Intro to Analytics

- What is the difference between OLAP and Data Mining?
- What is data mining? What is it not?
- Why do we need data mining - what is the bottom line?
- What are the different types of data mining tasks?
- For the types of tasks we covered in class (Decision Trees, Clustering, Association Rules), when is it appropriate to use each? What is the goal of each?

Decision Trees and Classification

- What do decision trees aim to do (classification)?
- What type of data do you use this for (binary variables – capturing whether an event occurs or does not occur – e.g., product purchase, loan default).
- How is the classification achieved? What is the general approach?
  - Group data points/observations based on splits in the predictors (e.g., one customer grouping might be > 25 and <= 25 years old – this would be a useful grouping if customers above 25 years of age are significantly more likely to exhibit the behavior you are looking at, than those customers that are less than or equal to 25).
- The concept of a training set and a validation/test set and what they are used for.
  - Take a portion of your data to create the decision tree with. Then see how well it does at predicting the outcomes for observations in the validation dataset.
- Terminology: predictors, classifiers, root nodes, child nodes, leaf nodes.
- How to validate a tree against test data (hit rate).
- What statistics does the software use to justify an additional branch/split (i.e., Chi-square and Gini Coefficient)?
  - What does it mean when a Chi-square grows larger or smaller?
  - What does it mean when a p-value gets bigger or smaller? How is this related to the Chi-square statistic?
  - What does it mean when a Gini coefficient gets bigger or smaller?
- How to read a decision tree based on predictor values (e.g., what is the probability of a behavior given these customer characteristics).

Clustering and Segmentation

- What are the two main uses of cluster analysis?
  - Grouping and Summarizing
- What types of data is this used for?
  - It can be used to identify similar sets of observations (e.g., customers or products) based on their observed attributes/characteristics.
- What is the difference between cluster analysis and manual classification?
• What are the two main types of cluster analysis?
  o Hierarchical and Partition
• What are the pros and cons of having only a few clusters? Of having many clusters?
• What are the steps involved in the k-means clustering algorithm (the one we did in class)?
  o Pick the number of clusters.
  o Pick starting points for the cluster centers.
  o Calculate cluster membership based on each data point’s nearest cluster mean/center (centroid).
  o Re-calculate the cluster mean (centroid).
  o Re-calculate cluster membership (go to last step).
  o Keep iterating until the centroid stops shifting very much.
• What are the two statistics that are used to evaluate clustering results?
  o SSE within cluster, SSE between centroids.
  o Do we want a high SSE within a cluster, or a low one? What does a high within-cluster SSE mean?
  o Do we want a high SSE between clusters? What does a high SSE between clusters mean?
• What does it mean to standardize variables (conceptually)? Why would we do this in cluster analysis?
• What is an outlier? Why do we want to avoid these?
• Post-processing concepts.
  o After you have your cluster results, what are the reasons you might merge two clusters? Split two clusters? Drop a cluster entirely?

Association Rules

• What type of data do you use this for?
  o The co-occurrence of events (e.g., products that are purchased together – Amazon recommendation agent).
• Terminology: market basket analysis, a basket, an item set, Antecedent and Consequent.
• What is a support count? Support percentage? What do these tell you?
  o Frequency of co-occurrence in the data.
• Does item ordering in the association rule impact the calculation of support (no)?
• What is confidence? What does it tell you?
  o Does the ordering of items in the rule impact confidence (yes)?
  o What does a higher or lower confidence mean (higher confidence means stronger support for the association rule)?
• Why might confidence be misleading (could be due to chance, if the events naturally occur with high frequency)?
• What is lift? What does it tell you?
How strong is the rule compared to what random chance would expect.

• What does it mean if lift = 1? What does it mean if lift << 1? What does it mean if lift >> 1?
• What might you do to act upon an example rule if you saw it had very high lift?

Data Visualization

• Understand Tafte’s tenets:
  o Tell a Story
  o Lie Factor
  o Data-ink Ratio
  o Chart Junk
• Don’t imply an ordering when one is not present.
• Don’t plot items that are not comparable (apples and oranges).
• When is a table potentially more useful than a plot?