CSE 512-HW4: Machine Learning Report and README

Name: AJAY GOPAL KRISHNA SBU ID: 112688765

1 Programming

1.1 Results

The below results was obtained when clustering using $k = 2$ and using euclidean distance metric. The results for 3 executions is shown

Run-1	$Y = 0$	$Y = 1$	% of Positive Diagnosis
Cluster-0	122	2	1.61
Cluster-1	90	355	79.775

Run-2	$Y = 0$	$Y = 1$	% of Positive Diagnosis
Cluster-0	90	355	79.775
Cluster-1	122	2	1.61

Run-3	$Y = 0$	$Y = 1$	% of Positive Diagnosis
Cluster-0	92	355	79.418
Cluster-1	120	2	1.639

From the above results we can clearly see that, clustering is not prefect based on diagnosis. We are able to find $Y = 0$ and $Y = 1$ values in both the clusters.

But majority of the points with $Y = 1$ i.e data points with positive diagnosis are falling into a single cluster (% positive diagnosis in the other cluster is small around 1.6%) and we see that this cluster has false positives as well ($Y = 0$ is divided among both the clusters).

The centroid values using euclidean distance for single run with k=2 are
Centroid-0: [$1.96183065e{+}01$, $2.18419355e{+}01$, $1.29708871e{+}02$, $1.21193629e{+}03$, $1.00311290e{-}01$]
Centroid-1: [$1.25972112e{+}01$, $1.85784494e{+}01$, $8.14527640e{+}01$, $4.99666966e{+}02$, $9.52593258e{-}02$]

2 README

kmeans.py supports the below parameters:

Option	Description	Default-Value
-h, –help	show this help message and exit	
-d, –dataset	***path_to_dataset***	Breast_cancer_data.csv
-k, –kcluster	K-Clusters	2
–distance	Distance Function used. euclidean/manhattan	euclidean
-e, –epsilon	Epsilon for Change in Centroids	0.0001

To run the k-means algorithm with $k = 2$ and euclidean distance execute
python kmeans.py -d path_to_data -k 2 –distance euclidean

Note not specifying the "−−distance" option will default to Euclidean distance

To run the k-means algorithm with $k = 2$ and manhattan distance execute
python kmeans.py -d path_to_data -k 2 –distance manhattan