* Review sheet for Data Mining
* Bootstrapping
* Number of observations per variable
  + 10 observations per variable
    - Categorical variables count as (k-1 of levels) as number of variables
* Dealing with transactional data
  + Figuring out what unit to role up to
* Dealing with Missing values
  + Bins
  + Imputation (be careful of, think of the pros and cons)
* Bonferroni correction
  + Multiply p-values by number of p-values in the collection (and should not go above 1)
* FWER and FDR
  + FWER 🡪 family wise error rate
    - Is the probability of making at least one type 1 error (false positive) among all hypothesis tests conducted
    - In other words it is the probability of incorrectly rejecting at least one true null hypothesis out of multiple tests
    - Goal is to minimize the chance of making a false discovery
  + FDR 🡪 expected proportion of type 1 errors (false positives)
  + Controlling number of false positives 🡪 different than Bonferroni
* Calculate support, confidence and lift for Association Analysis (interpret each of these values)
  + Formulas are given on the test
  + Be able to interpret all of them (and which are the same versus different given switching the antecedent and precedent)
  + Support 🡪 measures how often we find A and B in the dataset
    - How frequently an item or rule appears in a dataset
  + Confidence 🡪 measures what percent of transactions containing A also contains B
    - How often a rules consequent is found when the antecedent occurs
  + Lift 🡪 measures how much more likely we are to buy B given that we also buy A then we are to buy B at random
    - How much more likely is the consequent to occur with the antecedent that if the two were independent
* Antecedent, Consequence
  + What is meant by a rule
* Decision trees (CART 🡪 classification and regression trees)
  + Classification tree 🡪 using node purity to define when and where to split
    - Purity is defined by genie/entropy of the node
    - Regression tree 🡪 reducing the sums of squared error
  + Terminology
  + Advantages
    - Easy to use and understand
    - Handles numerical/categorical/missing data
    - No need for scaling
    - Non-parametric (no assumptions)
    - Can be used for both classification (predicting categorical) and regression (predicting continuous)
  + Disadvantages
    - Prone to over fitting
    - Unstable and sensitive to small data changes
    - Greedy algorithm (once something is assigned it stays that way)
  + Gini, information, SSE criteria
  + Predicted probabilities, predicted classes, predicted values
    - Predicted probability 🡪 probability that a sample belongs to a certain class
    - Predicted class 🡪 a majority vote of the types of class in a node
    - Predicted value 🡪 average target value of leaf node
  + How to split categorical variables in a decision tree
  + missing values
    - handled very easily (can be kept, and is an advantage)
  + What is meant by purity of a node
  + Pruning and prepruning a tree
    - Pruning is growing tree out to complete length
      * Will overestimate
      * Pruning happens after you build out the whole tree
    - Pre-pruning is defining before the algorithm starts
      * Ex: do not split if the resulting nodes will have less than 10 observations
  + Difference between CART and conditional trees
    - Trying to reduce SSE/node purity
    - Conditional tree 🡪 basing splits based on p-values of hypothesis tests
* Clustering: Hard versus fuzzy
  + Hard: every observation given a class (we did in class)
  + Fuzzy: probabilities you are in cluster x
* Clustering: Hierarchical versus flat
  + H: done in stages. Start with n clusters, and keep combining based on distance until there is one cluster
    - Done in stages
    - N = number of observations
  + Flat: done at one time (example would be k-means)
* How does k-means work/different clusters (random seeds)
  + Randomizes where the initial clusters are
* Kmeans—converges in small number of iterations
* Advantages/disadvantages of kmeans
  + Advantages:
    - Converges quickly
  + Disadvantages
    - Looking for spherical groups
    - Also need to specify the number of centers ahead of time
    - And initial starting point for each cluster can effect the results
* Data needed for clustering
  + Clustering based on distances
* How does hierarchical clustering work
* Linkage
  + Hierarchical clustering 🡪 you must know type of linkage you are doing
* Advantages/disadvantages for hierarchical clustering
  + Advantage
    - Do not need to specify number of clusters
  + Disadvantages
    - Not good for large datasets
* DBSCAN – what it is and what is it good for
  + Density based clustering algorithm
  + Advantage
    - Do not need to specify number of clusters ahead of time
    - Can handle many types of shapes
  + Disadvantage
    - Need to give distance and something…
* What is variable clustering and what is it used for
  + Good for dimension reduction
* k-nn
  + supervised for predictions and classifications
  + can use continuous or categorical
  + need to specify k (number of neighbors)
    - voting, average, median for predicting new observation
  + cant get feature importance (as it based on distances)
* advantages/disadvantages for knn
* MDS versus PCA
  + MDS: great for viz
  + Metric MDS using Euclidean is same as PCA
  + PCA
    - Good for dimension reduction
    - Look at percent of variation explained
* GOF for MDS
  + Goodness of fit for metric MDS (I think)
    - Higher is better
  + Non-metric mds
    - Between 0 and 1 🡪 want smaller number

**Clustering**

Hierarchical versus flat

* Hierarchical clusters form a tree so you can see which clusters are more similar to each other
* Flat clusters are created by some other process, usually iteratively updating cluster assignment

K-means

* Start with k “seed points”
  + Randomly initialized
* Assign each data point to the closest seed point
* The seed point then represents a cluster of data
* Reset seed points to the centroids of the clusters
* Repeat steps 2-4 updating the cluster centroids until they do not change
  + Advantages
    - Modest time/storage requirements
    - Good for a wide variety of data
    - Can terminate after a small number of iterations and still have good results
  + Disadvantages
    - Dependent on initial cluster anchors and initial number of clusters
    - Outcome depends on the starting point of each cluster
    - Can be sensitive to outliers (should standardize the data then)
    - Difficulty detecting non-spherical clusters

Hierarchical Clustering

* each point starts as its own cluster
* calculate the distance between each point
* choose two points that are the closest and form a cluster
* calculate all distance between all single points and clusters
* find smallest distance and combine to form a cluster
* repeat until all observations are in one cluster
* will result in a dendrogram
* Linkages
  + Single linkage 🡪 distance between the closest points in the clusters
  + Complete linkage 🡪 distance between the furthest points in the cluster
  + Centroid linkage 🡪 distance between the centroids of each cluster
  + Average linkage 🡪 average distance between all points in the cluster
* Advantages
  + Creates hierarchy (dendrogram) to help examine and choose number of clusters
  + Do not need to know number of clusters ahead of time
* Disadvantages
  + Computationally expensive, and not good for large datasets
  + Only makes decisions based on local criteria
  + Merging decisions are final
  + Poor performance on noisy or high dimensional data like text
* DBSCAN 🡪 groups together points that are close to one another based on a distance measure, a minimum number of points, and a “neighborhood” about each point
  + Advantage
    - Do not need to specify number of clusters ahead of time
    - Can handle many types of shapes
  + Disadvantage
    - Need to give distance and something…
    - Dependent on epsilon for distance
    - Need to specify number of clusters ahead of time

KNN

* Defines the K closest points
* Advantages
  + Easy to explain and understand
  + Applicable to any type of data
  + No assumptions have to be made
  + Only needs a large training set
* Disadvantages
  + Computationally expensive
  + Requires storage for the training set
  + Results are dependent on the choice of distance function, combination function, and number of neighbors
  + Susceptible to noise
  + Requires lots of data pre-processing and considerations for data metrics
  + Does not produce a model
    - And does not do feature importance
* Combination functions 🡪 how do we make a prediction for a certain group
  + Numeric target
    - Mean or median of target values in the group
  + Class target
    - Majority rules
    - Create probabilities of each class as the proportion of neighbors voting for each class
    - Weighted voting: nearer neighbors have stronger votes

PCA

* Is a data reduction technique
* Is useful to reduce the numbers of dimensions
* When you preform PCA, it creates d principle component vectors (d is the number of variables that you have), and then you chose K (how many to keep)
* Each principle component is independent of each other
* Can see how variation is explained by each variable/component

MDS

* Can be used to visualize data in a lower dimension
* To preform MDS, you need to give the algorithm a dissimilarity matrix (or distance matrix)
* Classical MDS
  + Will preserve the original distance between the data points (similar to PCA)
* Non-metric MDS
  + Constructs fitted distances that are in the same rank order as the original distances
    - Can be used with qualitive and quantitative data

PCA vs MDS

* PCA is more focused on the dimensions themselves and wants to maximize the explained variance
* MDS is more focused on the relationships between the scaled objects
* MDS is preferred for visualization
* PCA is preferred to data analysis