CS 445/545

Machine Learning

Winter 2016

Homework 5: *K*-Means Clustering

Due Tuesday, March 8, 2016, 2pm

In this homework you will experiment on using the *K*-means clustering algorithm to cluster and classify the OptDigits data, originally from the UCI ML repository.

Dataset: Download the data from our class website or from this link:
http://web.cecs.pdx.edu/~mm/MachineLearningWinter2016/optdigits.zip
Recall from the description in class that each instance has 64 attributes, each of which can have value 0−16.

Code to write: Write a program to implement *K*-means clustering using Euclidean distance, and to evaluate the resulting clustering using sum-squared error, sum-squared separation, and entropy. (Note you must write your own code for this, not use an existing package.)

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, in your report give the sum-squared error, sum-squared separation, and mean entropy of the resulting clustering. (See the class 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.

- Calculate the accuracy on the test data and create a confusion matrix for the results on the test data.

- 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)

- In your report, include the following:

  - Sum-squared error, sum-squared separation, and entropy of your resulting clustering in the best run of the five you did.

  - Classification accuracy on the test data and the confusion matrix over the test data (for this same run).

  - Visualization results (for this same run).

  - Discussion paragraph: Summarize your results. Answer: Do the visualized cluster centers look like their associated digits?

Experiment 2: Run *K*-means on the same data but with $K = 30$. In your report, include the same things that were asked for in Experiment 1, and in your discussion paragraph, compare the results of Experiments 1 and 2.

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 with instructions how to run it.

How to turn it in (read carefully!):

- Send these items in electronic format to mm@pdx.edu by 2pm on the due date. No hard copy please!
- The report should be in pdf format and the code should be in plain-text format.
- Put "MACHINE LEARNING HW 3" in the subject line.

If there are any questions, don't hesitate to ask me 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.