Searching for Similarity: Narrative Document

Isabelle Kirby, Bridgette Bryant, Rikita Patangay

Zuhayr Ali

# How kNN and Decision Trees Work for Classification and Regression

## Regression

### kNN

Summary

### Decision Trees

Summary

## Classification

### kNN

In classification, kNN works by finding the distances between the closest k neighbors of the data point to try to classify it based on how close it is to its neighbors (whom may be of different classes). In a way it basically groups data points together based on how close they are to each other. This makes it very easy to implement, useful for nonlinear data, and versatile. It usually has a high accuracy however, it sensitive to the scale of the data as well as irrelevant features of the data. It also requires a lot of computation and memory while being relatively slow to predict, especially with a large k value.

### Decision Trees

In classification, decision trees work by finding the best predictors which give good ‘splits. With these predictors it will create a split in the tree (a decision). Each split gives you a logical representation of which predictor value can easily classify it. It calculates each split using a greedy-recursive algorithm which split based on a top-down approach to partition and examine predictors (to see if they will make good splits). It will choose what seems best greedily and recursively continue until it hits a stopping point which can either be after a specific number of splits or until the cost of splitting isn’t worth the split for classification. Trees can often overfit the data, to avoid this you can attempt to prune the tree, which will remove some of the last few splits to make the tree a bit more generic for the test data.

# How the 3 Clustering Methods of Step 3 Work

# How PCA and LDA Work, and Why they Might be Useful Techniques for Machine Learning

## PCA

Principal Components Analysis is a data reduction technique that helps us reduce the dimensions of our datasets. PCA will manipulate the data and reduces the number of axes in a new coordinate space. In this reduced new coordinate space, each axis will represent a principal component. The first principal component (PC1) will represent the dimension of the most significant variance, and the other principal components represent decreasing variance. Since it is a data reduction technique, we will be losing data and may also lose accuracy in any models. Many times, in machine learning you will come upon high-dimensional datasets that can be hard to explore without reduction. This is when PCA is used in ML.

## LDA

LDA works by seeking to find a linear combination of the predictors that will maximize the separation of the classes while minimizing the within-class standard deviation. LDA is a supervised classification technique. Again, it is basically used to reduce the data, so we are able to see the numbers clearer and make predictions with the data.