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. By default, this uses a GINI index to split the data. It was possible to use an information gain entropy for splitting, but this did not make a meaningful difference in results for this dataset. A key aspect of sklearn is that 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.

To prune the tree, sklearn uses a minimal cost-complexity setting. Each node in a tree has an α value which measures the node misclassifications minus leaves misclassifications divided by the number of leaf nodes. The DecisionTreeClassifier then takes an α minimum setting, pruning all subtrees with a lower α. For both datasets, an accuracy curve was created by fitting the tree with different α values.

The Census Data, shown in Figure 1, mostly shows the expected curve. Accuracy improves as nodes are added, to a point, then accuracy declines due to over fitting as more nodes are added. Alpha, the pruning constant, leaves more nodes in the tree as it is decreased. The Census Data also showed a unique characteristic. One of its features (Capital Losses), showed a very high correlation with the labeled output, Income above $50k. Thus, a one node tree already showed 94% accuracy. Figure 1 shows this, with the the left side of the chart with few nodes already having a very high accuracy. Overall, the highest accuracy achieved was 95.3%.

The MNIST image data, shown in Figure 2, showed a more normal accuracy curve, with accuracy improving dramatically with added nodes until a maximum accuracy was reached, then accuracy declines due to over fitting. The maximum accuracy achieved was 88.7%, which was significantly below the census data. A decision tree is looking at one pixel at a time, so small variations in position or size would be very challenging for this algorithm. I think other classifiers, especially the neural network to better handle image data.

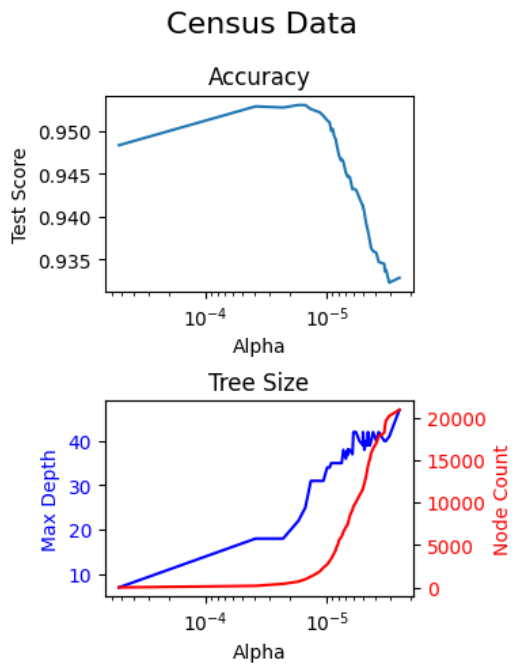

Figure 1: Census Data Decision Tree: The right side of the upper chart shows overfitting as the pruning is reduced and the number of nodes and tree depth increases. An unusual aspect is the left side of the chart shows very high accuracy with very few nodes. This stems from one feature being highly correlated to the labeled output.

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/.

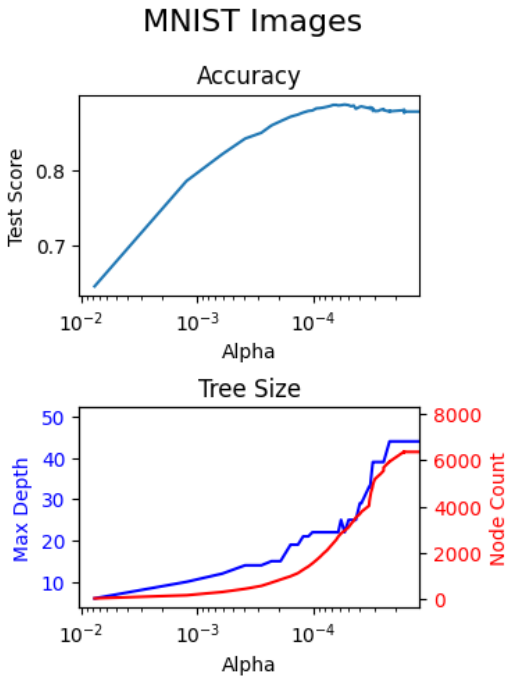

Figure 2: MNIST Images Decision Tree: The left side of the chart shows rapidly growing accuracy as nodes are added, eventually reaching a peak accuracy. The right side then shows overfitting with declining accuracy as less pruning increases the number of nodes and tree depth.