CSCI 446 Artificial Intelligence Project 2 Final Report

Roy Smart Nevin Leh Brian Marsh

November 21, 2016

Abstract

1 Introduction

In a broad sense, machine learning refers to efforts to give computers the ability to learn without explicit instructions and encompasses a wide range of problems. Classification in the realm of machine learning is a problem that can be solved using several algorithms, but the effectiveness of the algorithm depends greatly upon the specific dataset involved. We implemented four algorithms to attempt to classify new data points into sets based upon previous information gained through âĂIJtraining.âĂİ The algorithms used are k-Nearest Neighbors, NaÃŕve Bayes, Tree Augmented NaÃŕve Bayes (TAN), and Iterative Dichotomiser 3 (ID3). Additionally, each algorithm is tested with five different datasets: the Wisconsin Breast Cancer Database, the Glass Identification Database, the Iris Plants Database, the Small Soybean Database, and the 1984 United States Congressional Voting Records Database. The effectiveness of each algorithm is measured by the metrics of precision, recall, time-complexity, and convergence.

2 Datasets

2.1 Discretization

2.2 Stratified Cross-Validation

2.3 Missing Values

3 k-Nearest Neighbors

3.1 Training

K-Nearest Neighbors attempts to solve the classification problem by using already-known data points with similar characteristics to make an educated guess of the class. The training for k-Nearest Neighbors involves simply reading in all of the dataset and storing these records as points based on their values.

3.1.1 Constructing Probability Table

3.2 Validation

3.2.1 Value Distance Metric

3.2.2 Determining k and p

4 Naive Bayes

4.1 Training

4.1.1 Constructing Probability Table

4.2 Validation

4.2.1 Determining Class Probability Distribution

5 TAN

5.1 Training

5.1.1 Constructing Probability Table

5.1.2 Constructing Augmented Tree

5.2 Validation

5.2.1 Determining Class Probability Distribution

6 ID3

6.1 Training

6.1.1 Tree Construction

The `id3` method was the main logic behind tree construction. We closely followed the decision tree learning psuedocode in Figure 18.5 in [Russel and Norvig, 2010]. ID3 is an recursive process that generates a short tree by splitting the tree one attribute at a time. The first attribute to be used is the one with the most information gain. Information gain can described as how much entropy the system lost by splitting on an attribute. The equation for entropy is as follows:

$$H(S) = \sum_{i=1}^{n} \left(-p_i \log_2 p_i\right).$$

where p_i is the proportion of the number of datums in class i to the total number of datums in the set S. To use this equation the entropy of the whole system is determined first. Then the sum of the entropy for each new branch is calculated and subtracted from the total entropy. We try to branch on each attribute that has not been used yet. Since the goal is to maximize this difference so the tree will be small and more general, we choose the attribute that resulted in the greatest reduction of entropy. The reduction of entropy is defined as information gain. The rest of the algorithm is just iterating through the tree as it is made and recursively perform the calculation on an attribute that has not been split yet. The tree is finished when all leaf nodes are a single class or when attributes to split on have run out. In this case each leaf node that does not have a class is assigned the most common class in that leaf.

6.1.2 Pruning

We implemented reduced error pruning to help make the tree more general. The approach was to split the training set into a smaller training set and a validation set. The tree was then generated as usual using the new training set.

6.2 Validation

7 Results

7.1 Algorithm Convergence

7.2 Algorithm Precision

8 Conclusion

References

[Russel and Norvig, 2010] Russel, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson Education, Upper Saddle River, New Jersey 07458, 3rd edition.