Problem 1: Spectral Clustering

The Basic Algorithm:

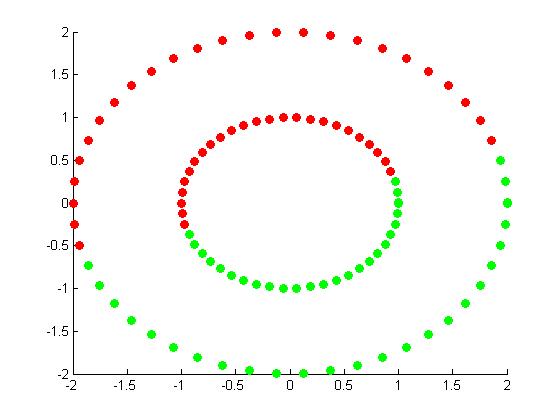
The algorithm is implemented in the Matlab file attached along with

A Simple Comparison:

Spectral Clustering for different values of sigma:

|  |  |
| --- | --- |
| Sigma | Scatter plot |
| 10 | C:\Users\Harshil\Downloads\sprectral_circle_10.jpg |
| 1 | C:\Users\Harshil\Downloads\spectral_circle_1.jpg |
| 0.1 | C:\Users\Harshil\Downloads\spectral_circle_0.1.jpg |

K-Means:



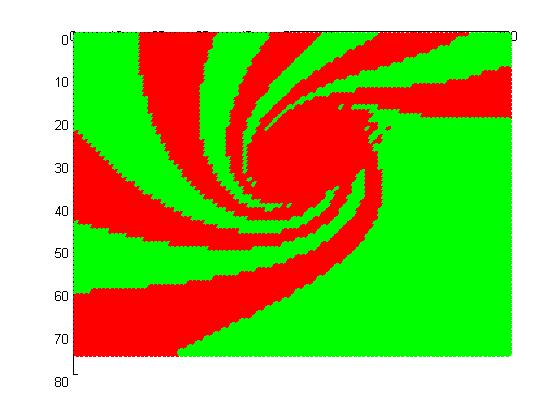
For sigma equal to 0.1 the spectral clustering outperforms k-means. As in spectral clustering the two concentric circles are clustered where as in k-means no matter what initial cluster mean you take, you will not separate the points into concentric clusters. This is because in 2dimension k-means with k=2 will divide the points by a line, thus concentric clusters will not be possible.

Partitioning Images:

Spectral Clustering in Image ‘bw.jpg’. We get the following scatter plots for different values of sigma

|  |  |
| --- | --- |
| Sigma | Scatter Plot |
| 10 | C:\Users\Harshil\Downloads\spectral_10.jpg |
| 1 | C:\Users\Harshil\Downloads\spectral_1.jpg |
| 0.1 | C:\Users\Harshil\Downloads\spectral_0.1.jpg |

K-means:



For higher values of sigma the center part of the image is all clustered into one cluster, but as the sigma is decreased the partitioned image is very similar to the original image. When Sigma is 0.1 it outperforms k-means.

Problem 2: Naïve Bayes and Logistic Regression

The training and test dataset both have attributes where the data is missing. One of the approach as used in the mixture models is that we can use all the information available in the dataset, and then omit the missing values of the attributes by not considering them into the naïve bayes model.

So the MLE estimate for Naïve Bayes will now be:

No there is no need to use EM algorithm in this case.

The above approach resulted in giving 88.7574% accuracy on the test data.

No the same approach cannot be extended in logistic regression as there is no way we know where to move when the missing value attribute comes.

The other alternative approach is to ignore the data point where there is a missing value. For this approach the accuracy in naïve bayes is 94.7917% on the test set and in logistic regression 93.75%.

One more approach is to replace the missing values with its mode (nominal) or mean (continuous). For this approach the accuracy in naïve bayes is 92.7083% on the test set and in logistic regression 90.62%.

This two above approaches gave a better test accuracy than the previous naïve bayes classifier.

Pros and cons of previous naïve bayes classifier is that we take all of the data values into consideration thus covering all the information we have. This might result into overfitting.

Pros and cons of the approach of ignore missing value data points is that we do not take unclear information in learning the classifier and stick to the information that is complete and proper. But the coon here is that we get very less complete information in real life application like in surveys where many participants avoid not give giving information they feel not giving it.

In the approach of mean/mode we take more information in to consideration for learning the classifier but it’s a bit greedy to use mean/mode. When there are many different possible nominal value for an attribute it is not the best approach to replace it with the most frequently occurring one.

To avoid overfitting we can use priors. The strategy that can be used to find a reasonable Dirichlet prior in the naïve bayes is by using cross validation. Use the training set to cross validate the alphas of the Dirichlet prior.