Homework 1:
Supervised Learning

Datasets and why are they important?

Nursery

The first dataset I have chosen comes from 1980's Ljubljana, Slovenia. During that period, there used to be an excessive number of applications to nursery schools but not enough school capacity to admit all of the applicants. Therefore, the government discretized some certain aspects of the applicants by assigning points and discrete values to them. In the end, there were a total of three main categories that the applications were evaluated: parents' occupations, family structure and financial standing, and social and health picture of the family. In this dataset, these three subcategories are described with 8 different features (attributes) and these attributes have around 3-5 options each. The target value (label) is also categorized in five different categories, ranging from '*not recommended*' to '*special priority*'. In this dataset, there are 12960 instances, meaning 12960 children who are classified. I separated my training and test data as 80% and 20% respectively of the total number of instances.

The reason I have chosen this dataset is because it comes from real life and the end result (labels) was something that had a great impact on thousands of families and their children. In reality, government formulized a way to classify the children and in this project, I will be trying to capture their classification methods and models. Throughout this paper, I will be referring this dataset as the 'nursery dataset.'

Letter Recognition (Letters)

This dataset contains 20,000 instances of digitally typed and distorted in different ways letters from 20 different fonts of the English alphabet, meaning the result (label) has 26 different available options. Each letter appears between 734-813 times in the dataset, suggesting a somewhat balanced distribution. Each instance is defined by 16 different features including height, width of the char box, the number of pixels available and other features that describe edges of the letters. Each attribute is discredited from a continuous scale to an integer range between 1-15.

The reason I have chosen this dataset was to be able to compare and contrast with the first dataset. Since this dataset contains attributes that are derived from a continuous range, intuitively, the classification process seems to be different compared to the first one, a dataset with very discreet and polarized features. Another reason I have chosen this dataset was to be able to work with one of the most popular concepts in the industry: OCR (optical character recognition). I wanted to see how different machine learning algorithms would be able to capture the basics of this method and be successful when it comes to recognizing shapes and classifying them as the correct letters. Throughout this paper, I will be referring this dataset as the 'letter dataset.'

Project Overview

This homework will have analyses of 5 different supervised learning algorithms. I will be exploring how different variables and kernels affect the results of these algorithms and how do they perform when ran on the two datasets I have. I separated each dataset as 80% training data, and 20% testing data to be able to test the trained classifiers on the test data. In the analyses part, there will be learning curve graphs and varying characteristic variables versus the accuracy graphs. Most of the classifier models are cross validated 7 times and this will be later explored.

Decision Trees

Nursery

Letter Recognition

Neural Nets

Nursery

Letter Recognition

Boosting

Nursery

Letter Recognition

Support Vector Machine

Nursery

Letter Recognition

K-Nearest Neighbors

K-nearest neighbors is a classification algorithm that tries to classify the instance by finding the closest k-number of neighbors and identifying that the instance belongs to that cluster. The first step I have taken was to find the optimum k value for each dataset and then I plotted learning curves to see if changing the training set size has a significant effect on the accuracy.

Nursery

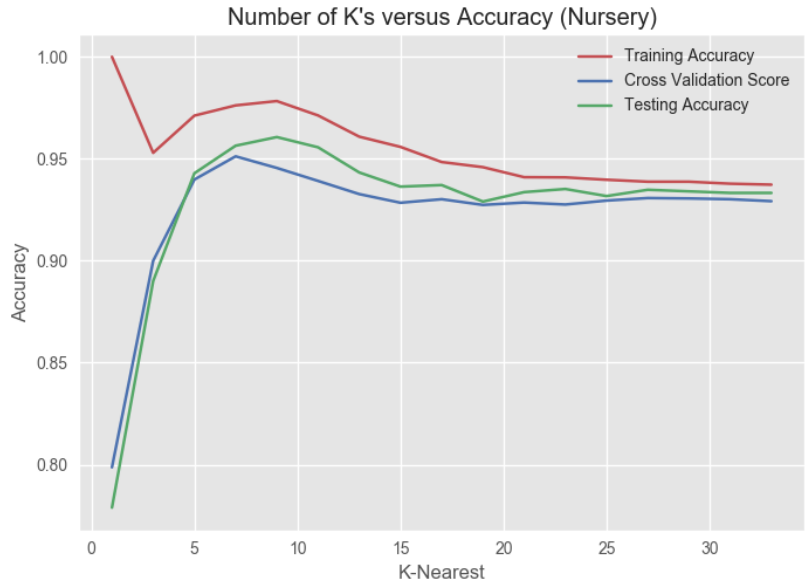


Figure: The model complexity graph for KNN algorithm on nursery dataset for k values.

Running the KNN algorithm on the nursery data set using a range of odd k values between 1 and 35, I obtained the model complexity graph given in figure. Looking at the graph, it can be seen that on lower k values, the trained model performs really well on the trained data, scoring close to 100% and over 95%. On the other hand, at these k values, the testing accuracy is low, suggesting the model performs much better on the seen data compared to unseen data. This is a typical example of overfitting since the model tailors its classifier very specifically for the trained data. The plot also reveals that the optimum k value for the nursery dataset is 9, and after this point, overfitting rate decreases and all three accuracy values start to converge but decrease, suggesting that increasing the k value all the way doesn't increase the accuracy. In the end, looking at a limited amount of neighbors only make sense and as k increases, the algorithm starts to merge different clusters into each other, decreasing the accuracy and causing false positives.

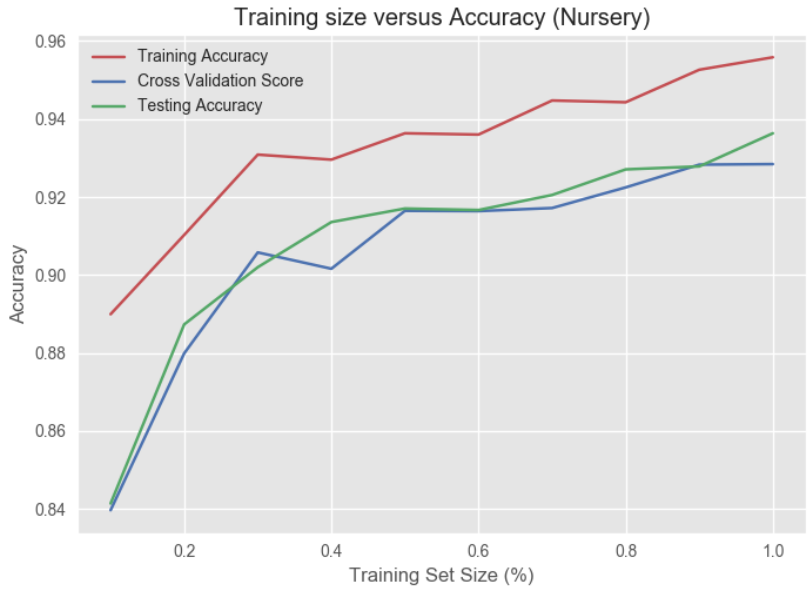


Figure: Learning curve of KNN on the nursery dataset

Looking at the figure, it can be observed that as the training set size increases, the accuracy of the model increases as well. However, it is possible to observe that the training accuracy is always higher than the testing accuracy, showing that the algorithm performs better on the seen data. A viable reason for this occasion is the KNN algorithm uses all the given training data rather than using the data to make an estimator, therefore, the entire training data is stored to run this algorithm, which leads to overfitting. Also, having very similar testing and cross validation curves show that there is not much bias in the testing set. In the final analysis, we see that, given the entire training set, the model performs with around 93.5% accuracy on the testing set, suggesting that with 9-nearest neighbors, the labels can be classified and clustered with a significant accuracy.

Letter Recognition

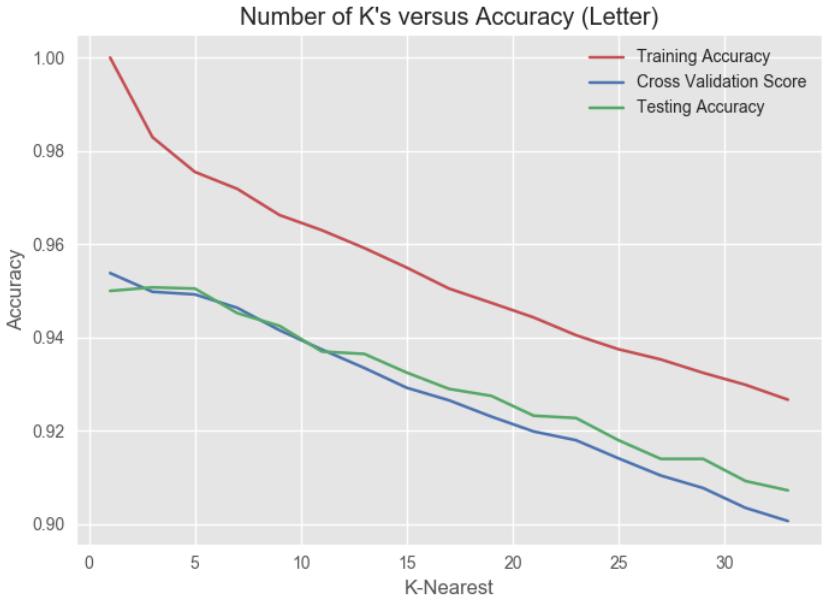
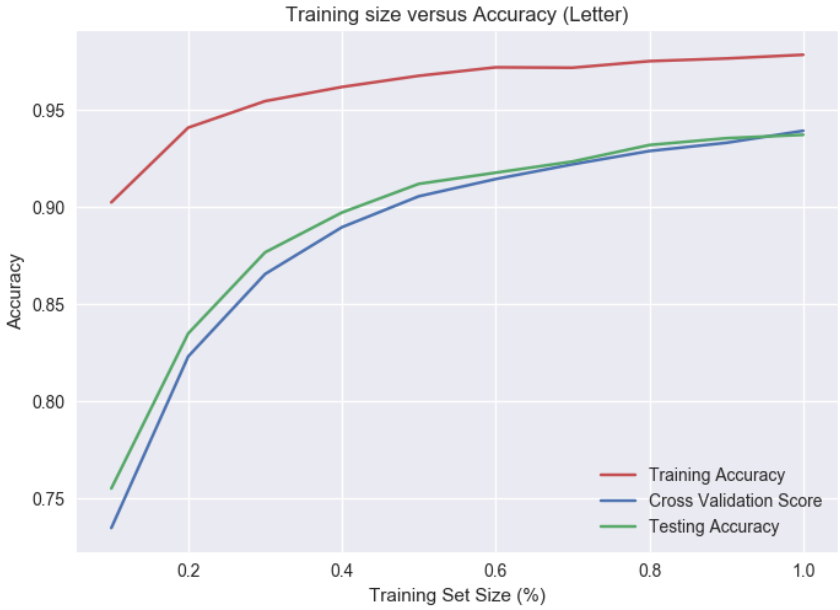


Figure: The model complexity graph for KNN algorithm on the letter dataset for k's

The figure shows the accuracy for KNN algorithm on the letter dataset, iterating over different k values ranging from 1 to 35. Unlike the previous dataset, the k value starts decreasing immediately, suggesting the optimum k value is around 3.


Training size versus Accuracy (Letter)
Accuracy
0.95
0.90
0.85
0.80
0.75
0.2
0.4
0.6
0.8
1.0
Training Set Size (%)
Training Accuracy
Cross Validation Score
Testing Accuracy