**Name:** Het Piyush Sheth **GTID:** 903393644

CS 7641-HW2

Introduction (Same as HW2\*)

Dataset description

The two datasets that I used for this analysis are QSAR Biodegradation Dataset and Digits Dataset.

Before I describe the datasets itself, I want to talk about why these datasets and why not the datasets that I used previously. Earlier, I used the Adult and Breast Cancer dataset. Now, both of these were binary classification problems and I wanted to explore the possibility of multiple classes in my analysis (thus the digits dataset). Now, I ran my clustering algorithm on the digits dataset and compared my results with the original classes and the performance was quite good. Same was observed with the cancer dataset. But when I tried it on the Biodegradation dataset, the performance wasn’t as impressive. And thus, I chose to swap in the biodegradation dataset for the Cancer one. This is because a poor initial performance means that there is a greater scope for improvement and thus a more enriched analysis can be performed on it.

The Biodegradation dataset has 1055 instances with 41 features and two output classes (biodegradable and not biodegradable). The 1055 instances correspond to different chemicals and each feature signifies a particular property of the chemical. The interesting part about this problem is that the clusters are not completely disjoint, thus reasoning out the initial poor performance. I believe that this will help enrich my analysis and provide some interesting results.

The Digits dataset has 1797 instances with 64 features and 10 classes (digits from 0 to 9). This dataset is derived from an image dataset of the handwritten digits. An image is split into 8x8 grids and each cell within that grid is considered as a feature (thus giving 8x8 = 64 features). In addition, there are approximately 180 instances of each digit adding up to a total of 1797 instances. This means that the dataset is balanced and produces good results on the initial algorithm. Thus the stark contrast between the two datasets will be interesting to analyze.

Clustering

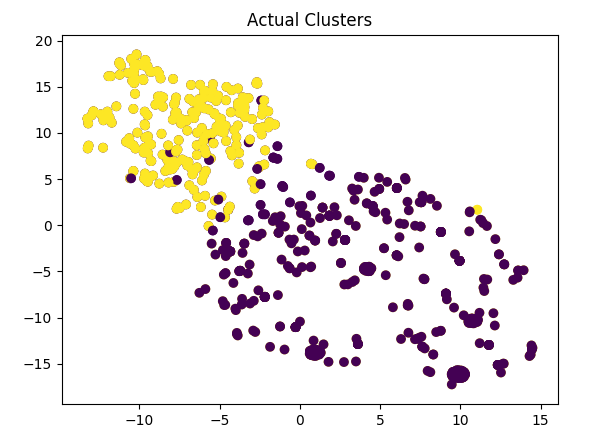
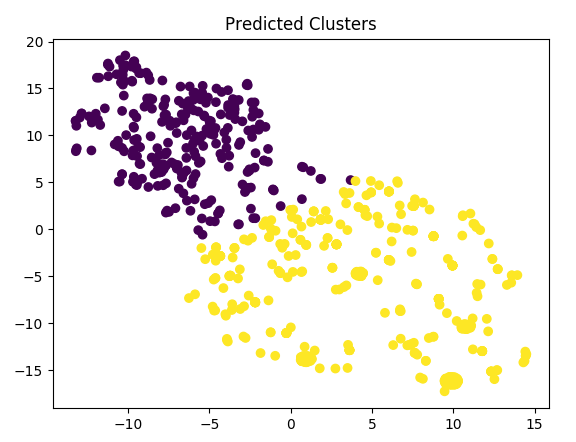
Clustering refers to forming groups from the given data points. The points that belong to the same cluster are supposed to be similar to each other than the ones far apart. It comes under unsupervised form of learning. The two clustering algorithms that I analyze in this report are k-means and Expectation Maximization.

K-means

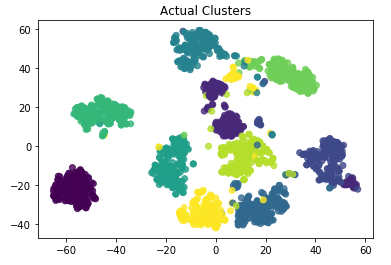
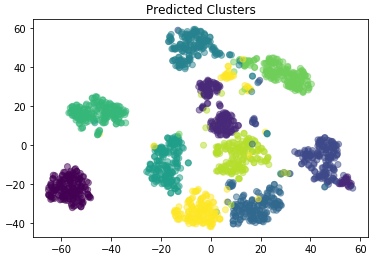
K-means algorithm segregates the data into k different clusters. Initially, it starts with random assignment of clusters but over time it iterates by taking centroids of clusters and re-computing the distances between data points and centroids of each cluster to form new clusters. Ideally, this continues iterating until it reaches a local optimum. I tried this algorithm on the two datasets and used several metrics to measure the performance on each dataset.

For both the datasets, I used Euclidian distance as a measure of the distance used by the K-Means algorithm. This is because the feature values for both the datasets are numerical values itself. I would consider using a different distance measure if the values were categorical or ordinal in nature.

Although, this is an unsupervised learning problem, I used the dataset with the labels as it would help me analyze results. First, I performed K-means on the breast cancer dataset and plotted out the actual vs predicted clusters. Overall, the performance of the algorithm is good but it fails to correctly segregate a few of the data points that lie in the vicinity of the second cluster. This is because K-means is primarily a distance based clustering algorithm. Thus if data points are sparsely placed as for the bottom left cluster in the image, then K-means might not give the best results.

A similar observation is made for the digits dataset. That is, the clusters that are less spread out are better obtained by the K-means algorithm (sea-green, purple) as opposed to the ones that are close together (green, light-blue and purple). Conclusively, we can observe 10 distinct clusters in the predicted graph. Thus, indicating that overall, the algorithm performs pretty well.

In order to evaluate the performance of the K-means algorithm I analyzed the Sum of Squared Errors (SSE) and Adjusted Mutual Index. I chose SSE as a metric of evaluation because K-means is inherently a distance based clustering algorithm. And thus SSE is the best way to measure the sparsity of the data points with respect to the clusters. In addition, mutual information represents the closeness between two clusters. The greater the value of MI, the closer are the clusters and vice versa. Thus, an ideal algorithm should produce clusters with low mutual information.

Below I have plotted the SSE and adjusted mutual information for the cancer dataset. We can see that for both the graph, there is a steep decrease initially and slow decrease after that. Now, eventually if we set the value of the number of cluster equal to that of the number of data points, then the error will be zero. But this only means that we are considering every data point as a cluster itself and thus, we are clearly overfitting the dataset. So our ideal number of clusters would be the point where the steep decrement stops and slow decrement starts as we treat the slow decrement as a case of overfitting. This method is also known as the elbow curve method. Clearly, from the graphs below, we can see that the ideal number of clusters based on SSE is somewhere between 0 and 5 (which is correct as the actual number of clusters should be 2). At the same time, mutual information keeps on decreasing steeply for a while greater than the SSE, this is because the clusters are not inherently separable (i.e. they have overlap data points). Thus, even though the best number of clusters that can be formed is true, the mutual information does not decrease for a while.

