Report_final

Thomas Harrison

Abstract

Introduction

Using the forest fires dataset taen from UCI machine learning repositry REFERENCE HERE, a series of classification machine learning alogorithms will be undertaken in order to identify their accuracy. These classification alogirthms will be a library implementation of a decision from SKLearn REFERNECE SK LEARN, and a decision tree written. The accuracy will be further tested through testing the parameters of the tree such as maximum depth of a tree and the minimal number of splits. Testing the implication of these paramters will allow for a better understanding on how trees function.

```
library(patchwork)
library(nnet)
library("knitr")
library(ggplot2)
```

Methods

Data-preprocessing

Data pre-processing is an important step and consists of the following algorithm:

- Standardise
- Remove anomalies
- Principal component analysis First standardising the data is a stem in which brings all the individual features so that they lie within the same sacle, standardisation is a method in which scales all the data so they lie cenetered around the mean in units of the standard deviations. This is important as many machine learning algorithms use the Euclidean distance between two data points within their computation, without feature scaling the features with high magnitudes will be more weighted leading the skewed and biased results.

  Removing anomalies is another important step, if anomalies were not removed it will lead to skewed results. If there was highly influential point left within the dataset this would lead to highly biased data and negatively impact the final accuracy when performing classification alogorithms.

  Principal component analysis(pca) is a dimentionality reduction technique, it works by taking a large dataset and transforming it into a smaller that contains the information of the larger set. # Clustering In order to performing classification(supervised learning techinique), first the data needs to undergo clustering the calculate the labels for the classification to train from. The clustering algorithm choses was Gaussian mixture model. IMPORT GRAPH HERE

  From visually inspecting figure #reference here#, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are minimal at seven indicating that this is the optimal number of clusters.

Decision tree

Two different approacher were taken to compare the performace of decision trees. A library implimintation from SKLearn #rfernce here# and a decision tree written from scratch were compared and how changing the tree paramters such as max depth affected the performance and computation times. # SKLearn decision tree classifier The SKLearn implimantation has different paramters that effect the construction of the tree, altering these can help reduce any issues such as over or underfitting. This is important as this will negatively impact the prediction of the labels. Overfitting is a phenomena in which the model is designed to fit the training data perfectly but would not predict the labels accurately for the untrained data. Underfitting is a phenomena in which the decision tree cannot capture the underlying trend of the data, this will result in poor accuracy when predicting labels for both the training dataset and the un-trained data. SKLearn has the option to alter parameters in order to change the structure of the decision tree, the paramters chose to test were:

- criterion - Measure quality of split.
- max_depth - Maximum depth of the tree.
- min_sample_split - Minimum number of samples required to split.
- min_sample_leaf - Minimal number of samples required to be at a leaf node.
  Raw Tree

Results

Maximum depth characteristics
Min tree split
Min leaf Split
Test Train Split

Discussion

Bibliography