Supervised Learning Algorithm Comparison and Analysis

Yi Zhao

[Yzhao644@gatech.edu](mailto:Yzhao644@gatech.edu)

Introduction

DataSet 1: Income level classifier

1. Data Description

The data is obtained from UCI Adult Income Dataset. What I like about the data set is there are decent amount of data points, about 30k data points. For feature wise, it has a mix of categorical data and numerical data. The data is clean, most of the data points have valid data points, except a few question marks, which I will explain how to impute later.

1.1 data details

1.2 impute method

1.3 historical plot

1.4 categoical data combine

1.5 one hot encoding

2. Model Training details

2.1 data pipeline

GridSearch, cross-validation, learning curve

2.2 Decision Tree

2.3 MLP

2.4 Boosting

2.4 SVM

2.4 KNN

2.5

3. Result and reasoning

4. Future work

Unbalanced label

Reference

1. UCI adult income data set http://archive.ics.uci.edu/ml/datasets/Adult