Review sheet for Data Mining

Bootstrapping

Number of observations per variable

Dealing with transactional data

Dealing with Missing values

Bonferroni correction

FWER and FDR

Calculate support, confidence and lift for Association Analysis (interpret each of these values)

Antecedent, Consequence

Decision trees

- Terminology
- Advantages and Disadvantages
- Gini, information, SSE criteria
- Predicted probabilities, predicted classes, predicted values
- How to split categorical variables in a decision tree
- missing values
- What is meant by purity of a node
- Pruning and prepruning a tree
- Difference between CART and conditional trees

Clustering: Hard versus fuzzy

Clustering: Hierarchical versus flat

How does k-means work/different clusters (random seeds)

Kmeans—converges in small number of iterations

Advantages/disadvantages of kmeans

Data needed for clustering

How does hierarchical clustering work

Linkage

Advantages/disadvantages for hierarchical clustering

DBSCAN – what it is and what is it good for

What is variable clustering and what is it used for

k-nn

advantages/disadvantages for knn

MDS versus PCA

GOF for MDS