CSCI 447 Project 2 Design Document

George Engle, Troy Oster, Dana Parker, Henry Soule

September 25, 2019

1 Introduction

For this assignment, we are required to implement five different algorithms: k-nearest neighbors, edited k-nearest neighbors, condensed k-nearest neighbors, k-means, and k-medoids. We are required to implement all of these algorithms on six different datasets, three of which are classification datasets, and three of which are regression datasets. We are then required to asses the performance of each algorithm using 10-fold cross validation and loss functions of our choosing. The essential purpose of the latter four algorithms is to produce a reduced data set with which to run k-nearest neighbors. Because we are working with reduced data on all algorithms except k-nearest neighbors, we hypothesize that the average performance of k-nearest neighbors across all six datasets without reduction will be worse than the average performance across the six datasets once reduction has been performed via edited k-nearest neighbors, condensed k-nearest neighbors, k-means, and k-medoids.

2 Experimental Design

Components

Our design has 5 primary components. *database.py* is a wrapper class that will handle all functionality of each dataset. Each instance of the database class will store the processed data of one our six datasets, the the index of the class attribute of its respective database. *knn.py* will store our implementations of k-nearest neighbors, edited k-nearest neighbors, and condensed k-nearest neighbors. *clusters.py* will store the implementation of both clustering algorithms– k-means and k-medoids. *validation.py* will perform our 10-fold cross validation and our loss functions. *main.py* will perform the execution of our program.

Design Decisions

Each of the five algorithms we are implementing compute distance between datapoints in each dataset. We have chosen to use euclidean distance for each algorithm. We will be using 0-1 Loss to compute performance of each algorithm

on each classification dataset. We will use mean squared error to compute performance of each algorithm on each regression dataset.

3 Plan

To test our hypothesis, we need to compute the average performance of k-nearest neighbors across all six of our datasets, and the average performance of each of the other four algorithms across all six datasets. Our program will first perform k-nearest neighbors on all six our datasets, performing 10-fold cross validation for each dataset. As we perform our cross-validation for k-nearest neighbors on each dataset, we will compute the average loss function (0-1 for classification data, mean squared error for regression) for the current dataset. We will then compute the average overall loss function across all six datasets once we have completed our cross validation for each dataset.

We will then perform 10-fold cross validation using the four other algorithms. Each of these algorithms produces a reduced dataset with which to work. Thus, we will implement each of these algorithms to output the reduced dataset. For each of the reduced datasets produced by each of the four algorithms, we will perform 10-fold cross validation