CS 445/545, Machine Learning, Spring 2013

Homework 5: *K*-Means Clustering

Due Wednesday June 12, 2:00pm.

In this homework you will experiment with classification of handwritten digits using *K*-means clustering.

You will be using the Optdigits data set from Homeworks 1 and 4.

K-Means Clustering

Write a program to implement *K*-Means clustering using Euclidean distance, and to evaluate the resulting clustering using sum-squared error, cohesion, separation, and entropy.

Experiment 1: Repeat the following 5 times, with different random number seeds. Run your clustering program on the training data (optdigits.train) with $K = 10$, obtaining 10 final cluster centers. (Remember not to use the *class* attribute in the clustering!) Your initial cluster centers should be chosen at random, with each attribute A_i being an integer in the range [**0,16**].

Stop iterating *K*-Means when all cluster centers stop changing or if the algorithm is stuck in an oscillation.

Choose the run (out of 5) that yields the smallest sum-squared error (SSE).

- For this best run, give the sum-squared error, average cohesion, average pair-wise cluster separation, and entropy of the resulting clustering. (See the slides for definitions.)

- Now use this clustering to classify the test data, as follows:
  - Associate each cluster center with the most frequent class it contains. If there is a tie for most frequent class, break the tie at random.
  - Assign each test instance the class of the closest cluster center. Again, ties are broken at random. Give the accuracy on the test data as well a confusion matrix.
  - **Note:** It's possible that a particular class won't be the most common one for any cluster, and therefore no test digit will ever get that label.

- Visualize the resulting cluster centers. That is, for each of the 10 cluster centers, use the cluster center's attributes to draw the corresponding digit on an 8 x 8 grid.

(You can do this using any matrix-to-bit-map format – e.g., pgm: http://en.wikipedia.org/wiki/Netpbm_format#PGM_example)

- Write a paragraph describing your results and your visualization, and compare the accuracies you obtained here on the test set with those you obtained on Homeworks 1 (decision trees) and 4 (naïve Bayes). Do the visualized cluster centers look like their associated digits?

Experiment 2: Run *K*-means on the same data but with *K* = 30. Give the entropy of the resulting clustering. Also give a visualization of the resulting 30 cluster centers, as in Experiment 1. Write a paragraph describing your entropy and visualization results, and how they compare with those of Experiment 1.

Here is what you need to turn in:

Your spell-checked, double-spaced report with the information requested above. Also, your commented *K*-Means code and code for calculating cohesion, separation, and entropy, with instructions how to run it.

How to turn it in:

Send these items in electronic format to mm@cs.pdx.edu by 2pm on the due date. No hard-copy please! Your report should be in pdf format.

If there are any questions on this assignment, don't hesitate to ask me, Max (our TA), or e-mail the class mailing list.

Policy on late homework: If you are having trouble completing the assignment on time for any reason, please see me before the due date to find out if you can get an extension. Any homework turned in late without an extension from me will have 5% of the grade subtracted for each day the assignment is late.