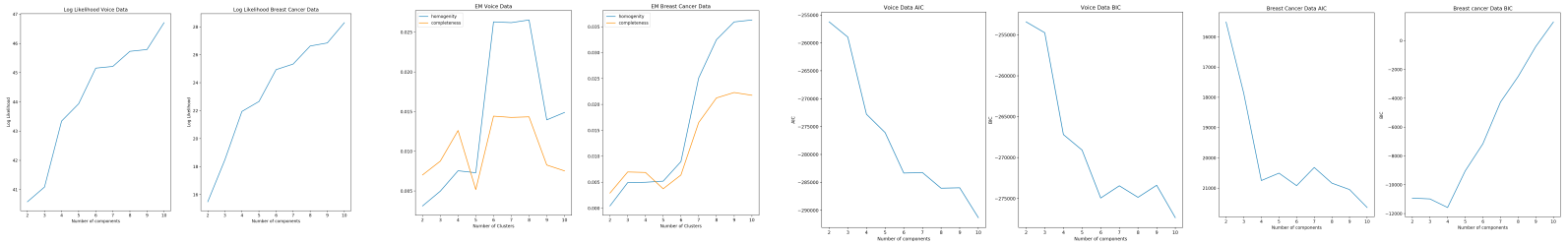
As can be seen from the above graphs that as we increase the number of clusters from 2 to 3 the percentage variance for both 'Total within-cluster sum of squares' and 'Betweenss/Totalss' decreases to a great extent. So, by elbow method by looking at these two plots we can say that the number of clusters identified by k-means are 2 for voice data which aligns with the ground classification of voice data as well. If we look at the average silhouettes produced for voice data we can see that the maximum average silhouettes are produced for k=2. So, by average silhouette method as well we can conclude that the number of clusters formed for voice data are So, the k-means algorithm for voice data suggests there should be at maximum 2 clusters, although most of the data points are clustered in one cluster. So, we can say that the data points do not have much variance among each other or are very close to each other in the dimensional space

Breast Cancer Data

If we look at the graphs of 'Total within-cluster sum of squares', 'Ratio of Betweenss/Totalss' we can see by the elbow method that the percentage variance for both of these parameters decreases rapidly as the number of clusters increases from 2 to 3. So, by elbow method we can say that the appropriate number of clusters for breast cancer data is 2 which aligns with the ground classification of the breast cancer data as well. However, if we look at the average number of silhouettes for the breast cancer data we can see that the number of average silhouettes for both k=2 and k=4 is the same. So, silhouette method suggests that the number of clusters for breast cancer data can be 2 or 4

Expectation- Maximization

Expectation Maximization is implemented using python's sklearn library. Expectation Maximization method is a probabilistic method which finds the probability of each of the data points belonging to a particular cluster. In other words it finds k distributions of data (Expectation step) so that the likelihood of data that they are generated from these k distributions is maximum(Maximization). EM is an iterative algorithm where it is iterated between the expectations and maximization step using random hill climbing approach. No of random restarts are chosen to be 25 with a maximum iteration of 100 to prevent the algorithm to converge to a local minima and the number of components(distributions) are varied from 1 to 10. The parameters chosen to evaluate the distributions is the per sample average log likelihood of the Gaussian mixture X. So, the log likelihood of the sample should be as high as possible for the given gaussian distributions. Other parameters chosen are Akaike Information Criterion (AIC) and Bayesian Information criterion(BIC). Both of these parameters are a measure of the information loss as we increase the number of distributions. So, the lower the better because we want minimum information loss.The penalty for increasing the number of parameters in BIC is more as compared to AIC. The penalty help in avoiding overfitting and the increase in parameters helps to avoid underfitting. Other parameters that are chosen are homogeneity and completeness of the distributions.
A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class and on the other hand clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster. So, two clusters which have all the data points belonging to the same class will be homogenous but not complete.



Voice Data- As we can see from the AIC for Voice Data as we increase the number of clusters from 6 to 7 the percentage variance in the AIC values does not change a lot and thus by the elbow method we can say that expectation maximization soft clusters the voice data into 6 clusters. Similarly by looking at the BIC graph for voice data we can see that as we move number of components from 6 to 7 the BIC value increases which is unfavorable because we don't want any information loss as we increase the number of clusters as that suggests overfitting. Similarly it can be seen from the log likelihood graph of voice data that the percentage variance of log likelihood is minimum as we increase the number of clusters from 6 to 7. Similarly, we can see that the percentage variance decreases for both homogeneity and completeness minimizes as we increase the number of clusters from 6 to 7.So, we can say that EM algorithm suggests 6 soft clusters for voice data which does make sense because as we observed from k-means that there is not much high variance in the data points and as k-means does hard clustering so it will not separate the data points according to these small variations and thus suggested just 2 clusters for the voice data whereas as EM only gives probability of a data point belonging to a cluster so it could be possible that the data points are probable to come from different distributions and EM is thus sensitive to the small variations of the voice data.

Breast cancer data - We can see from log likelihood graph that the percentage variance becomes minimum as we increase the number of distributions from 4 to 5 . Similarly , we can see from the AIC and BIC graph that they start increasing as we increase the number of distributions from 4 to 5 which is not favorable as we want minimum information loss and this suggests that there is just overfitting of the data after 4 distributions. Similarly, we can see that completeness decreases as we increase the number of clusters from 4 to 5 and both homogeneity and completeness has minimum variance as we increase the number of clusters from 3 to 4 . So, we can say that the number of distributions suggested according to homogeneity and completeness graph are 3 but looking at the other graphs AIC, BIC, log likelihood we can say that the number of distributions suggested are 4. This however, does not align with the ground classification but aligns with one of the suggestions from k-means clustering.The reason why the number of distributions suggested by EM or the number of clusters suggested by k-means does not

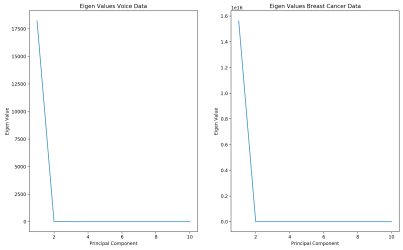
align for breast cancer data with it's ground classification could be because there may some irrelevant features (i.e. which do not provide information) in the dataset which if removed can help the clustering algorithms to take better decisions with regard to clustering. So, using an appropriate feature selection method can help the clustering algorithms to perform better. I have done this using Information gain as the feature selection method later in this assignment. We can also decrease the number of dimensions by using various feature transformation methods(PCA, ICA, Randomized projections) which can also help to choose components of maximum relevance and thus improve clustering. This is done in the next section of dimensionality reduction

Dimensionality Reduction

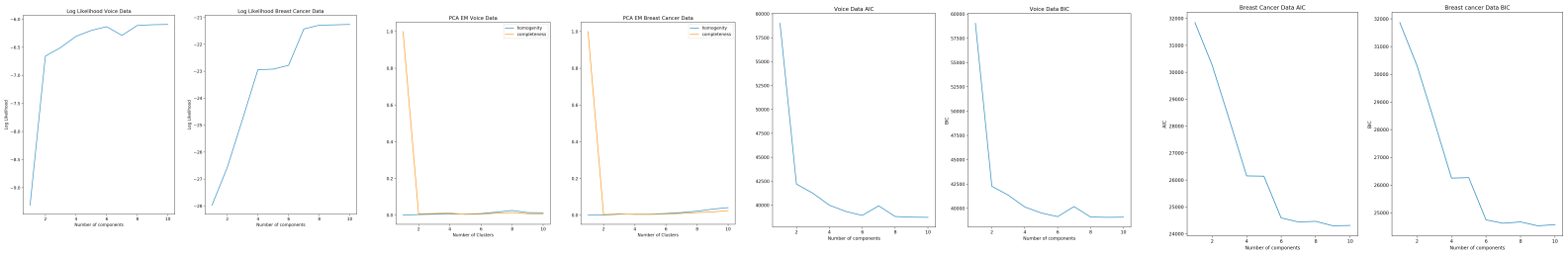
PCA- Principal Component Analysis

PCA is a feature transformation method which produces new features for the dataset in the direction of maximum variance and such that they are orthogonal to each other. Eigen values provide the magnitude of the eigenvectors(the principal components). So, the greater the magnitude of the Principal component more is it's relevance so we can say that the higher the eigenvalue of a principal component , more it is favoured.

PCA was implemented using sklearn. Initially to find the relevant number of components PCA was performed using 10 projections. As we can see from the Eigen value graph for both Voice data and Breast Cancer data that the Eigen value for component 2 or higher is almost negligible so we can conclude only first of the projected principal component is important. Next, I performed EM and k-means again with the reduced number of transformed variables to see if they can perform better by using relevant variables. I chose 2 principal components to perform clustering again although there was only one principal component relevant for both datasets. I got comparable results for clustering with both 1 or 2 principal components.



EM

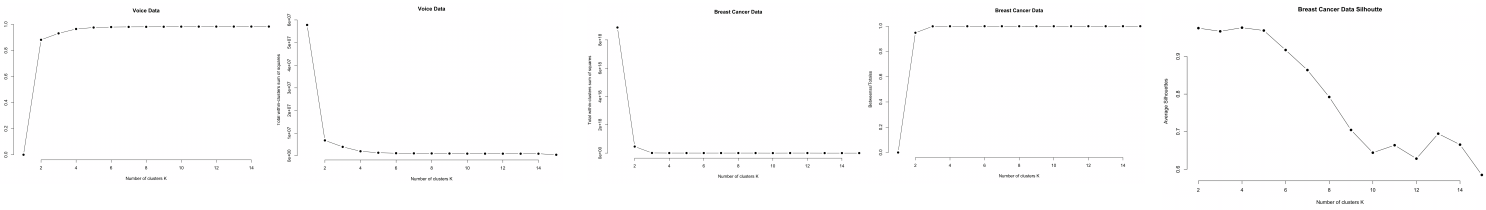


Voice Data- It can be seen from the log likelihood , AIC , BIC , homogeneity and completeness graph that there is minimum percentage variance as increase the number of components (distributions) from 2 to 3 . So, we can say that from all of these graphs 2 distributions are suggested for the voice data which does align with the ground classification of the voice data as well so we can say that PCA does a pretty good job by transforming the features to a component which gives the maximum variance and thus helps the EM clustering algorithm to identify the number of distributions correctly. As we observed that previously that without dimensional reduction EM was identifying 6 soft clusters for voice data this was majorly due to all the data points were located very close to each other in the dimensional space as they had very less variance among each other but after PCA transformed the features such that they have maximum variance among each other so EM could easily eliminate the 4 extra soft clusters which it had previously suggested. Thus, we can say that PCA helped EM to correctly identify 2 ground classification distributions for Voice data

Breast cancer data-It can be seen from log likelihood AIC and BIC graphs that the EM suggests 4 distributions for breast cancer data which is consistent with its previous suggestion without PCA. However, homogeneity and completeness graph suggest if there are greater than 1 distribution then the completeness drops considerably which means that the two or 4 distributions that are formed are not pure distributions and

have a lot of outliers. We can conclude that PCA does not help EM to identify correct distributions for breast cancer data probably because wrt to breast cancer data we don't need a component which has maximum variance but it could be possible that we require components that are independent or plain just choosing random projected components may work out or maybe no feature transformation is required and we need a good feature selection method for the original features. All of these cases will be tested in the next sections.

K-Means

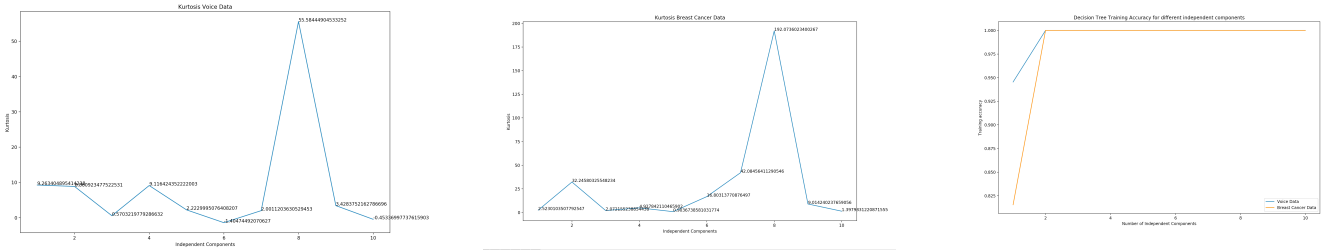


Voice Data- We can observe from the above plots that the within cluster sum of squares(plot2) by elbow method suggests that the number of clusters should be 2 which aligns with the ground classification. We can also observe that the ratio of between cluster sum of squares and total sum of squares does not increase much as we increase the number of clusters from 2 to 3 so this plot (plot 1) also suggests that the number of clusters for voice data should be 2 . We can say that PCA helped means to classify the clusters by providing features which have maximum variance.

Breast cancer data

However for breast cancer although the number of clusters suggested by elbow method from plot 3 and plots 4 is 2 but the silhouette method for breast cancer still suggests that the number of clusters can be 2 or 4 .So, we can say that for breast cancer data the target variable does not require features that have maximum variance but may require some other kind of transformed features such as independent or just random projections which will be analyzed in the further sections.

Independent Component Analysis

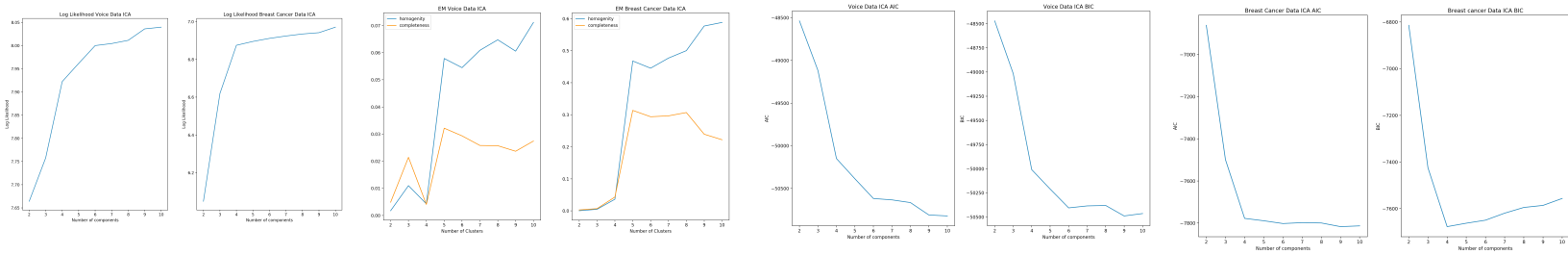


ICA is impelented using sklearn's FastICA and kurtosis is calculated using kurtosis function in python from the spicy module. This implementation of kurtosis in python has 0 value for normal distributiontion.The higher the kurtosis value from 0 it means the lesser the weight at the tails and thus a more sharp peak. The lesser the kurtosis value as compared to 0 more is the weight at the tails and a more flat distribution is obtained.

5

Decision tree is used as a sample classifier to choose the number of independent components that give the highest training accuracy. We can see from the plot (extreme right)that the sample classifier suggests that the number of independent components for both voice and breast cancer data should be 2. We can see from the kurtosis values(extreme left plot) that component 8 and component 4 are the ones that have the most non-gaussian distribution. Similarly for breast cancer data (middle) component 8 and component 7 are the ones with most non-gaussian distribution.

EM



Voice Data

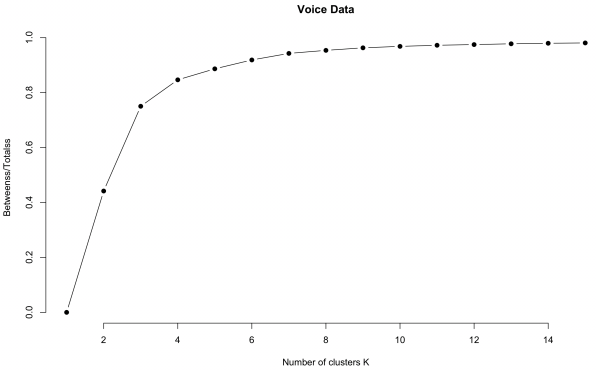
We can see from the log likelihood(Plot1), AIC(plot5), BIC(plot6) that the suggested number of components are 6 as the percentage variance as we increase the number of components from 6 to 7 does not changes. The homogeneity and completeness plots (plot3) suggests that if the number of distributions are increase from 6 to 7 there is no further increase. So, from all of these plots we can say that the suggested number of distributions by EM are 6 based on the two independent components that ICA suggested. We can say that the two independent components that ICA suggested does not help to predict the clusters for voice data in sync with the ground classification probably because the data points in voice data are very close to each other so they would be better classified on the basis of maximum variance components as PCA suggested as compared to independent components.

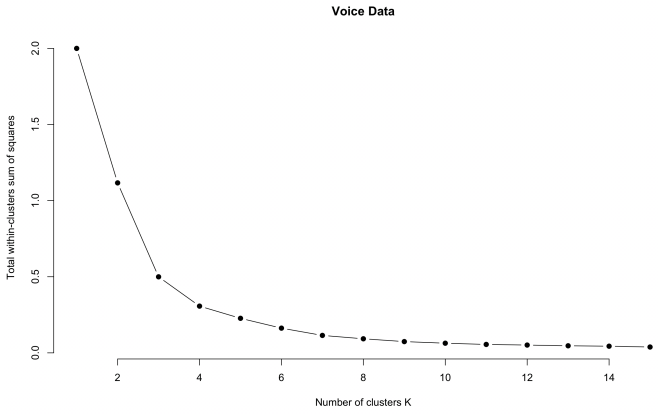
Breast Cancer Data

We can see from the log likelihood(Plot2), AIC(plot7), BIC(plot8) that the suggested number of components are 6 as the percentage variance as we increase the number of components from 4 to 5 does not changes.The homogeneity and completeness plots (plot4) suggests that if the number of distributions are increase from 4 to 5 there is no further increase.So, from all of these plots we can say that the suggested number of distributions by EM are 4 based on the two independent components that ICA suggested. We can say that the two independent components that ICA suggested does not help to predict the clusters for breast cancer data in sync with the ground classification probably because the data points in breast cancer data can be correctly classified based on some random feature transformation or simple feature selection from the original features (both of these will be tested in next sections) and not maximum variance feature transformation nor independent transformed features as suggested by PCA and ICA respectively.

K-means

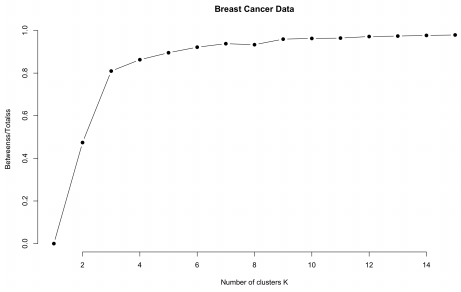
Voice Data

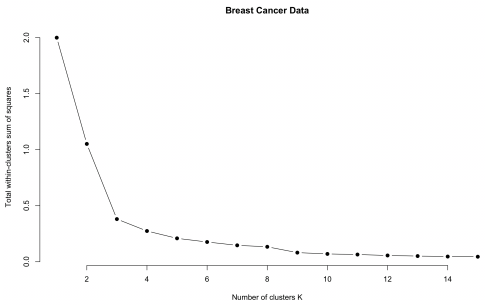


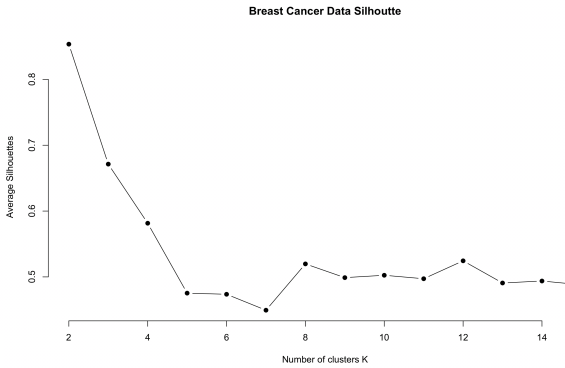


We can observe by the above plots that elbow method suggests that the number of clusters for voice data should be 3 whereas average silhouettes method suggests that the number of clusters should be 2. As the ground classification of breast cancer data is 2 so we can say that the independent components suggested by ICA may or may not be helpful for k-means to cluster correctly

Breast Cancer Data

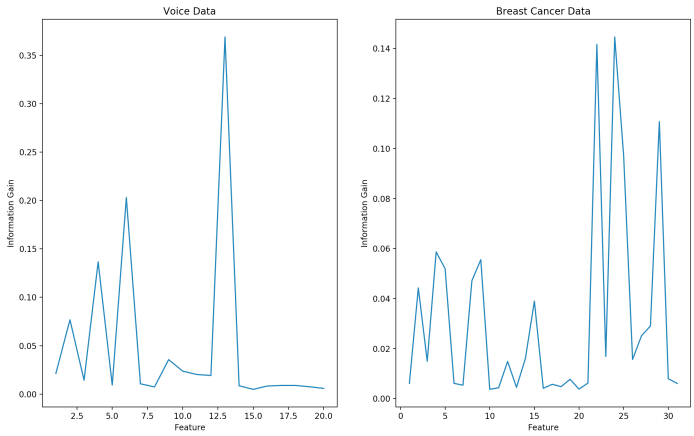






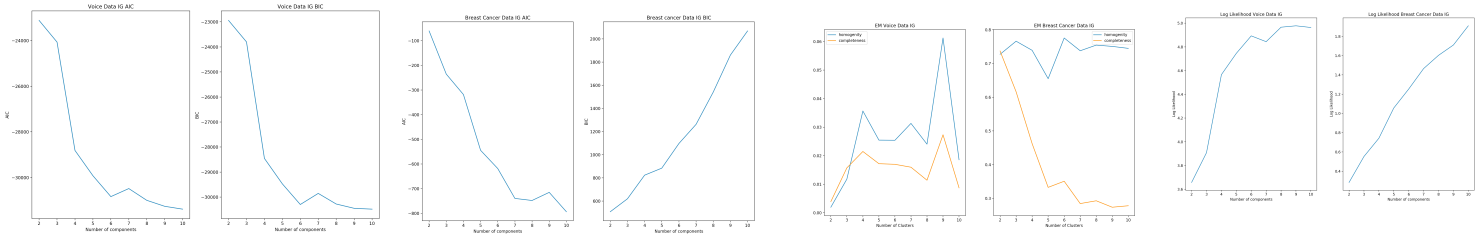
Information gain

I chose information gain as the feature selection method because it helps to choose the feature which decreases the entropy in the child nodes wrt to their classification labels. I implemented information gain using Random forest Classifier in python and then using the feature importance attribute which returns the information gain for every attribute.



We can see from the above plots that for Voice Data variable 2,4,7 and 13 provide maximum information whereas for breast cancer data variables 2,4,5,8,9,15,22,24,25,29 provide maximum information. So, I selected these features to perform clustering.

EM



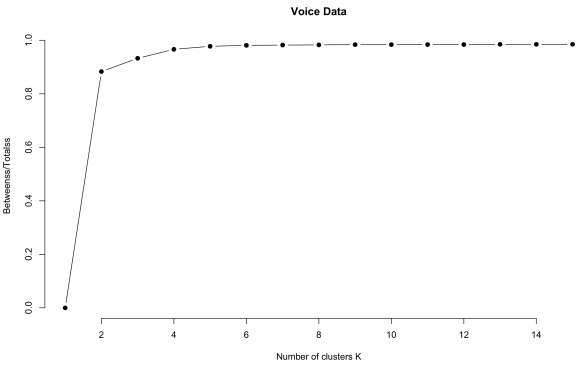
Voice data

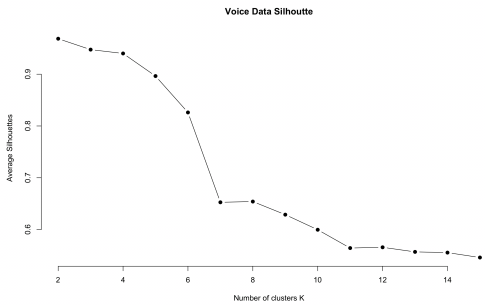
It is observed that from AIC(plot1) , BIC(plot 2), log likelihood(plot 7) that the number of components suggested by EM are 6 based on the 4 features which information gain suggested but it can se said that clearly EM does not require the features that have maximum information gain nor features that are independent but requires features that have maximum variance as we observed that EM classified the data points into 2 distributions(aligning with the ground classification) based on the maximum variance components suggested by PCA.

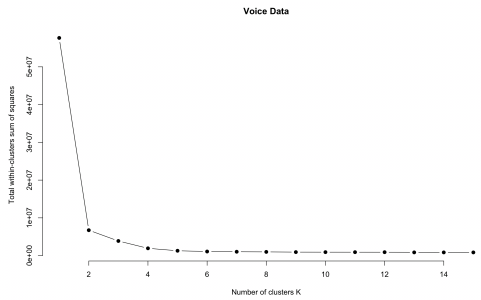
Breast cancer data

Results for breast cancer data were kind of inconsistent. By looking at log likelihood(plot 8) we cannot clearly find any elbow AIC (plot 3) suggests the number of components to be 7 and BIC (plot 4)suggests the number of components to be 4. However, completeness(plot 6) graph suggests that completeness drops considerably if the number of clusters are increased more than 2 . Thus, we can say the clusters formed by EM after using features suggested by information gain we would need some other method other than elbow method to quantify the number of clusters formed.

K-Means



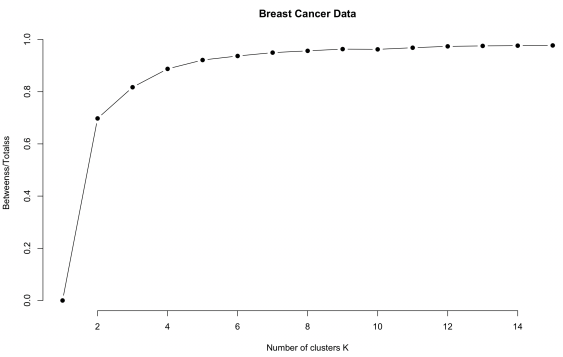


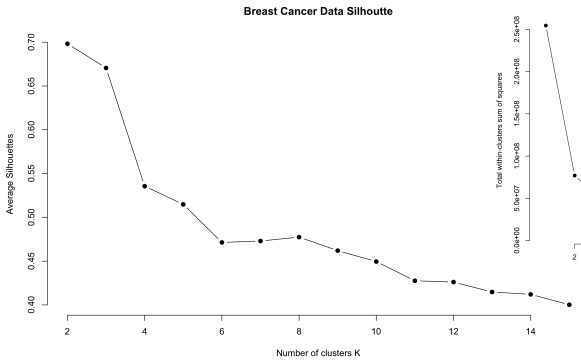


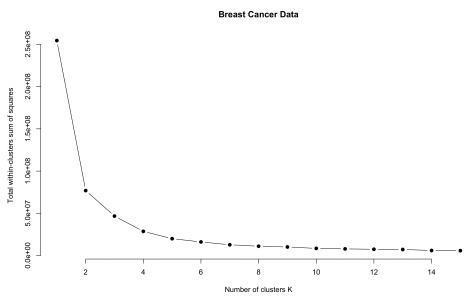
We can observe from all the above 3 plots that the number of clusters suggested by k-means are 2 as by elbow method we can that the percentage variance for ratio of betweenss/totalss and for total within-cluster sum of squares does not vary much as we change the number of clusters from 2 to 3. Also, by silhouette method we can that the maximum value is for k=2. So, we can say that the 4 features suggested by Information gain did help K-means to cluster voice data correctly even though they did not help EM to form correct distributions. A reason for this could be that since EM works on soft clustering so it is sensitive to less variance as well .So, it will require features that have maximum variance so that it can eliminate soft clusters which are formed on the basis of less variance among features. However, as k-means does hard clustering so it clusters data points with less variance in the same cluster and thus do not form separate clusters for such data points. So, we can say that EM surely requires features with maximum variance so that it can cluster correctly but k-means works fine even with simple feature selection with high information gain for voice data.

Breast Cancer data

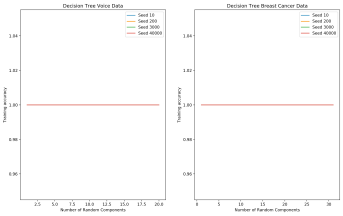
We can observe from all below 3 plots that the number of clusters suggested by k-means are 2 as by elbow method we can that the percentage variance for ratio of betweenss/totalss and for total within-cluster sum of squares does not vary much as we change the number of clusters from 2 to 3. Also, by silhouette method we can that the maximum value is for k=2. So, we can say that the 10 features suggested by Information gain did help K-means to cluster breast cancer data correctly even though they did not help EM to form correct distributions.



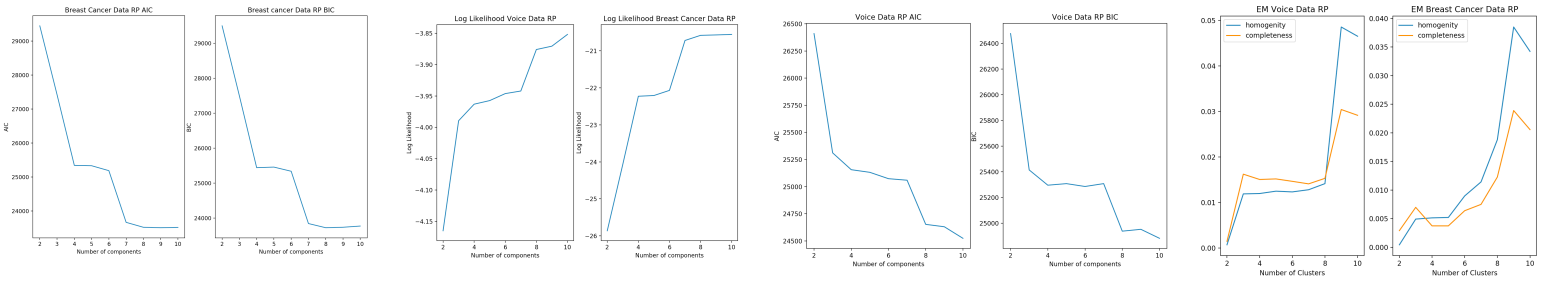




Randomized Projections



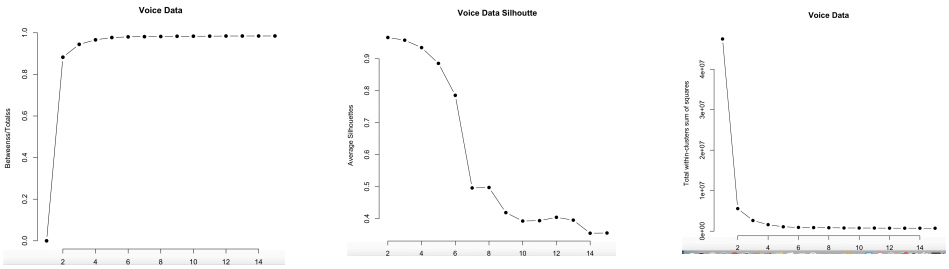
Randomized projections transforms features in directions which are completely random. I used decision tree as a sample classifier to choose for the appropriate number of projections. However the results were weird I got the same accuracy with whatever the number of random projections I chose . So, I just transformed the features to two components because the accuracy was the same irrespective of the number of projections with different seeds. I used 4 different seeds and varied the number of projections from 2 to 11. As cane seen from the below plot of accuracy vs number of random projections as the accuracy does not vary so I just used 2 random projection to perform EM



Voice Data - Looking at AIC (plot 5) BIC(plot 6) , log likelihood(plot 3) and homogeneity and completeness(plot 7) using elbow method we can say that it suggests that there are 4 soft clusters for voice data using elbow method. However, this does not align with the ground classification of voice data. So, we can say that the two randomly projected features are not helping EM to correctly classify this data. As we have seen previously that EM performs the best for Voice dat with PCA suggested features because those features have maximum variance among them and those are the kind of features that are required for classifying the voice data since the data points are very close to each other and these randomly projected features may or may not have maximum variance among them.

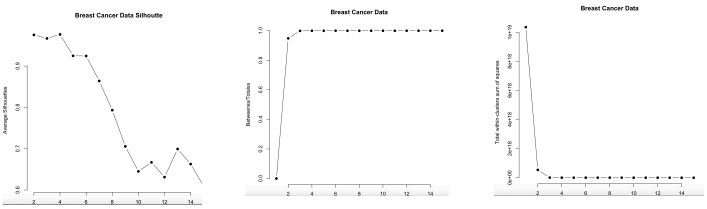
Breast Cancer data - Looking at AIC (plot 1) BIC(plot 2) , log likelihood(plot 4) using elbow method we can say that it suggests that there are 4 soft clusters for breast cancer data using elbow method. However, this does not align with the ground classification of breast cancer data . So, we can say that the two randomly projected features are not helping EM to correctly classify this data. However the homogeneity and completeness graphs show that as we increase the number of distributions from 3 to 4 then there is a considerable drop in homogeneity and completeness and after that there is just overfitting of the data.

K-means: 1. Voice Data



I used 5 random projections as input features for K-means and we can see the 5 randomly projected features help k-means to very well identify the two ground classification clusters as using the elbow for Total within cluster ss and ratio of Betweenss/Totalss the number of clusters suggested are 2 .

2. Breast Cancer Data

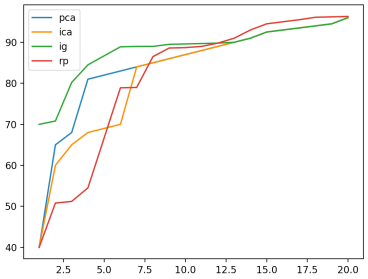


I used 5 random projection as input features for breast cancer data as well. As we can see from the Total within cluster ss and Betweenss/Total ss the breast cancer data has been correctly classified into its ground classification of 2 . This is probably the best result that I have got for breast cancer data. So, this suggests that for assessing breast cancer data randomly transformed 5 features are required for forming clusters which align with the ground classification.

Neural Network

Dimension reduction
I used Weka to run neural network

I am using breast cancer dat to train the neural network after dimensionality reduction of the original dataset. I am performing forward search to train the neural network i.e. I increase the number of components(Attributes) from 1 to 20 one by one and measure the accuracy. I am using same values for momentum and learning rate i.e. 0.4 and 0.2 which I used while running neural network for assignment 1 . These values gave the highest accuracy when I ran neural network for assignment 1 . The number of hidden layers used is the default (attriutes+Classes)/2=11 . I measured training accuracy and training time in seconds to assess performance. Training accuracy for the dataset without dimension reduction is 96.2. From the plot and the tale we can see that the accuracy is maximum with the original dataset but information gain also gave comparable accuracy but with a greater training time. From the plot we can that for PCA to achieve a high accuracy only 4 components are required whereas for ICA 7 are required, For RP we can say 6 are required and for IG 7 are required. Information gain has the maximum training time and PCA is the fastest. Random Projection also has a comparable training time. Thus we can say we can use IG for comprabale performance and PCA for faster training.



	Training accuracy percentage	Training time in milli seconds
Original dataset	96.25	1000.5
ICA	95.2	990.2
IG	96	1131
PCA	94.25	902.5
Random Projection	94.3	1127

Clustering as dimensionality reduction
I used AddCluster for using Cluster as an additional attribute and Cluster Membership for using cluster as the only attribute.

Cluster as additional attribute	Training time(milliseconds)	Training accuracy	Cluster as only attribute	Training time (milliseconds)	Training accuracy
Cluster number =2 KMeans	1002.3	96.5	Cluster number =2 KMeans	-	-

Cluster as additional attribute	Training time(milliseconds)	Training accuracy	Cluster as only attribute	Training time (milliseconds)	Training accuracy
Cluster number=5 kMEANS	1120.5	97.2	Cluster number=5 kMEANS	-	-
Cluster number =10 KMeans	1120.6	98.4	Cluster number =10 KMeans	-	-
Cluster number =2 EM	1003.4	96.6	Cluster number =2 EM	-	54.3
Cluster number=5 EM	1004.5	97.4	Cluster number=5 EM	-	60.4
Cluster number =10 EM	1009.5	98.6	Cluster number =10 EM	-	70.2

We can see that adding clusters as an additional attribute increases the accuracy with both kmeans and. EM and we can say that the training time is still comparable to what we got for original dataset. However when we add clusters as the only attribute then the the accuracy decreases considerably as compared to the baseline accuracy of 96.25 . Thus we can say that adding clusters as an additional attribute is a profitable method to attain higher accuracy for this particular dataset. KMeans was not present as a fileer for ClusterMembership and only EM was available.

Conclusion

We can conclude that for voice data the best clusters are obtained when PCA is used as the dimensionality reduction algorithm because as stated above several times the data points for voice data are very close to each other so it requires features which can distinguish these close points in dimensional space . As PCA provides features in a new dimensional space which has maximum variance so PCA works very well for this dataset. However for Breast cancer data it has been observed none of independent components or components with maximum variance(from PCA) nor simple feature selection using Information gain provides clustering which aligns with ground classification. However, 5 randomly projected features gave the best correspondence to the ground classification.

It has also been observed that if EM is used with dimensionality reduction algorithms for Voice data then it always suggests 4 to 5 clusters this is also probably because of the fact because data points are close in dimensional space so EM is unable to choose a single distribution for majority of the data points.

Datasets - Voice dataset-It has 20 attributes and one class,. It has 3168 instances. This dataset identifies whether the voice is of female or male on the basis of 20 attributes like mean frequency, spectral flatnesss etc. This dataset is interesting because it can be used for voice recognition by AI bots to give a personalised exporience to the user by automatically addressing the user as him/her on the basis of their gender recognized by the 20 attributes and even wrt to Machine Learning it is interesting because it has data points which are very close to each other so it is interesting to see how unsupervised learning algorithms will be able to distinguish among them

Breast cancer dataset-consists of 31 attributes on the basis of which it is decided if the cancer is Malignant or benign. It has 569 instances . This is an interesting dataset because it covers attributes like radius of the cell nucleus , texture , perimeter , area , compactness etc which can help in the diagnosis of breast tumor with the help of images of the cancer from the patients. This data can be used in the medical field by doctors to diagnose patients for breast cancer. This dataset is interesting wrt to machine learning because it has very few instances so it can be a good test for the models to see how well they perform on a dataset which is not large