CSCI 446 Artificial Intelligence Project 3 Design Report

Roy Smart Nevin Leh Brian Marsh

November 3, 2016

1 Introduction

Introduce Machine Learning Algorithm (MLA)!!!!

2 Datasets

2.1 Dataset Representation

We will define a *datum* to be a vector consisting of the class as the zeroth element and the associated attributes as the rest of the elements. Classes and attributes will be represented by an integer. Continuous data will be binned before it is inserted into each datum. The resolution of the bins will be a variable that will have to be tuned. Each dataset will be represented as a vector of datums.

2.2 Data Imputation

Imputation is the process of approximating missing values in the datasets. To our knowledge there is only one dataset that has real missing values: the Wisconsin Breast Cancer Database. The 1984 United States Congressional Voting Records Database appears to have missing values, but these can actually be interpreted as a stance on a particular issue. Since the breast cancer database has a small proportion of missing values, it is appropriate to simply eliminate datums with missing values. The authors assert that trying to train a MLA with imputed values would only create unnecessary bias in the network.

However, it is a common real-world problem to attempt to classify an unknown, incomplete datum. Therefore we will perform imputation on the validation datasets. To approximate the missing values we will first try a hot-deck imputation, where missing attribute values of a given datum are constructed by selecting a random member of that datum's class and copying the value of the attribute. If the hot-deck is unsuccessful, we will attempt to fill in the missing values using a regression model developed in *Mathematica*.

2.3 Cross-validation

To partition the full datasets into test and training datasets, we will use 10-fold cross validation. This method partitions the data into ten *folds* and uses one fold for testing data and the remaining nine folds for the training dataset. This process is repeated nine more times until every fold has been used as a test dataset.