CIS400/600 Fundamentals of Data and Knowledge Mining

Homework #3

Spring, 2017

Problem 1 (10 pts): Consider the market basket transactions shown in the below table.

Transaction	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support) and why?

(b) What is the maximum size of frequent itemsets that can be extracted (assuming *minsup* > 0)?

(c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

(d) Find an itemset (of size 2 or larger) that has the largest support.

(e) Find two pairs of items, *a* and *b*, such that the rules $\{a\} \longrightarrow \{b\}$ and $\{b\} \longrightarrow \{a\}$ have the same confidence.

Problem 2 (10 pts): In a electronic store transaction data, {*movie*} refers to the transactions containing {*movie*}, $\{\overline{movie}\}$ refers to the transactions that do not contain *movie*, and {*video game*} refers to the transactions containing *video game*. Assume that the confidence of {*movie*} -> {*video game*} is less than the support of {*video game*},

(a) Is confidence of {*movie*} -> {*video game*} greater or less than that of $\{\overline{movie}\}$ -> {*video game*}? Why? Please show your steps.

(b) Is the support of {*video game*} greater or less than the confidence of $\{\overline{movie}\}$ -> {*video game*}? Why? Please show your steps.

Problem 3 (10 pts): Hierarchical clustering is sometimes used to generate K clusters, K > 1 by taking the clusters at the Kth level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: {6, 12, 18, 24, 30, 42, 48}.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

i. {18, 45}

ii. {15, 40}

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

(c) What are the two clusters produced by single link?

(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.)

(e) What well-known characteristic of the K-means algorithm explains the previous behavior?

Problem 4 (20 pts): For this problem, you need to explore the bank data, available on the BlackBoard as a csv file and an accompanying description of the attributes and their values. The dataset contains attributes on each person's demographics and banking information in order to determine they will want to obtain the new PEP (Personal Equity Plan). Your goal is to perform Association Rule discovery on the dataset using appropriate packages in R or Weka.

First perform the necessary preprocessing steps required for association rule mining, for example, the id field needs to be removed and a number of numeric fields need discretization or otherwise converted to nominal. Next perform association rule discovery on the preprocessed data. Set PEP as the right hand side of the rules, and see what rules are generated. Experiment with different parameters and preprocessing so that you get on the order of 20-30 strong rules, e.g. rules with high lift and confidence which at the same time have relatively good support.

Select the top 3 most "interesting" rules and for each specify the following:

- Support, Confidence and Lift values
- An explanation of the pattern and why you believe it is interesting based on the business objectives of the company.
- Any recommendations based on the discovered rule that might help the company to better understand behavior of its customers or to develop a business opportunity.

Note that the top 3 most interesting rules are most likely not the top 3 in the strong rules. They are rules, that in addition to having high lift and confidence, also provide some non-trivial, actionable knowledge based on underlying business objectives.

In more detail, your answers should include:

- Description of preprocessing steps
- Description of parameters and experiments in order to obtain strong rules
- Give the top 3 most interesting rules and the 3 items listed above for each rule
- For at least one rule, manually calculate the support, confidence and lift numbers.