

Machine Learning - CS 7641 Assignment 1

Anthony Menninger

Georgia Tech OMSCS Program
tmenninger@gatech.edu

Abstract

FILL IN?

Introduction

Data Sets

The first data set is from 1994 and 1995 current population surveys conducted by the U.S. Census Bureau [1] from the UC Irvine Machine learning site. The classification task is to determine if the user type for each instance has an income above \$50,000 or below (\$94,000 in today's dollars). The 1994 median annual salary was \$16,000 (2020 median salary was \$34,600)

There are 40 usable features, with both continuous features, such as age, and discrete, such as married status or sex. The training set has 199,523 instances and the test set has 99,762 instances. Because many of the features in the file are string based, I created transformed instances using the sklearn preprocessing LabelEncoder to translate discrete string values into consistent numeric values. This was necessary for algorithms that only used numeric features. I chose this data set because it seemed relatively large and I expect somewhat noisy with a smallish number of features.

The second data set is The MNIST Database of Handwritten Digits [2]. This consists of instances of 28 X 28 greyscale images of the digits 0-9. The only transformation needed was that each of these images was flattened, creating 784 features, one for each pixel. There are 60,000 training instances and 10,000 test instances. I chose this because it is a good comparison to the first data set, with a very different type of data (images), which leads to significantly more features (784). In addition, each of the features can be thought of as related to each other, as they are all positional in a two dimensional grid, while the first data set features do not have any necessary relation between themselves ie: age is not related to sex.

Decision Trees

Decision Trees. For the decision tree, you should implement or steal a decision tree algorithm (and by "implement or steal" I mean "steal"). Be sure to use some form of pruning. You are not required to use information gain (for example, there is something called the GINI index that is sometimes

used) to split attributes, but you should describe whatever it is that you do use.

I used the DecisionTreeClassifier algorithm from the sklearn library. This has several distinct aspects. First it only accepts numeric data for features and treats all features as continuous. Because of this, a feature can appear in multiple nodes within the same path, with different thresholds. It also means this is a binary tree, with each node having only two edges. Other decision tree implementations might allow for discreet features, meaning a feature could only appear once in any tree path and the tree might not be binary.

The

References

- 1 Census Income (KDD). url: <https://archive-beta.ics.uci.edu/ml/datasets/census+income+kdd>.
- 2 The MNIST Database of Handwritten Digits. url: <http://yann.lecun.com/exdb/mnist/>.