Konteya Joshi
kjoshi78@gatech.edu

Abstract

This report analyzes the two datasets using supervised learning techniques. The techniques used are KNNClassifier, SVM Classifier and Neural Networks. The datasets used are Dry Bean [1] and Wine Quality [2]. The classifiers used are binary classifiers for Wine Quality Dataset and multiclass classifier for Dry Bean dataset.

Why Are The Problems Interesting?

[3]

Dataset1

Dataset = 13,611; Type = 7;Feature = 16 ; 12;target = 4 shapes Multiclass Classification

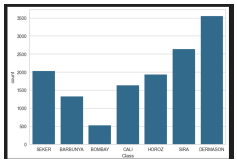

Figure 1: Target - Dry Bean Dataset

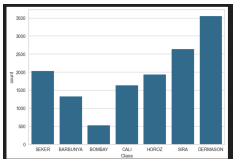

Figure 2: Target - Quality Dataset

Figure 3: Main Caption

Dataset2

Data size = 1599 red;Dataset = 4898; Bad = 3 to 6; Good : 6-9 Binary Classification

Methodology

[3] [3] Data is split into train and test datasets with .95 and .05 split. The testdatasets is preserved for the testing purposes. The training datasets was further split into training and validation datasets using .80 and .20 split. The train, validation data is used for training and validation of the algorithm

KNN Classifier

Classifier implementing the k-nearest neighbors vote.

Dataset1

Learning Curve

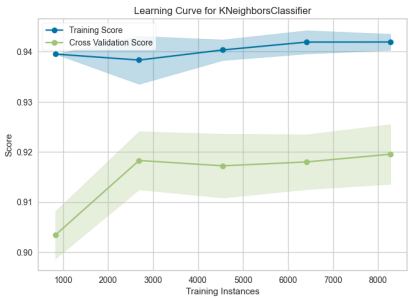

Figure 4: Learning Curve - Yellow-brick

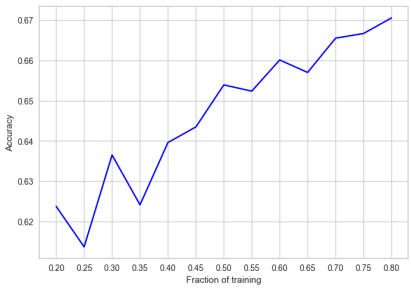

Figure 5: Learning Curve - Manual

Figure 6: Main Caption

- The learning curve Fig 7, 8 show that accuracy score increase with the increase in the size of the data.

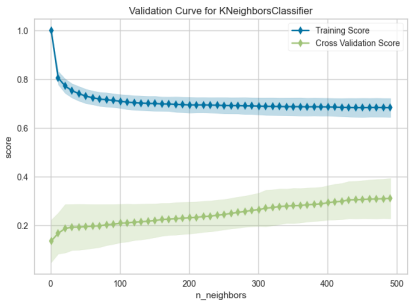

Figure 7: Caption for Figure 1

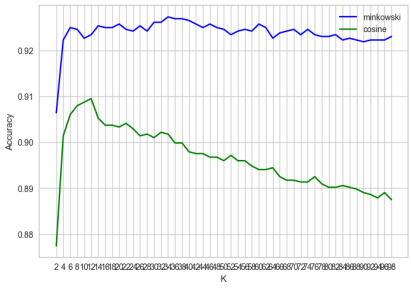

Figure 8: Caption for Figure 2

Figure 9: Main Caption

Validation Curve

- The training and the cross validation score has converges for a k value of around a dozen 12.
- Comparison of the different distance metrics for KNN with minkowski and cosine distance. The minkowski performs much better on the accuracy scores than the cosine.

Dataset2

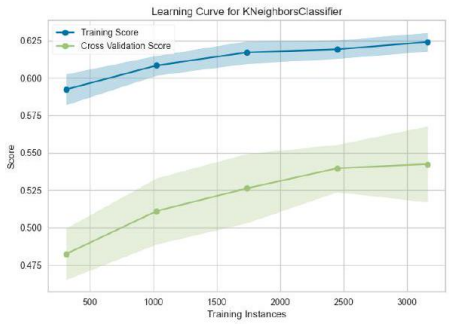

Figure 10: Learning Curver

- It is increasing but very little. but not that significant. The reason why we are not seeing the improvememnt here because our neighbourhood is very small 11 . so increase the in this data in thousands is not affecting our neighbourhood of 11

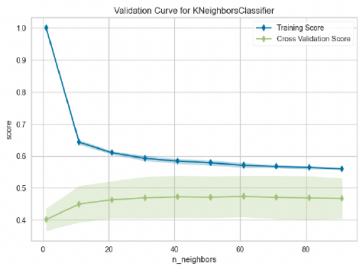

Figure 11: Validation Curve

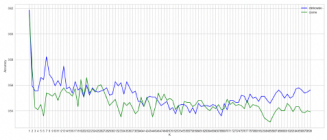

Figure 12: cosine and minkowski

Figure 13: Main Caption

Validation Curve

- The training and the cross validation score has converges for a k value of around a dozen 11.
- Comparison of the different distance metrics for KNN with minkowski and cosine distance. The minkowski performs much better on the accuracy scores than the cosine.

Neural Network

Dataset1

Learning Curve

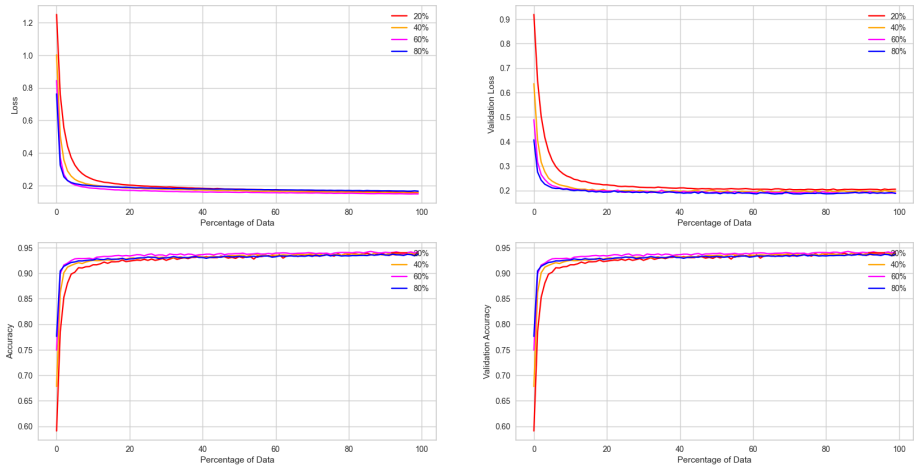

Figure 14: Learning Curve

- The training from 20, 40, 60 and 80 percentage of the training data shows that the accuracy increases with the increase in the size of the dataset.

Validation Curve

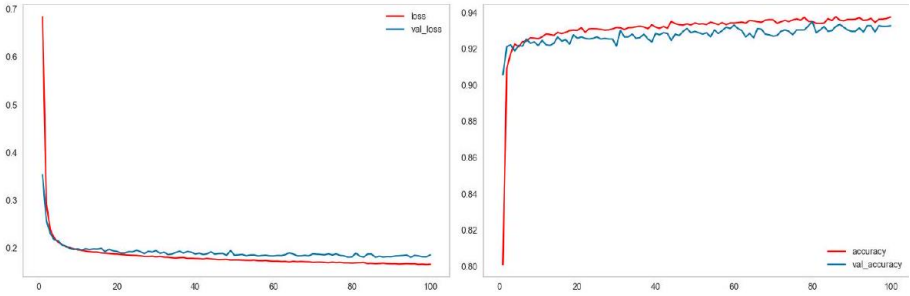

Figure 15: Dataset 1 - 1 layers(hidden neurons = 32

- The training was done NN architecture of two layer network(with 1 hidden layer) of 4 neurons, 16 neurons, 32 neurons, 50 neurons and 3 layer 2 hidden layers with 16 neuron and 10 neuron.

- Accuracy doesn't improve on adding the multiple layers e.g 2 architecture with more neurons outperforms 3 layer architecture

- The curve of the adam optimizer is relatively smoother than the sgd optimizer. In the sgd curve we see the training loss and accuracy diverage as lot with the increased training.

Dataset2

Learning Curve

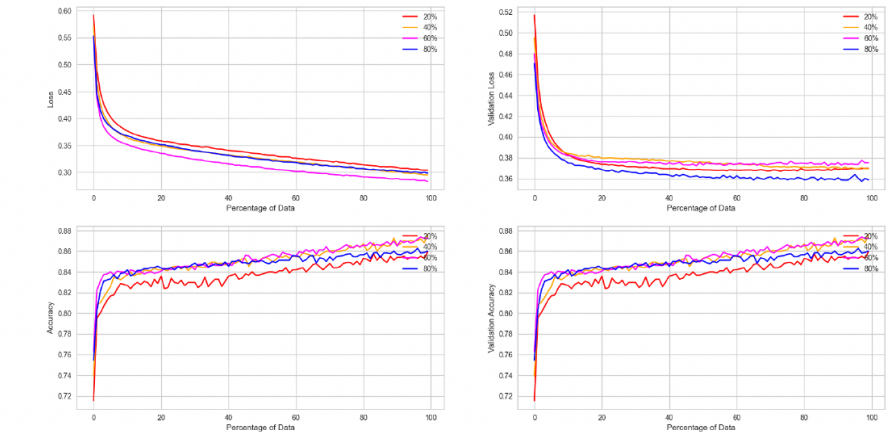

Figure 16: Learning Curve

Figure 16: Learning Curve

- two layer network(with 1 hidden layer) of 50 neurons

Validation Curve

- The model with 50 neurons in the hidden layer shows decent accuracy of : loss: 0.3044 - accuracy: 0.8582 - val loss: 0.3696 - val accuracy: 0.8300; Starts overfitting after 20 epochs

SVM Classifier

Classifier implementing the SVM. It explores the Linear,rbf and ploynomial kernel trick as well the tunig of C and gamma for the rbf.

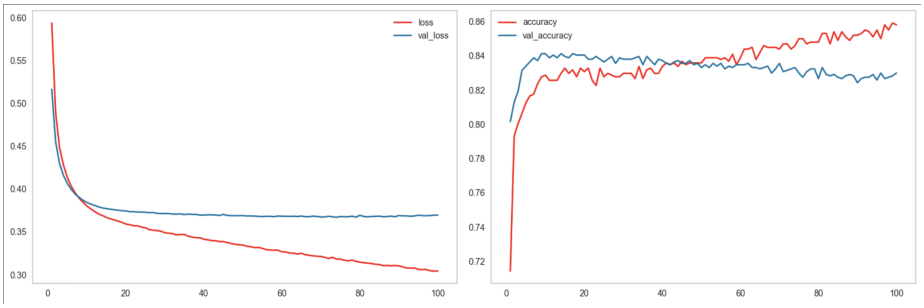

Figure 17: 1 layers(hidden neurons = 50

Figure 18: Neural Network - Wine Quality

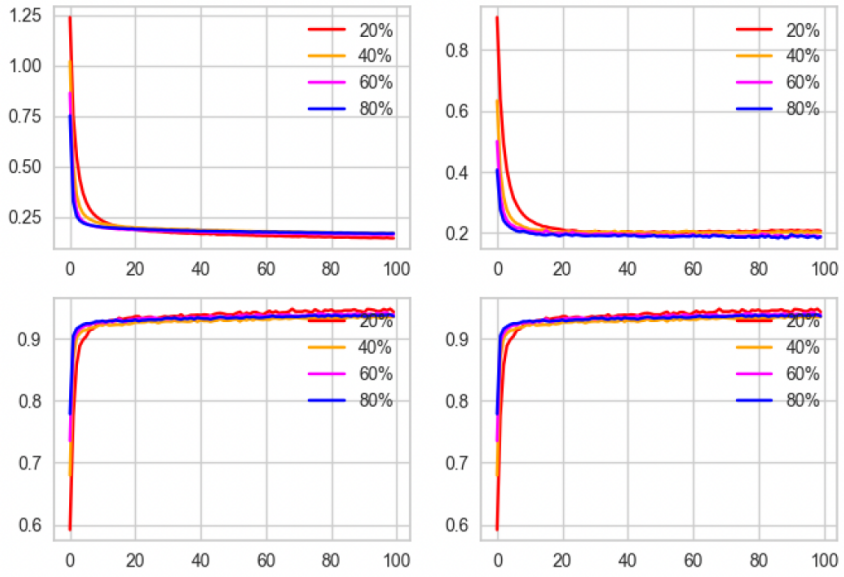

Figure 19: Learning Curve

Figure 19: Learning Curve

Dataset1

Learning Curve

- The learning curve Fig 19show that accuracy score increase with the increase in the size of the data.

Validation Curve

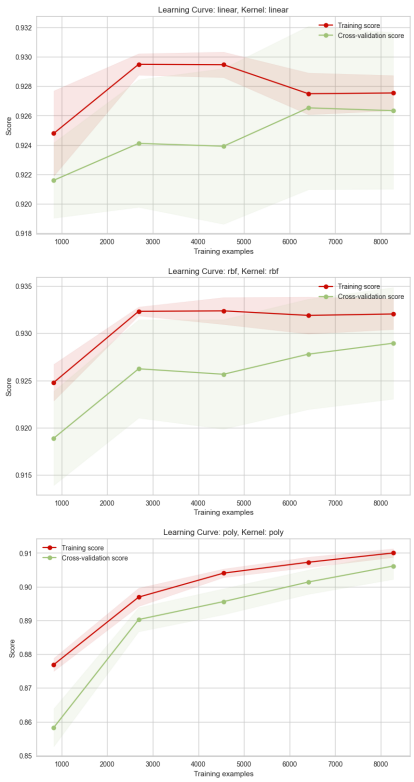

Figure 20: Validation Curve

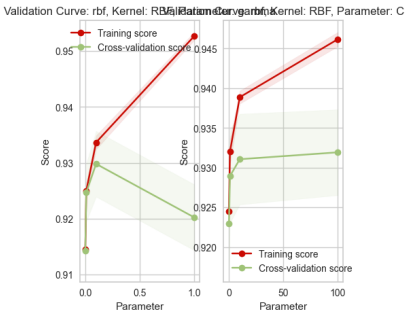

Figure 21: Hyperparamter tuning - rbf (C, gamma)

Figure 22:

Dataset2

Learning Curve

Similar to Dataset 1

Validation Curve

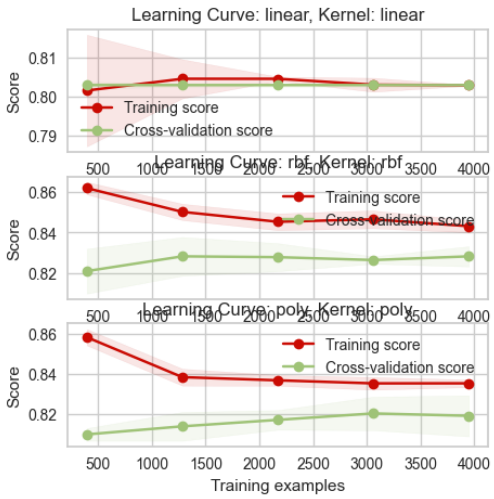

Figure 23: Validation Curve

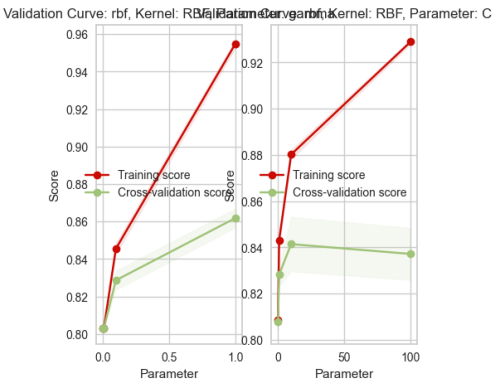

Figure 24: Hyperparamter tuning - rbf (C, gamma)

Figure 25:

Comparasion

[3]

	BeanData 1Training A	BeanData Test Accurac	Wine 1Training Acc	Wine Test Accuracy	Training Speed	Hyperparameter tuning
KNN	0.9262367565	0.7319061918	0.8308	0.640952284	Fast	Little
NN	0.9327	0.9383	0.8397	0.8184615374	Fast	Large
SVM	0.9285255036	0.9319407894	0.8738461538	0.8738461538	Fast	Medium

Figure 26: Algorithms comparison

1 References

References

1 - https://archive.ics.uci.edu/dataset/602/dry+bean+dataset
2 - https://archive.ics.uci.edu/dataset/186/wine+quality
3 - Some Heading taken from - https://github.com/danielcy715/CS7641-Machine-Learning/blob/master/Assignment1/ycai87-analysis.pdf