Assignment: Clustering & Decision Trees

Due Date: Friday August 7th.

Homework Lecture 5a

The first homework this week is in Lecture 5 notes. See pages 76 – 79 for instructions.

Homework Lecture 5b

An online shopping site has the following primary pages or sections: **Home, Products, Search, Prod_A, Prod_B, Prod_C, Cart, Purchase**. A user may browse from "Home" to "Products" and then to one of the individual products. The user may also search for a specific product by using the "Search" function. A visit to "Cart" implies that the user has placed an item in the shopping cart, and "Purchase" indicates the user completed the purchase of items in the shopping cart. The site has collect the hypothetical session data for 100 sessions. This data is available in CSV format, Sessions.CSV, on course website.

Use WEKA's K-means clustering algorithm to cluster these user sessions into segments. Try different clustering runs with various numbers of clusters (e.g., between 4 and 8), and select the result set(s) that seem to best answer as many of the following questions as possible.

- If a new user is observed to access the following pages: **Home => Search => Prod_B**, according to your clusters, what other product should be recommended to this user? Explain your answer based on your clustering results. What if the new user has accessed the following sequence instead: **Products => Prod_C**?
- Can clustering help us identify casual browsers ("window shoppers"), focused browsers (those who seem to know what products they are looking for), and searchers (those using the search function to find items they want)? If so, are any of these groups show a higher or lower propensity to make a purchase?
- Do any of the segments show particular interest in one or more products, and if so, can we identify any special characteristics about their navigational behavior or their purchase propensity?
- If we know that, during the time of data collection, independent banner ads had been placed on some popular sites pointing to products A and B, can we identify segments corresponding to visitors that respond to the ads? If so, can we determine if either of these promotional campaigns are having any success?

For this problem, you should submit your clustering result summary (including the cluster centroids), the final data set which shows the final assignment of these sessions to clusters, and your answers to the above questions along with your justification based on the clustering results. **Other Notes:** You may also want to use WEKA's cluster visualization capabilities to identify interesting distributions of various page visits among and within clusters.