CS 345: Data Mining Names:____

Assignment 5 ____

1. The *k*-means algorithm

The *k*-means algorithm is a simple, straightforward algorithm to assign instances to clusters. Each cluster is defined by a cluster centroid, and instances belong to the cluster for which their Euclidian distance to the centroid is the smallest. For each cluster a new centroid is found by taking the average over the cluster instances, which may lead to shifts of instances between clusters. This iterative process ends when the centroids stop changing.

Start the Weka Explorer and on the Preprocess page load the file “iris.arff”.

a. Study the data set to remind yourself of the attributes and what each attribute is doing. What’s the most natural way to cluster this dataset?

Go to the Cluster tab. In the Clusterer box, click the “Choose” button. You can now select the SimpleKMeans” clusterer. In the cluster mode box, choose “Use training set”. Change the number of clusters to 3, and click Start.

b. What do you notice about the results? How good are the results? Are they too good to be true? What might be causing this?

To fix the above issue, click on the Ignore Attributes button and select class. Then rerun the clustering.

c. Are your results as expected? What are your results (how many instances fall into each cluster?) Since we know that there are 50 instances of each class in the dataset, what is the minimum number of incorrectly clustered instances?

Another way to “fix” this issue and to also gain more information is to do the following. In the cluster mode box, change to use “Classes to clusters evaluation”, and be sure that Nom (class) is highlighted in the dropdown box. This will make the clusterer ignore the “class” attribute, but will use this attribute for comparison in the generated report. Click “Start” to activate the clustering.

d. Are your results as expected? What are your results (how many incorrectly clustered instances were there)? Did you get the same results as you did in part c?

Click on the “Clusterer” box (the text right next to the “Choose” button) and change the seed, but leave the parameter k (which is the number of clusters) at 3. Changing the seed will make the clusterer use different random numbers. Rerun the clustering.

e. Study the results. Are they equal to the results you got previously? If so, try a couple more times until the results change. Why are there slightly different results depending on the seed?

To visualize your clusters, right-click on the last SimpleKMeans run that you did and select Visualize Cluster Assignment. For the x-axis select Instance Number and for the y-axis, select Class. Play around with the settings until you’re able to best see the visualization.

f. Based only on the visualization, how many instances are incorrectly clustered?

If you like, you can apply the k-means algorithm to some of the other datasets. Be warned that if you use a dataset with a great many instances and ask the algorithm to split it into a dozen or more classes, this task can take a very, very long time (so, stay away from the “Soybean” data set).

2. The Expectation-Maximization (EM) algorithm

The EM algorithm is a probabilistic clustering algorithm. Each cluster is defined by probabilities for instances to have certain values for their attributes, and a probability for instances to reside in the cluster. For numerical values it consists of a mean value and a standard deviation for each attribute value, for discrete values it consists of a probability for each attribute value.

Because discrete values are easier to evaluate in this respect, and also to facilitate comparison with the previous algorithm, we will apply the EM algorithm again to the "iris" data set.

Go to the Cluster page. In the Clusterer box, click on the text you see there. You can now select the "EM" clusterer. For "Cluster mode" select "Classes to clusters evaluation". This will make the clusterer ignore the "class" attribute, but will use this attribute for comparison in the generated report. In the default setup, the EM algorithm will determine the number of clusters automatically. Click "Start" to activate the clustering.

a. Study the results. How many clusters are generated? Why is this? Can you get a different result with a different seed?

When you click on the "Clusterer" box you can change the "numClusters" value. This value is -1 by default, which allows the algorithm to determine by itself the needed number of clusters. If you set it to a specific value, the algorithm will try to derive that number of clusters.

b. Change the number of clusters to the number of clusters you'd like to get (or a bit higher) and rerun the experiment. Also rerun the experiment with different seeds (and a desired number of clusters). How does this impact the results?

c. Did the EM algorithm get better, the same, or worse results than k-means?

3. Using the zoo.arff dataset from today’s Classification Weka activity (on course website), answer the following questions.

a. How many animal types are represented in this dataset?

b. Start using the SimpleKMeans clusterer choosing 7 clusters. How well does this algorithm cluster the data? Think about the best way from above to tell this. Which cluster mode did you use?

c. Compare results with EM clusterer (with 7 clusters), MakeDensityBasedClusterer (this is a wrapper algorithm so have it use SimpleKMeans and set to 7 clusters), FarthestFirst (with 7 clusters), and HierarchicalClusterer (7 clusters). Which algorithm seems to provide the best clustering match for this dataset?

Another option for clustering is to use the ClassificationViaClustering meta-classifier. This allows you to evaluate your clusters using the same metrics as you would for other classification techniques. It ignores the classes, clusters, then assigns classes based on the clusters.
To use this algorithm, you must first install it. Go to the Weka GUI Chooser and click on Tools -> Package Manager. In the Package Search, type in ClassificationViaClustering. Close all other Weka windows, then click install. Once it’s installed, reopen Weka Explorer. Reload the zoo.arff file and go to the Classify tab. Choose meta.ClassificationViaClustering, then select it to use SimpleKMeans with 7 clusters. Choose Cross-Validation in the Test Options area.

d. How many instances were correctly classified using this technique? Compare this with your favorite Classification technique (pick any classification algorithm in Weka and tell me which one and how well it did.)