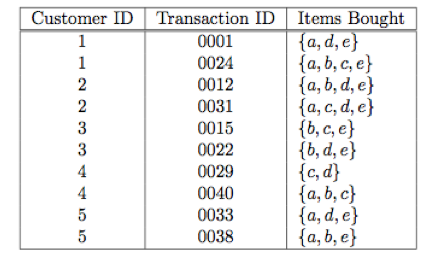
**Week 3 Homework**

**Please provide clear and complete explanations for your answers, and show your work.**

## NAME:

## Week 3 Homework (150 points) – Submit to Dropbox

**Exercises (100 points)**

1. (40 points) Consider the data set of market basket transactions shown in following table:
   1. Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.
   2. Use the results in part (a) to compute the confidence for the association rules {b,d} −→ {e} and {e} −→ {b,d}. Is confidence a symmetric measure?
   3. Repeat part (a) by treating each customer ID as a market basket. Each  item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)
   4. Use the results in part (c) to compute the confidence for the association rules {b, d} −→ {e} and {e} −→ {b, d}.
2. (20 points)
3. What is the confidence for the rules ∅ −→ A and A −→ ∅?
4. Let c1, c2, and c3 be the confidence values of the rules *{p} −→ {q}*, *{p} −→ {q,r},* and *{p,r} −→ {q},* respectively. If we assume that c1, c2, and c3 have different values, what are the possible relationships that may exist among c1, c2, and c3? Which rule has the lowest confidence?
5. (10 points) Suppose that for a data set,

* there are m points and K clusters;
* half the points and clusters are in “more dense” regions;
* half the points and clusters are in “less dense” regions; and
* the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

1. Centroids should be equally distributed between more dense and less dense regions.
2. More centroids should be allocated to the less dense region.
3. More centroids should be allocated to the denser region.

Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

1. (15 points) When a comprehensive training set is available, a supervised anomaly detection technique can typically outperform an unsupervised anomaly technique when performance is evaluated using measures such as the detection and false alarm rate. However, in some cases, such as fraud detection, new types of anomalies are always developing. Performance can be evaluated according to the detection and false alarm rates, because it is usually possible to determine, upon investigation, whether an object (transaction) is anomalous. Discuss the relative merits of supervised and unsupervised anomaly detection under such conditions.
2. (15 points) An analyst applies an anomaly detection algorithm to a data set and finds a set of anomalies. Being curious, the analyst then applies the anomaly detection algorithm to the set of anomalies. What do you think the behavior of an anomaly detection algorithm should be when applied to a set of anomalous objects?

**Lab (50 points):** SAS e-Learning Course: Pattern Discovery Using SAS Enterprise Miner

* P. 15: How do you determine the number of observations in each cluster? (5 points)
* P. 20: (25 points)
  + Provide screen shot of Results window from running the Segment Profile node on the Dungaree data.
  + Provide 3 observations about the segments created.
* P. 32: (20 points)
  + Provide screen shot of Results window from running the Association node on the Transactions data.
  + What is the 3-item rule with the highest lift value?