CSI 5387: Data Mining and Concept Learning

Winter 2020

Assignment 3

Submission Deadline: 22nd March 2020 on Brightspace.

Part A: Association Analysis (30 points)

Perform a market basket analysis of transactional data on the *groceries* dataset provided using Weka, R or Python.

(a) With the Apriori algorithm, evaluate the set of association rules. Generate association rules using minimum support of 0.5% minimum confidence of 50%, and length of 2. Display the rules, sorted by descending lift value.

(b) Select the rule from (a) with the greatest lift. Compare this rule with the highest lift rule from using minimum support of 0.1%, minimum confidence of 50% and length of 2. Which rule has the better lift?

(c) If you were a marketing manager, and could fund only one of these rules, which would it be, and why?

Part B: Clustering (70 points)

1) Given the points $x_1 = \{1, 0\}$, $x_2 = \{0,1\}$, $x_3 = \{2, 1\}$, and $x_4 = \{3, 3\}$, suppose that these points are randomly clustered into two clusters: $C_1 = \{x_1, x_3\}$ and $C_2 = \{x_2, x_4\}$ (10 points).
   a) Apply one iteration of k-means partitional clustering algorithm and find new distribution of elements in clusters. What is the change in sum of square errors?
   b) Apply the second iteration of the k-means algorithm and discuss the changes in clusters.
2) Use the similarity matrix in Table 1 to perform complete link hierarchical clustering (assuming Euclidean distance is used). Show the steps for constructing the dendrogram. Plot the graph, which should clearly show the order in which the points are merged (20 points).
3) If DBSCAN algorithm is applied with *eps* of 0.5, and $MinPts \geq 2$ (required density), what are core, border, and noise points in the set of points p_i given in the table. Explain (10 points).

Table 1

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

4) Consider the *Absenteeism_at_work* dataset from UCI machine learning repository provided. Use only the *age* and *workload average/day* fields. Apply three clustering algorithms: k-means, agglomerative, and EM for Gaussian mixture models using Weka, R or Python for the following (30 points):
   a) Determine the optimal number of clusters *k* using the elbow method.
   b) Standardize the attributes, run k-means on the data, using the derived *k* value. Plot the data.
   c) Run agglomerative clustering on the data, using single linkage. Plot the final assignment of the clusters.
   d) Run the EM Gaussian mixture model with the optimal number of *k*
   e) Compare the similarities / differences of the k- means with the other clustering from (c) and (d). Which do you think is the most reasonable?