CPSC 392– Introduction to Data Science

Exam II Review

* Supervised Classifiers
  + k-Nearest Neighbors
  + Decision Trees
  + Naïve Bayes
* Unsupervised Clustering Algorithms
  + kMeans
  + Hierarchical Clustering
* Preprocessing Algorithms
  + Principal Component Analysis

For supervised classifiers, you need to know:

* How they work, what aspect of the data they are targeting (probabilities, distance metrics etc.)
* What kind of data they work best with
* What the underlying assumptions are for applying each algorithm
* Pros and cons of each algorithm
* Preprocessing steps that need to be taken before applying these algorithms
* Computing the accuracy of these models

For unsupervised clustering algorithms, you need to know:

* How they work, what aspect of data they are targeting
* How to measure accuracy of these models
* Elbow-method for kMeans and how to compute the optimal k value
* Dendrogram generation for Hierarchical clustering and how to use it to find the k value
* Linkage and distance methods
* Pros and cos of each algorithm

For PCA, you need to know the math behind reducing a feature space using eigen decomposition. And when should PCA be applied.

I can also give you a scenario where you are given a data set and you need to figure out which machine learning technique to apply, or comment on if it is viable to use either one given a special condition. (Similar to Exam I)