CIS400/600 Fundamentals of Data and Knowledge Mining

Homework #4

Spring, 2017

Problem 1 (18 pts): Please apply Agglomerative Clustering Algorithm to establish a hierarchical grouping structure using the below one-dimensional training dataset:

ID	x1
y_1	1
y_2	5
y_3	8
y_4	10
y_5	2

a) Using the complete linkage method with Euclidean distance to measure inter-cluster distance, please show all the steps to infer a hierarchical cluster from the training dataset.
b) Please draw the dendrogram graph showing the sequence of how subclusters are merged together with the lifetime information (defined as the difference between the distances at two successive nodes).
c) How many clusters will you discover if you use the maximal lifetime as the cutting criterion and what are the cluster memberships for all the data points in the training set?

Problem 2 (12 pts): Hierarchical clustering algorithms require $O(m^2 \log(m))$ time, and consequently, are impractical to use directly on larger data sets. One possible technique for reducing the time required is to sample the data set. For example, if K clusters are desired and $\sqrt{m}$ points are sampled from the $m$points, then a hierarchical clustering algorithm will produce a hierarchical clustering in roughly $O(m)$ time. K clusters can be extracted from this hierarchical clustering by taking the clusters on the Kth level of the dendrogram. The remaining points can then be assigned to a cluster in linear time, by using various strategies. To give a specific example, the centroids of the K clusters can be computed, and then each of the $m - \sqrt{m}$ remaining points can be assigned to the cluster associated with the closest centroid.

For each of the following types of data or clusters, discuss briefly if (1) sampling will cause problems for this approach and (2) what those problems are. Assume that the sampling

technique randomly chooses points from the total set of m points and that any unmentioned characteristics of the data or clusters are as optimal as possible. In other words, focus only on problems caused by the particular characteristic mentioned. Finally, assume that K is very much less than m.

(a) Data with very different sized clusters.

(b) High-dimensional data.

(c) Data with outliers, i.e., atypical points.

(d) Data with highly irregular regions.

(e) Data with globular clusters.

(f) Data with widely different densities.

Problem 3 (20 pts): The data set attached comes from the Kaggle Digit Recognizer competition. The goal is to recognize digits 0 to 9 in handwriting images.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive. The provided dataset has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. Each pixel column in the training set has a name like pixelx, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as x = i * 28 + j, where i and j are integers between 0 and 27, inclusive. Then pixelx is located on row i and column j of a 28 x 28 matrix, (indexing by zero).For example, pixel31 indicates the pixel that is in the fourth column from the left, and the second row from the top.

Apply HAC and k-Means clustering algorithms to the digit recognition data set (as provided), and answer the following questions:

a) Using Weka's "classes to clusters evaluation", what is the best accuracy you achieved by using k-Means to predict the digits? What parameter tuning helped improve the accuracy? Then examine the confusion matrix of your best clustering model: what digits are easy to distinguish and what digits look confusing to the k-Means algorithm?
b) Repeat the same analysis using the HAC algorithm. For your best HAC model, which digits are considered similar? Hint: by setting the number of clusters to a number smaller than 10, such as 3, the model will be forced to cluster the digits to three clusters, see what digits are clustered together.
c) Do you think the two models captured the same patterns to distinguish the digits, or did they find different patterns? Why?