

# CSCI 446 Artificial Intelligence Project 2 Final Report

ROY SMART

NEVIN LEH

BRIAN MARSH

November 21, 2016

## Abstract

## 1 INTRODUCTION

Define MLA acronym here!!!!

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 training. The algorithms used are k-Nearest Neighbors, Naïve Bayes, Tree Augmented Naïve 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 and convergence.

## 2 DATASETS

In this project we are asked to train our MLAs on 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. Each dataset has a different set of challenges to our MLAs, for example, the soybean dataset is really small, the cancer database is reasonably large, and the attributes in the glass dataset are not very well correlated (as we will soon see).

### 2.1 Discretization

Two datasets, the iris dataset and the glass dataset have continuous values. We are asked to discretize these dataset to more accurately compare the different MLAs. For our project, we adopted *binning* as our discretization scheme. Binning is a simple method for discretization that defines a certain number of adjacent intervals, known as bins, and then replaces points that fall within each bin by a value representative of that bin.

For the data in this report the number of bins was set to 10. In principle, the ideal number of bins in terms of MLA precision could be any integer larger than one, but through tests we found that the ideal number of bins is proportional to the number of classes in the dataset.

### 2.2 Cross-Validation

Cross-validation is statistical bootstrapping technique used to estimate the generality of a statistical model using a dataset independent of the dataset used to create the model. In the context of machine learning, this means that we will partition the full datasets into training and validation datasets. For this project we are

asked to use the so-called  $k$ -fold cross validation technique, where the full dataset is partitioned into training and testing datasets through the use of *folds*. The folds are found by splitting the dataset into  $k$  equal sized pieces, and assigning each to be a fold. One fold is selected as the validation set, and the rest of the folds are assigned to the training set. This process is then repeated  $k$  times, where each fold is taken to be the validation set once.

To further increase the accuracy of our cross-validation process, we will *stratify* the folds described by  $k$ -fold cross-validation technique. Stratification in this context means that each fold will have an equal distribution of the classes in each dataset.. This process makes cross-validation more accurate, because it gives each MLA adequate information on each class for every training dataset. Without stratification, some training datasets could give a MLA zero examples of a particular class, hurting its ability to generalize.

In  $k$ -fold cross-validation,  $k$  is a free parameter that can be tuned by the user. For this project, we have selected  $k = 10$ , because it will allow us to perform the convergence tests discussed in Section 7.1 over a larger range.

## 2.3 Missing Values

Some of the datasets have missing values. This is problematic since most of the MLA's will yield vastly different answers if some data is missing. However, we note that the appearance of missing values in the voting dataset is merely an illusion, missing values in this dataset can signify a stance on a particular issue and this information should be used by our algorithms.

Since the voting set contains no missing values, the only dataset with missing values is the cancer dataset. Since there are so few missing values and since the cancer dataset is so large, we have decided to simply delete any data with missing values from the cancer dataset. Therefore, we have found that data imputation is not necessary for this project.

# 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

During the training phase for naive Bayes, the algorithm,

#### 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 pseudocode 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 be described as how much entropy the system lost by splitting on an attribute. The equation for entropy is as follows:

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

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 performing 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.

To deal with continuous variables we just used the discretization method outlined above.

#### 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. We then removed non-leaf nodes one at a time, checking how well they performed on the validation set. If they performed better, the nodes remained removed. Otherwise they would be restored to their previous state.

	<b>Cancer</b>	<b>Glass</b>	<b>Iris</b>	<b>Soybean</b>	<b>Vote</b>
<b>kNN</b>	0.953154	0.579221	0.96	1	0.942389
<b>NB</b>	0.975085	0.649783	0.95333	1	0.903436
<b>TAN</b>	0.961893	0.66342	0.96	0.98	0.944767
<b>ID3</b>	0.950171	0.662771	0.98	0.975	0.95629

**Table 1:** *1x10 Cross Validation Precision Average*

## 6.2 Validation

The process of classifying a datum according to the tree is a simple process. The tree is traversed according to the values of each variable in the datum. When a leaf node is reached, that leaf's class is returned. It is easy to determine leaf nodes because they lack children nodes.

## 7 RESULTS

### 7.1 Algorithm Convergence

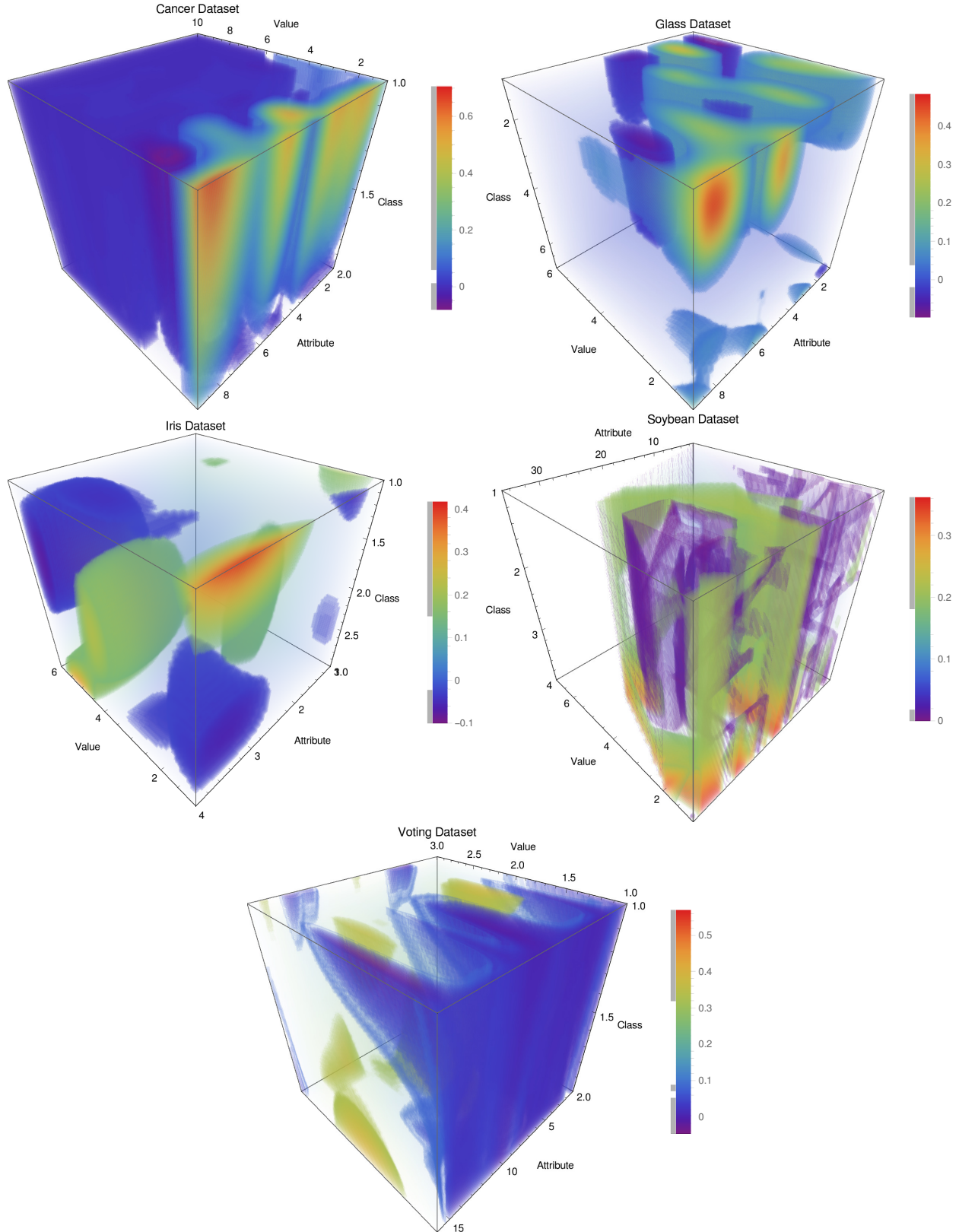
To show the relationship between training set size and precision we created the graphs seen in Figure 3.

### 7.2 Algorithm Precision

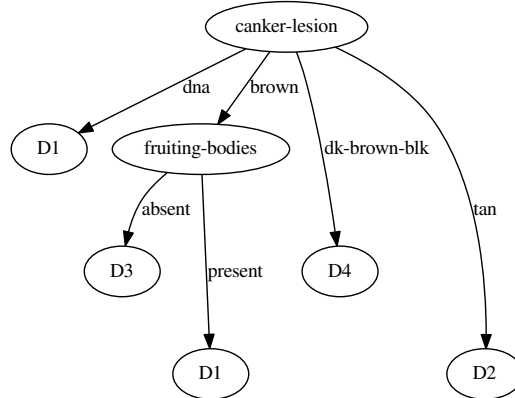
## 8 CONCLUSION

## REFERENCES

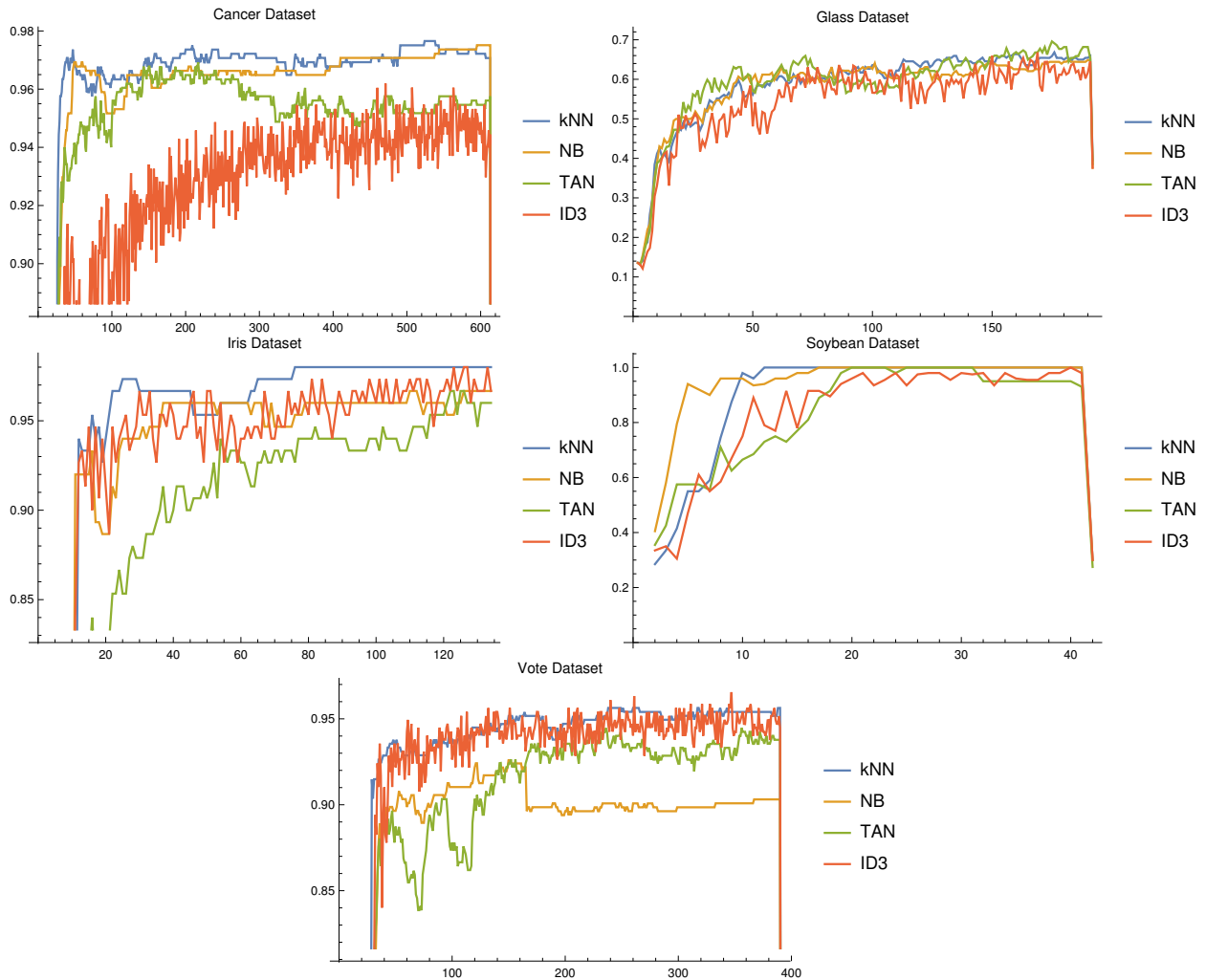
- [Russel and Norvig, 2010] Russel, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson Education, Upper Saddle River, New Jersey 07458, 3rd edition.



**Figure 1:** 3D density plots of the naive Bayes probability table for all five datasets. The vertical class axis represents the value of the class,  $c$ , the attribute axis denotes the index of an attribute,  $x$ , and the value axis represents the value of that attribute,  $a$ . Finally the color is equivalent to the probability of the attribute  $x$  having value  $a$ , and the class having the value  $c$



**Figure 2:** An example id3 tree generated for the soybean database.



**Figure 3**