CS 4641 Project 1 Report

HU Heng

February 7, 2019

1 Overview

This report is for CS 4641 Machine Learning project 1 supervised learning. In the following pages, you will see analysis of five different learning algorithms. I will show the results by some figures which is about the performance of each algorithm on different hyperparameters. I will first introduce the datasets I used in my experiences. Then the five algorithms and the corresponding results will be described and analyzed. Finally, I will show how the size of training dataset influences the classification result.

2 Datasets

The datasets I choose include wine dataset and adult dataset. I will introduce these two datasets and explain the reason why I choose these two datasets in detail.

2.1 Adult Dataset

Adult dataset is extracted from the 1994 Census database. There are 14 attributes and the result is whether the income is greater than $50k or not. It is a binary classification problem and there are 48842 instances in total. I believe there should be enough data to train a good model using different algorithms. In addition, binary classification problem is not complicated and it would be suitable for training. The training result should be reasonably good.

2.2 Wine Dataset

Wine data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. There are 13 di-

mensions and three possible classes. There are 178 instances in total. Previously, working on adult dataset is a binary classification task and there are more than 10k training data. I would like to move from binary classification problem to multi-class classfication problem. In addition, I am not sure whether 178 instances are large enough for classification. So I would like to do classification on this dataset and see the preformance of different algorithms.

3 Adult Dataset

3.1 K Nearest Neighbor

3.2 Decision Tree

3.3 Boosting

3.4 Support Vector Machine

3.5 Neural Network

3.6 Summary

4 Wine Dataset

4.1 K Nearest Neighbor

4.2 Decision Tree

4.3 Boosting

4.4 Support Vector Machine

4.5 Neural Network

4.6 Summary

5 Influence of Size