Supervised Learning

Writeup for Assignment 01 - CS 6741

Magahet Mendiola

ABSTRACT

An analysis of two machine learning classification problems, including an evaluation of various learning algorithms and an exploration into the

1. CLASSIFICATION PROBLEMS

Two classification datasets have been chosen based on the following criteria. First, each dataset was required to include a minimum of 3,000 sample instances to insure that sufficient data was available to evaluate each learning algorithm. A standardized test set was created by taking a uniform random sample of 33% of the original instances without replacement. The remaining 66% was then used as a training superset. Smaller subsets were pulled from this training partition as needed.

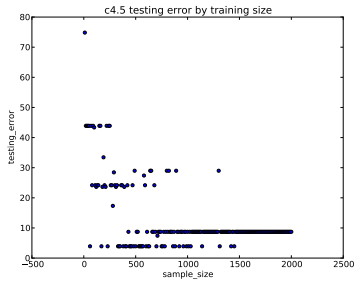
1.1 Classification Task 1 - Poison Mushrooms

The first dataset is titled, Mushroom [1] and was chosen from the UCI ML Repository. The classification task for this dataset is to determine whether a given mushroom is edible or poisonous based on the specimen's physical attributes. There are 22 recorded attributes which describe the physical appearance and olfactory perception of each sample. A full description of attributes can be found at http://archive.ics.uci.edu/ml/datasets/Mushroom

This classification task was chosen to explore the various learning algorithms' behavior on attributes with only discrete values. The attributes are based on direct physical observations, which makes the data easier to comprehend without domain knowledge.

1.2 Classification Task 2 - Income

Our second dataset was also chosen from the UCI ML Repository and is titled, Adult [1]. The classification task in this case is to determine if one's household income exeeds $50,000/yr based on 14 biographical attributes. The data was collected from a census database from 1994. Attributes include the subject's age, level of education, marital-status, occupation, race, sex, etc. The full description of attributes can be found at http://archive.ics.uci.edu/ml/datasets/Adult.



(a) Test Error by Training Size - C4.5

(b) Test Error by Training Size - ANN

2. ALGORITHM EVALUATIONS

2.1 Learning Curve

2.2 Decision Trees

2.3 Neural networks

2.4 Boosting

2.5 Support Vector Machines

2.6 k-nearest neighbor

3. ANALYSIS

3.1 Overview of Results

3.2 Algorithm Comparison

4. REFERENCES

[1] K. Bache and M. Lichman UCI Machine Learning Repository 2013 http://archive.ics.uci.edu/ml University of California, Irvine, School of Information and Computer Sciences