An Emprical Comparison of Supervised Machine Learning Algorithms

Andres Christian Baez

June 14th, 2019

Abstract

The increasing number of supervised machine learning algorithms in recent years necessitates a thorough comparison between such approaches. As inspired by Caruana and Niculescu-Mizil, we will be comparing K Nearest Neighbors (KNN), Support Vector Machines (SVMs), and Decision Trees on the basis of testing accuracy. In addition, we will be examining the effect of various training/testing partitions on the overall performance of said models.

1 Introduction

In the present day, technology relies increasingly on supervised learning algorithms to complete a wide range of tasks in many application areas. As such, it is important for us to draw comparisons between such algorithms in a wide variety of problem settings across a range of criteria to develop a better understanding of each algorithm's strengths and weaknesses. Here we will discuss the overall performance of three such learning models in three different binary classification tasks. We will observe each algorithm's performance across multiple trials as well as multiple data partitions to create a more thorough analysis.

2 Datasets and Problem Description

2.1 Breast Cancer Dataset

The Breast Cancer dataset contains 10 features pertaining to cell nuclei as well as a label for whether or not it is malignant or benign. We define Malignant to be the positive class, indicated with a 1. There are 639 samples in this dataset.

2.2 Heart Disease Dataset

The Heart Disease Dataset contains 13 features pertaining to each patient's cardiovascular health and overall health habits. In addition each entry has a label of either 0,1,2,3, or 4. A value of 0 indicates the absence of heart disease while a value of 1,2,3, or 4 indicates the presence of heart disease as well as the severity of said disease. For the purpose of this study, we have grouped all labels 1,2,3,4 to be the positive class (indicated with a 1) and all labels 0 to be the negative class. There are 297 samples in this dataset.

2.3 Wine Quality Dataset

The Wine Quality Dataset contains 11 features and 1,599 samples. Wine quality is ranked on a scale of 1-10, but for the purpose of binary classification we consider all value 1-5 to be the negative class and all values 6-10 to be the positive class.

3 Methodology

3.1 Learning Algorithms

SVM (RBF Kernel)

For our SVM we use a radial basis function (rbf) kernel, and perform a grid search over the following C and Gamma Lists: C = 0.1, 1, 10, 100 and G = 1e-7, 1e-6, 1e-5, 1e-4.

Decision Tree

Our decision tree uses Entropy criterion and performs grid search over a list of depth in order to find the optimal max-depth parameter for a given classification task. We use the following depths in our search: D = 1, 2, 3, 4, 5.

K Nearest Neighbors (KNN)

In our observations we use a uniform weight setting where all points in the neighborhood are weighted equally. In addition, we use Euclidean distance as our distance metric. We perform a grid search over the following list of K's to find the optimal hyperparameter for each classification: K = 1, 2, 3, 4, 5, 6.

3.2 Data Partitions

We use the following three different partitions in our comparison: 0.2/0.8, 0.5/0.5, and 0.8/0.2. All dataset are relatively small and thus can be partitioned using the proportions listed above without any special accommodations.

3.3 Average Over Trials

We conduct 3 trials for each partition, dataset, and learning algorithm. We then average testing accuracy across all trials to account for the possibility of observing inaccurate results. Each dataset is shuffled with each trial, and each accuracy is computing using optimal hyperparameters obtained by 3-fold cross validation.

4 Conclusion

4.1 Performance Across Partitions

Algorithm Accuracy Across Partitions			
Algorithm	80/20	50/50	20/80
SVM (RBF Kernel)	.66	.71	.63
Decision Tree	.83	.80	.76
KNN	.65	.65	.61

Figure 1: The average testing accuracy for each learning algorithm across 3 datasets, 3 partitions, and 3 trials.

4.2 Performance Across Datasets

Algorithm Accuracy Across Datasets			
Algorithm	Breast Cancer	Heart Disease	Wine Quality
SVM (RBF Kernel)	.637, .666, .631	.638, .738, .568	.710, .731, .704
Decision Tree	.956, .947, .942	.788, .751, .661	.730, .723, .690
KNN	.637, .655, .605	.594, .619, .617	.717, .676, .612

Figure 2: Training, validation, and training accuracy for each learning algorithm across datasets
.

As observed in Figure 1 our findings indicate that in most cases a larger training to test data ratio results in a more accurate classification. We can also observe this in figure 2 as well where our accuracy measurements generally decrease as the size of our training dataset decreases. While there are some exceptions, it is evident in our study that a partition with substantially more testing data than training data will not yield optimal testing accuracy. In addition, we see our Decision Tree Classifier repeatedly outperforms both our SVM and K Nearest Neighbors classifiers. In addition, all of our classifiers were significantly more accurate on the wine quality dataset, which was substantially larger than the previous two.

References

Blake, C., Merz, C. (1998). UCI repository of machine learning databases.

Caruana, R., Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proc. 23rd International Conference on machine Learning (ICML'06).