**CS7641 - Machine Learning**

**Project 1 – Supervised Learning**

**I. Introduction**

The motivation of this paper is to survey various supervised machine learning algorithms. These algorithms are: decision trees, neural networks, boosting, support vector machines, and k-nearest neighbors. The algorithms are to be applied to classification problems. We will be exploring the performance of these algorithms by looking into training vs. testing error on two different datasets.

The first data set records the physiochemical characteristics of red wine and a final dependent variable that is the quality rating of the wine. This dataset contains information on the wines’ various types of acidities, amount of citric acid, residual sugar, chlorides, free and total sulfur oxide, density, pH (potential of hydrogen – a numeric scale used to specify the acidity of basicity of an aqueous solution), amount of sulphates, alcohol content, and it’s quality as a score between 0 and 10.

The second data set is the result of shape feature extractors encoding the characteristics of 3D objects, in this case cars, within 2D images into 2D silhouettes. These characteristics are then represented as various features calculated based on measurements such as average perimeter, average radius, average distance from border, various ratios between maximum and minimum radii, perpendicular and horizontal lengths, major and minor axes, kurtosis about minor axis, kurtosis about major maxis, ratio between area of hollows and area of the bounding polygon, and its classification as being manufactured by being an opel, a saab (both car manufacturers), a bus, or a van.

In the following section, these two datasets will be described more in depth. It is to be noted that these data sets were not kept as they were and were, in fact, “massaged” to be more easily “digestible” by the algorithms. The manner in which they were massaged will be further described in the following section.

**II. The Datasets**

The first dataset I will talk about is the red wine data set. I found this dataset interesting, first of all, because wine is a product, and the methodology behind examining this data could be extended to any product available for mass consumption. Second of all, because there’s a mixture between quantitative, measurable, features as well as a final qualitative one., the wine’s rating. This, in my opinion, lends itself to many kids of real world problems such as deciding whether or not a stock is good to invest in based on actual measurable features, such as its alpha, volatility, 30 day return, historical mean, etc.

Since the assignment was concerned with classification problems, and the original wine data set had a quality rating represented within the domain of integers, changes had to be made. I decided to, first of all, normalize all features to be in the [0, 1] range, and second of all, quality would be encoded into two classes: POOR and GOOD. Anything under a 0.5 would be considered a wine of POOR quality, while anything above is considered to be GOOD. This might be an over-simplification of the quality rating system, but since the second dataset naturally lent itself to multi-class (4 different classes, to be precise) problems, I really wanted to have a problem that was simply a binary classification problem since these are the ones that are easiest for beginner students of machine learning to “wrap their head around”.

The second dataset was the vehicle silhouettes dataset. As I already mentioned, this was a multi-class dataset possessing 4 different classes. All other features were numeric in nature and were also normalized, same as with the previous dataset. I found this data set interesting because of its obvious relationship to problems in the fields of robotics, and also general computer vision. These applications are very relevant today, and I was interested in the way they encoded 3D objects within images into numeric features based on measurements representative of the actual object’s physical appearance rather than trying to encode it based on color, or other features more strongly related to the images rather than the physical object itself.

The end result of the vehicles dataset is simply a set of ratios of different dimensions of the shape of the vehicle and its final classification as a saab, opel, van, or a bus. I believe, since image recognition and deep learning are such huge topics, this would be an interesting introductory, yet relevant problem to examine.

The way the numeric values were normalized in both data sets was the same for both. It was a basic normalization consisting of:

for all , whereis the value of a particular feature (so each “row” is a vector of x’s), is the smallest value present in the entire dataset for that particular feature (the smallest value found for that entire column), and is the largest one.

In the next section, I will detail the performance of the five algorithms when being trained, then tested on these data sets. We will look at the confusion matrices, as well as training vs. testing errors and talk about possible explanations for the results observed.

**III. The Algorithms**

**1. Decision Trees**