Jered Hightower, Haniyyah Hamid, Sai Gonuguntla

Decision trees and kNN - Regression

For regression, kNN takes an average of the target values of the closest Neighbours to predict the value. For example, if we wanted to predict house prices we would take the average of the neighbours k value.

For decision trees, a top-down, greedy approach is used for partitioning the data. First , we examine the predictors to see whether they make good splits in the data. And for each predictor we have to determine the numerical value at which the split will be.

Decision trees and kNN - Classification

In terms of classification, kNN will take the target value that is to be predicted and then attempt to group it with a category or class of values of similar label by looking at the k nearest neighbours. For example, if the target value that is being predicted, X, is surrounded by training data that are labelled “Red”, then that target value will be labelled “Red” as well due to the closest k neighbours in its proximity being labelled as that. It does so by calculating the approximate conditional probability for a class j as the fraction of neighbours that are labelled such.

Classification with decision trees also uses the same top-down greedy approach used for regression. However the difference is that the counts of classes in regions would replace the RSS. Also, rather than using accuracy as a measure for splitting regions, entropy and the Gini index are utilised. Entropy would measure the uncertainty in the data while the Gini index measures homogeneity amongst regions. With these metrics, the goal is to be able to see from the decision tree how homogenous the regions are.

Clustering

How kMeans Clustering Works

kMeans uses an iterative algorithm that randomly selects k observations to centroids. It then assigns each observation to its closest centroid and recalculates the centroids. This is repeated until convergence.

How Hierarchical Clustering Works

Hierarchical clustering places each observation in its own cluster and calculates the distance between each cluster and every other cluster. The two closest clusters are combined and the two previous steps are repeated until all observations are in one cluster. This forms a dendogram.

How Model-Based Clustering Works

Model-based clustering attempts use a variety of data models in an attempt to find one that fits the data best and applies maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters.

PCA and LDA

Our accuracy with PCA and LDA was slightly lower than the results we got from logistic regression, kNN and decision trees, but not by much.

PCA is an unsupervised algorithm and is a data reduction technique to reduce the number of predictor columns. It transforms the data into a new coordinate space while reducing the number of axes and each axis of the reduced space represents a principal component. PC1 represents the dimension of greatest variance and others PC's represent decreasing variance.

LDA is a supervised algorithm and tries to find a linear combination of the predictors that maximises the separation of the classes while minimising within class standard deviation. Works better than PCA when the class is known.

They might be useful in machine learning because they can help reduce the number of features so there is a more manageable amount and make it easier to plot the data.