CSE 4510/5400 Data Mining
Due 5pm, Apr 27
Submit Server: Class = datamining, Assignment = hw5

1. Written assignment (from textbook) [pdf file or hardcopy in class]:
   (a) Ch10, Q10, p682, use different scenarios in your comparison; plus how would you use Silhouette coefficient as an anomaly score?
   (b) Ch10, Q11, p682
   (c) Ch10, Q12, p682
   (d) Related to parts 2(f) and 2(g) of the programming assignment (CSE 5400: for each algorithm and option):
      i. Experiment 1
         A. plot the anomaly scores of all instances in descending order (y-axis: anomaly score; x-axis: rank)
         B. explain your recommendation for j, where top-j are considered anomalies.
      ii. Experiment 2
         A. discuss your expected behavior of AUC when k increases and your reasoning,
         B. plot k vs AUC,
         C. discuss your observed behavior of AUC from your plot when k increases and possible reasons for the difference (if any) in your observed and expected behavior, and
         D. explain your recommendation for k.
   (e) CSE 5400 only: In the spirit of Ch10, Q4, p680, but use Algorithms 6.1-6.3. How would you use association rules for anomaly detection? How would you use their support and/or confidence values to generate an anomaly score?

2. Programming assignment:
   (a) Implement Algorithm 10.2 (p669)
      i. line 7: use the reciprocal of Equation 10.7 [see lecture notes]
   (b) CSE 5400 only:
      i. Implement Example 10.3 (p672)–using k-means, allow both options for the outlier score.
   (c) Allow k and score threshold as user-specified parameters
   (d) Print performance of an algorithm based on:
      i. true-positive (TP) rate and false-postive (FP) rate
      ii. accuracy
   (e) Datasets are on the course website:
      i. toy-anom
      ii. Breast cancer
   (f) Experiment 1: for the Breast Cancer data and $k = 5$, print the anomaly scores of all instances in descending order
   (g) Experiment 2: for the Breast Cancer data, vary k from 3 to 10 and for each k, print:
      i. k,
      ii. pairs of (TP and FP rates) and
      iii. Area under the curve (AUC) of Receiver Operating Characteristic (ROC) Curve (p298-301) [add up the area of trapezoids]
   (h) Implementation:
      i. The same implementation should be able to handle the two different data sets
      ii. Use C (GNU gcc), C++ (GNU g++), Java (Oracle Java), LISP (CLISP), or Pythoan. If you don't have a preference, use Java since it's more portable.
      iii. Your program preferably runs on code01.fit.edu (linux).
   (i) Submission:
      i. README.txt:
         - what the different files are
         - how to compile and run your program with the two data sets with a different k for k nearest neighbors (and for k-means for CSE 5400) (preferably on code01.fit.edu).
         - how to compile and run the experiments parts 2(f) and 2(g) above.
      ii. source code files