Rahul Srikonda

rahulsrikonda@u.boisestate.edu

Abstract

Generating decision tree from scratch using Python and NumPy and matplotlib  
Breakdown data set into test and training dataset  
Compare accuracy of performance between training dataset and test dataset

Implementing decision tree from scratch

CS334 – HW#6

* Data set used 🡪 diabetes data set.
* Label column has been coverted o YES/NO
* Helper functions were created to help perform below activities:
  + Check\_purity 🡪 helps check if the node is pure or impure
  + Potential\_splits 🡪helps generate splits between data points in all features – Fig(1)
  + Plotting the data doesn’t show any clear correlation that was visible to eyes – Fig(2)
  + Split\_data 🡪helps split data in features based on split\_value
  + Calculate\_entropy and calculate\_overall\_entropy 🡪 helps calculate entropy based on split
  + Determine\_best\_split 🡪 helps define best split on data. This appears to be glucose for our data set. – Fig(3)
* Above helper functions are then used to generate the Decision Tree Algorithm. Hyper parameters used are max\_depth=7 and min\_samples = 2. Fig(4)
* Classification function helps classify the output for an example file Fig(5)
* Accuracy is calculated using calculate\_accuracy, which shows the test data set has 10% lower accuracy thean training data set. Fig(6)











