Lab 06 – Clustering and Lazy Classification

Clustering

Weka provides a number of clustering algorithms. We are going to try three of them using the iris and diabetes datasets.

1. Fire up Weka, launch the explorer window and select the “Pre-process" tab. Open the iris dataset. Select the “Cluster” tab to get into the clustering perspective of Weka. Under “Clusterer” select and run the clustering algorithms - *EM, farthest first* and *simple K-means* - in turn.
2. Note that:

   - under “Cluster mode” leave “Use training set” highlighted;

   - under “Cluster mode” select “Classes to clusters evaluation”;

   - For each algorithm, make sure you set the number of cluster to 3 reflecting that there are three distinct class values.

3. In the “Classes to clusters evaluation” Weka ignores the class attribute and generates the clustering. Then during the test phase, it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. Then it computes the classification error, based on this assignment it also shows the corresponding confusion matrix. For each clustering algorithm inspect the confusion matrix and classification error. Which algorithm gives the best result?
4. Under the “Results list” you have three entries, right-clicking on an entry will give you the option to “Visualise cluster assignments”. Inspect each of the entries for the plot (varying the x- and y- axes). Can you see why some of the instances are in the wrong cluster?
5. Repeat steps 1 to 3 above using the diabetes data set. Remember to set the number of clusters to 2 instead of 3 because there are only two outcomes – tested positive and tested negative.

Lazy Classifier

We are going to try a lazy classifier, namely IBk, that implements the k nearest neighbour method.

1. Fire up Weka, launch the explorer window and select the “Pre-process" tab. Open the iris dataset. Select the “Classify” tab.
2. Choose the lazy classifier IBk
3. Click ‘Start’ and note the result presented in the confusion matrix.
4. Right click on the ‘Choose’ window and change the value for K.
5. Click ‘Start’ and note the result presented in the confusion matrix.
6. Repeat steps 4 and 5 several times. What is the best value for K?
7. Repeat steps 1 to 6 using the diabetes data set. What is the best value for K?