CS412

Introduction to Machine Learning

Homework Assignment - 1

Submitted By

Saikrishna Kalahasti Karthik

UIN: 655128301

Due Date: 30th January 2019

Important Details and Assumptions

Features Used

1. Variance of the means of 16 vertical columns in 16x16 pixel matrix (Plotted on x-axis)
2. Kurtosis of the means of 16 vertical columns in 16x16 pixel matrix (Plotted on y-axis)

Cross Validation Error (E_{cv})

Weighted F1 score has been used to measure the accuracy of the model. This was chosen to account for any difference in number of training samples among different classes. Every model is evaluated with 10-Fold Cross Validation(CV) to minimize sampling bias.

Train-Test Ratio

The given dataset was split in the following ratio

- 20% Training Data
- 80% Test Data

Drawing the First Graph

Select a way to graphically represent the data in two-dimensions

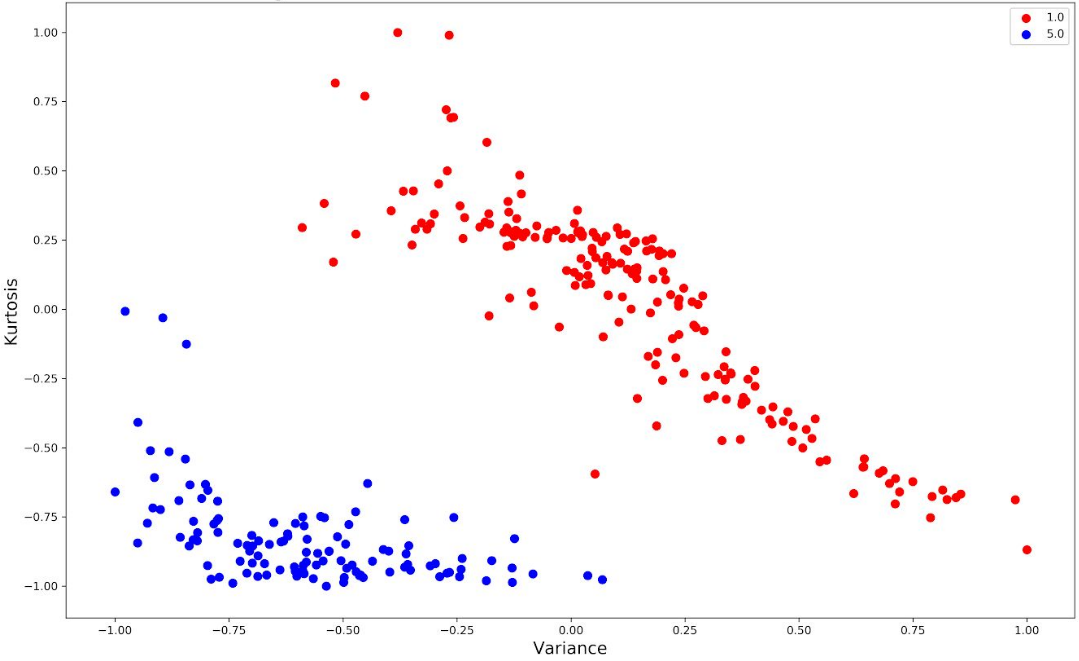
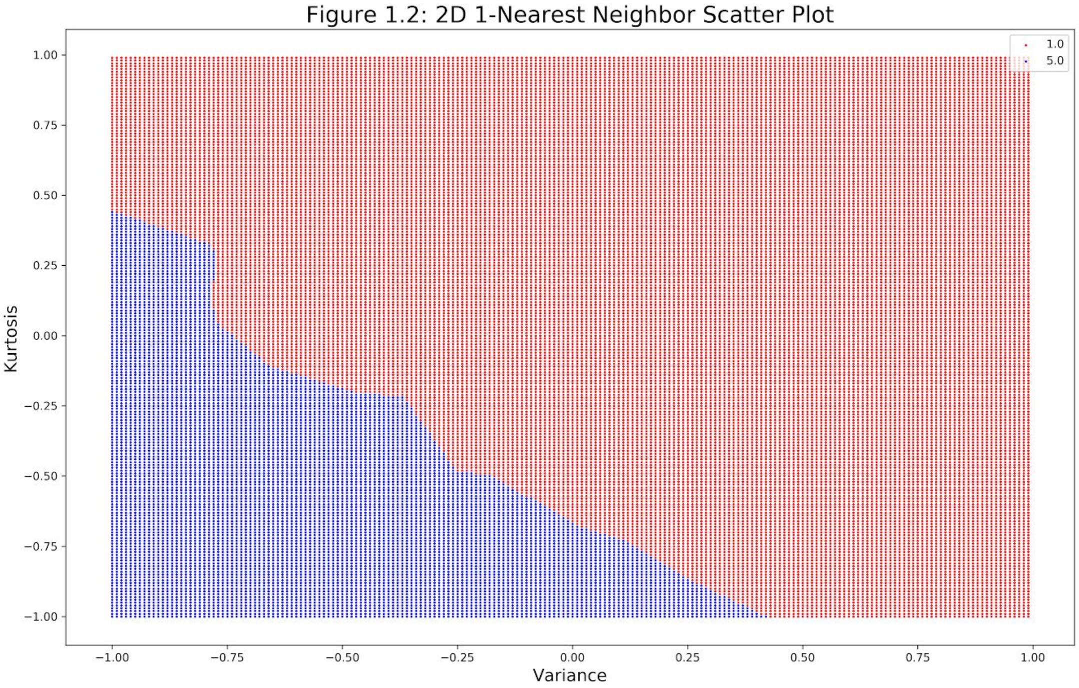


Figure 1.1: Variance - Kurtosis Vertical Data Scatter Plot

1-Nearest Neighbor

Color the regions of the graph



Do you believe that this model suffers from underfitting or overfitting? Why or why not?

This model overfits the data. The selected features themselves do a fair job of classifying 1.0 and 5.0. However, the 1-NN model is highly prone to noises in the data and outliers. Since the model considers only one nearest point during classification, it is prone to wrongly classify the points around outliers.

Is the error for this model equivalent to or different from the 1-Nearest Neighbor model in the 256-dimensioned space? Explain your answer.

Computationally, the CV F1-Scores of the 2D and the 256D 1-NN models are very similar (~1). Also, the 256D CV scores have higher variance than those of 2D model. One reasoning follows from curse of dimensionality. When the number of dimensions are high, the model overfits the data, which could be the reason for higher variance in the CV score.

However, representing numbers in a 256D space retains their shape as it is. When we plot the pixel intensities in such a space, the plots of 1 and 5 should be separable by a hyper-plane because 1 and 5 will have their own set of non-negative pixels. So it logically follows that, similar numbers will be grouped together based on the nearest point on the Euclidean space. Moreover, in the 2D space, we are using the same 256D points in a derived form. These points in the 2D Euclidean plane should be distributed similar to the points in 256D Euclidean space, although there could be significant losses when dimensions are reduced.

Use 10-fold cross validation to determine the cross validation error for the set under the following conditions and present their error for the following 1-Nearest-Neighbor problems

Sno	**Distance Metric**	**Feature Space**	**10-Fold Cross Validation F1 Score**
a)	*Euclidean*	2D	1.0
b)	*Manhattan*	2D	1.0
c)	*Chebyshev*	2D	1.0
d)	*Euclidean*	256D	0.9936
e)	*Manhattan*	256D	0.9936
f)	*Chebyshev*	256D	0.8949

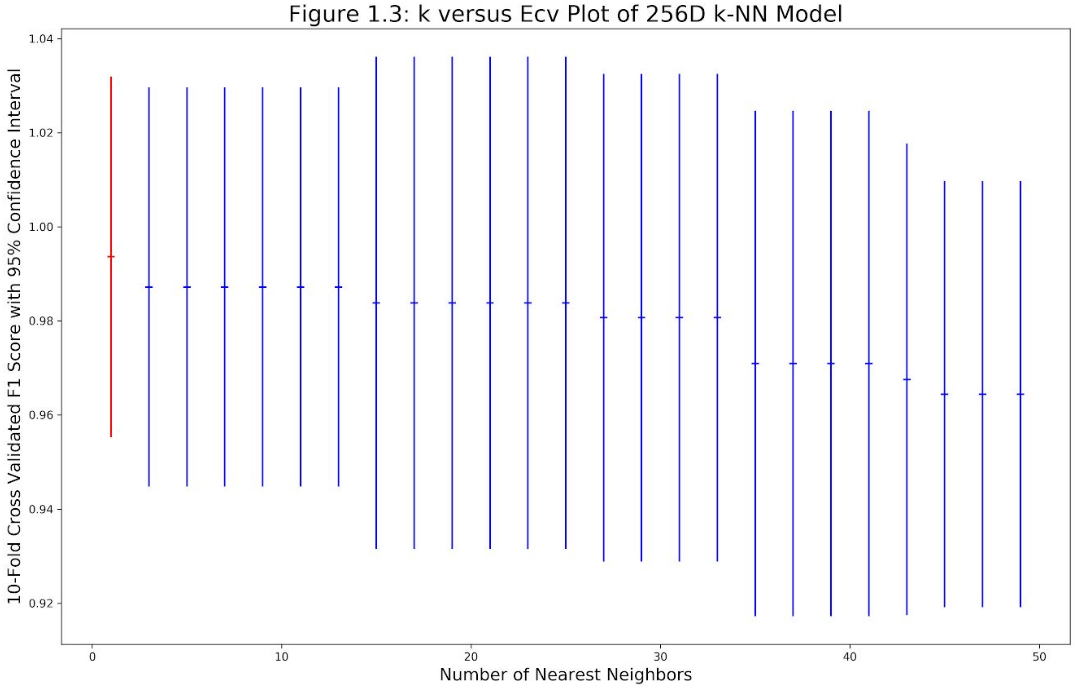
Comment on any differences you see in the results and what may have resulted in them

The only model with a different result is *256D Chebyshev 1-NN model*. As we know Chebyshev distance is the maximum distance along any co-ordinate dimension. Because of this property, it is quite possible that many points will have the nearest point by Chebyshev distance from another class.

Chebyshev distance will possibly work if two numbers have **no** overlapping pixels by the virtue of its shape, i.e same pixel **cannot** be non-negative for different numbers. Since this is not true for 1 and 5, i.e some pixels are non-negative for both 1 and 5, using Chebyshev distance increases error in the model.

k-Nearest Neighbor

Consider all of the odd k-Neighbor models between 1-49. Produce a graph of the 10-fold cross validation results for each of the 25 candidates and show their result



Explain your answer

The **best F1 Score 0.9936** occurs at **k=1** *(marked in red in the graph)*, i.e the 1-NN model gets the lowest CV error. I think this result is valid due to the following reasons:

1. As mentioned in the previous section (refer here), 1-NN model does not behave well in case of outliers. However, when the data is pure, 1-NN model performs the best because it classifies based on the nearest point. Also, such high accuracies for 1-NN model could be indicative of low noise in the data.

2. The training sample size is as low as 312 data points. When sample size is low, the data points are going to be sparse (assuming the population data is uniformly distributed), which will again result in low k. In other words, the points in the cluster boundaries will be wrongly classified when the data is sparse because, second or third nearest points themselves could be from a different class. I think when the sample size increases, higher k yields optimal E_{cv}.

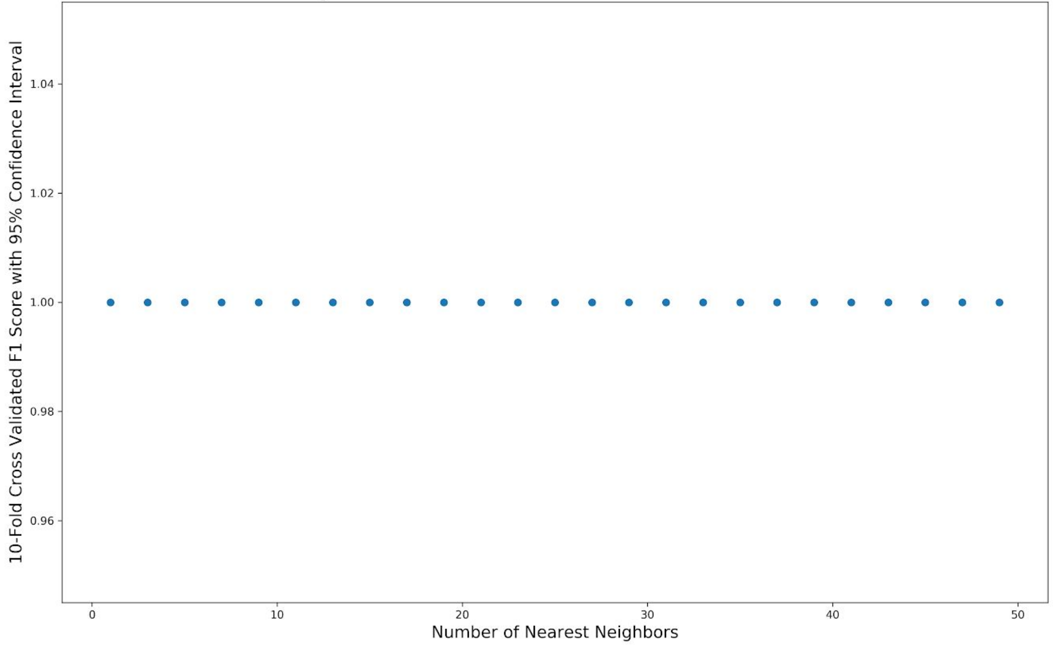
Give a graph of the 2-dimensional region for your optimal k-Nearest neighbor model and label this Figure 1.4. Does this model suffer from overfitting or underfitting? Explain your answer

Since k=1, Figure 1.4 will be same as Figure 1.2 (refer here). The answer for overfitting or underfitting is again identical to the answer given here.

Provide the estimated error of the 25 models at the 95% confidence level by utilizing the variance of the region. Select your model as that with the lowest 95% upper bound. Does this make a model more likely to overfit or underfit the data? Explain

The F1 score of the 2D k-NN model with the aforementioned features results in **1** for all folds of cross validation and for all values of k. We have a mean of 1, standard deviation of 0. This is represented graphically in figure 1.5. The best model selected is when k=1. The answer for overfitting or underfitting is again identical to the answer given here.

Figure 1.5: k versus Ecv Plot of 2D k-NN Model



A comment on choosing models with lowest upper bound of CV Error

When we select a model with the lowest upper bound of 95% confidence interval, we are more likely to minimize the risk of choosing models that overfit or underfit.

Extra Credit

Iris Dataset - Dataset for high k

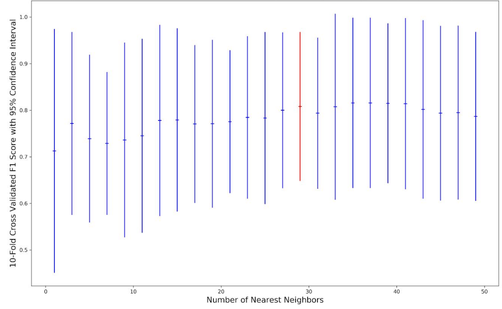


Figure 1.6: k versus Ecv Plot of k-NN Model on Sepal Length and Sepal Width

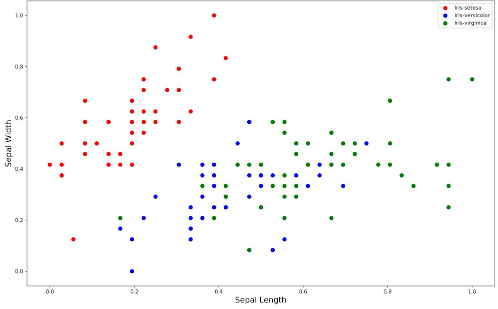


Figure 1.8: Sepal Length - Sepal Width Iris Scatter Plot

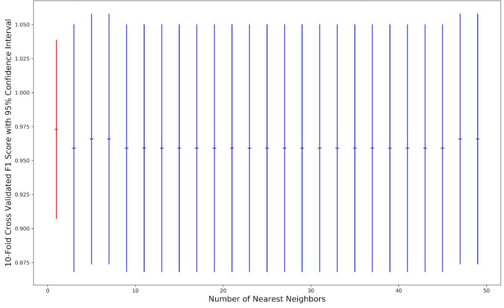


Figure 1.7: k versus Ecv Plot of k-NN Model on Petal Length and Petal Width

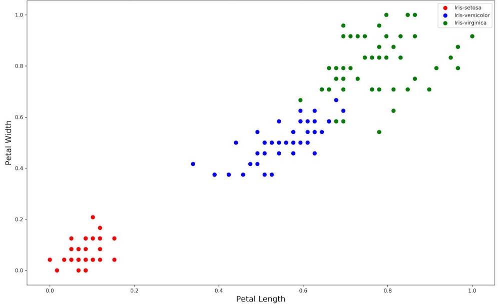


Figure 1.9: Petal Length - Petal Width Iris Scatter Plot

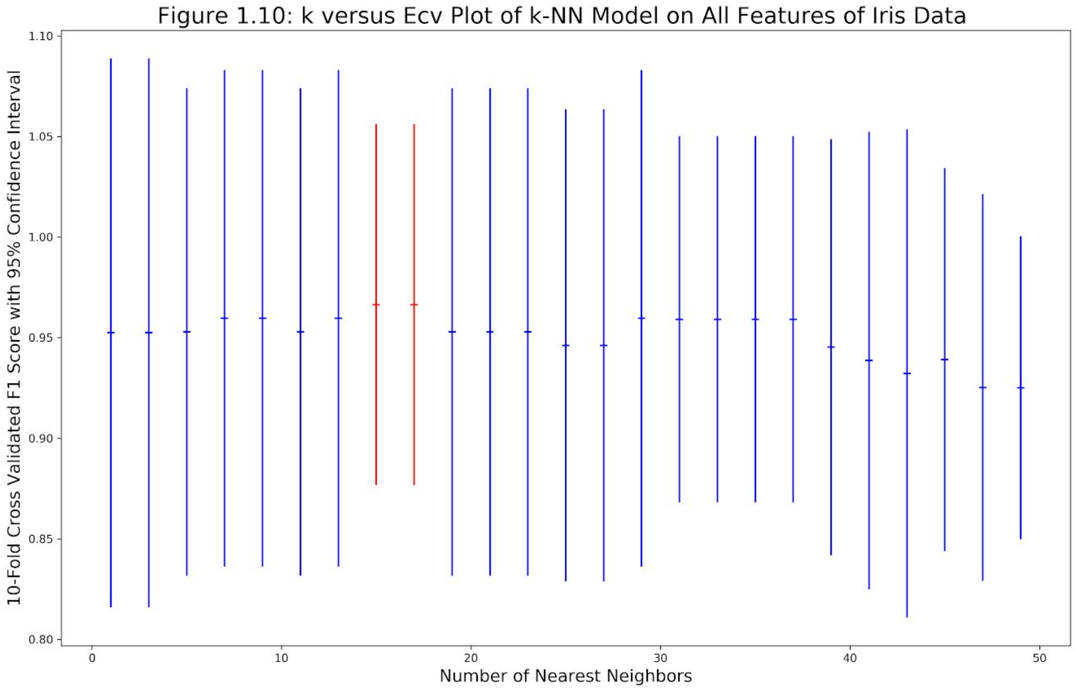
Explain what in the data may have led to this result and support your answer with figures as necessary.

When Sepal Length and Sepal Width were used as the features, an optimal k-NN model was obtained at k = 29 with an F1 score of 0.8081 (Refer Figure 1.6). However, when Petal Length and Petal Width were used as the features, an optimal k-NN model was obtained at k = 1 with an F1 score of 0.9731 (Refer Figure 1.7).

One direct explanation can be inferred from Figure 1.8 and Figure 1.9. When the selected features can create non-overlapping clusters with well-defined separating regions, low k is sufficient in the k-NN model to classify the points. When the clusters overlap in a certain feature space, then we need more data points to classify a given point, thus resulting in higher k. In this dataset, petal length and petal width create non-overlapping clusters, resulting in a lower k.

Additionally, give the E_{cv} for these models

The 10-Fold Cross Validated F1 score for this dataset when all features were used was 0.9665 obtained at k=15,17 (Refer Figure 1.10).

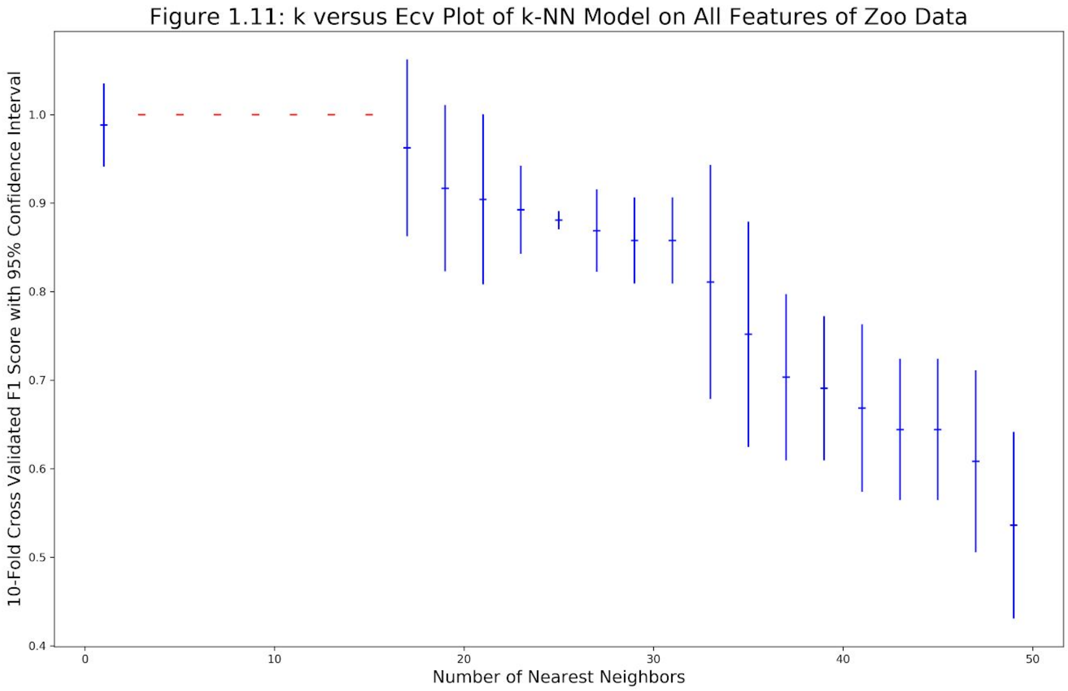


Zoo Dataset - Dataset for low k

This dataset originally had 7 classes of data. Classes 3,5 and 6 had less than 10 data points, therefore unsuitable for 10-Fold cross validation. These classes were removed from the dataset during the k-NN analysis.

Give the E_{cv} for these models

The 10-Fold Cross Validated F1 score for this dataset when all features were used was 1.0 obtained at k=3 (Refer Figure 1.11).



Explain what in the data may have led to this result and support your answer with figures as necessary.

As described here, more successful the features are in forming distinct clusters, more likely it is to achieve an optimal model in a lower k value.

The nature of this data is very deterministic, i.e the features can very easily classify across the classes. The classes of animals are really broad, and every class has a large number of features to tell them apart from the rest. This is the reason for a low k value.