Assignment 1 Report

Prepared by

SM Afiqur Rahman

ID: 201926862

&

Jubaer Ahmed Bhuiyan

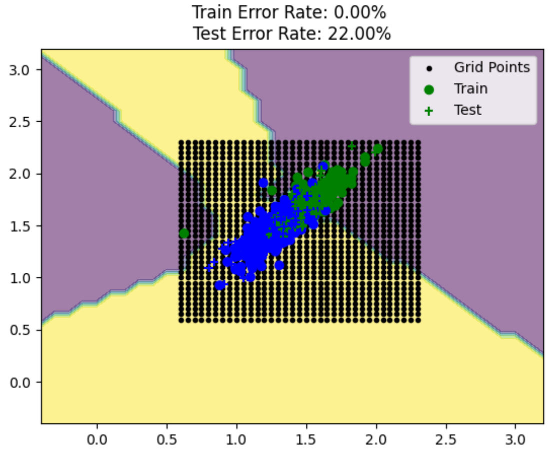
ID:

Course: Introduction to Machine Learning

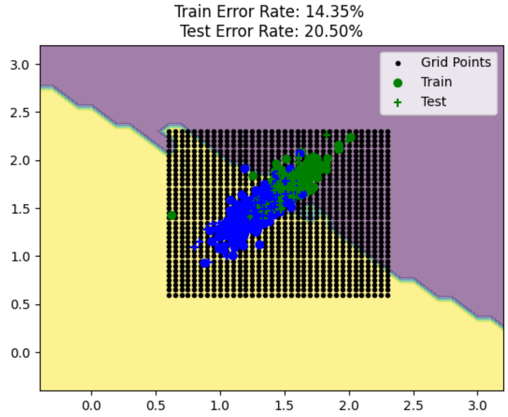
Instructor: Karteek Popuri

Question 1:

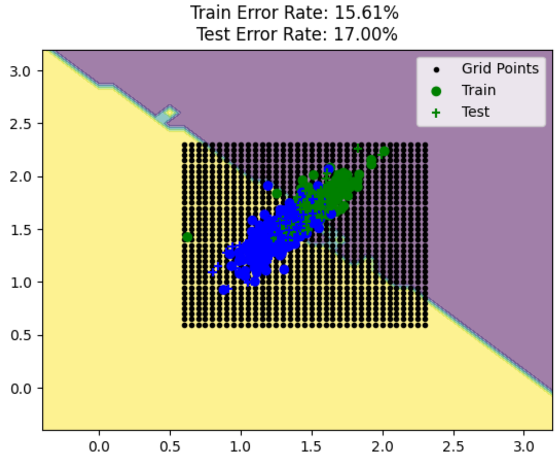
Images:



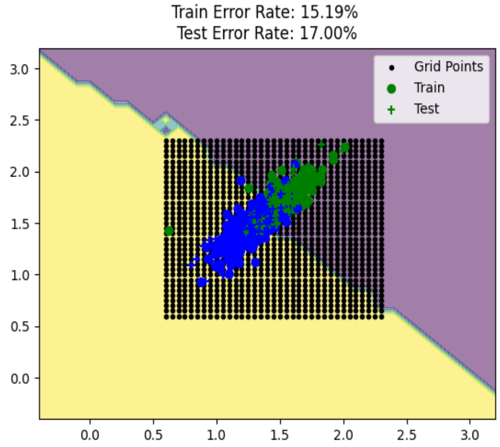
K=1



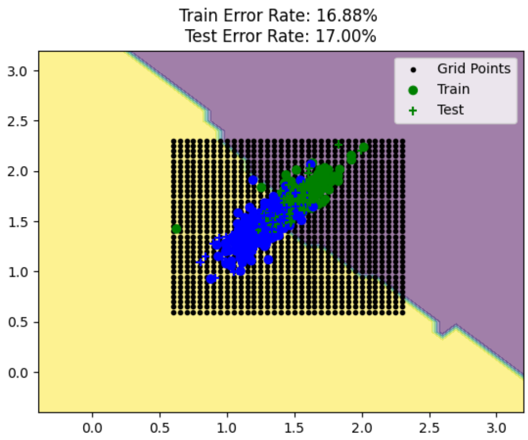
K =3



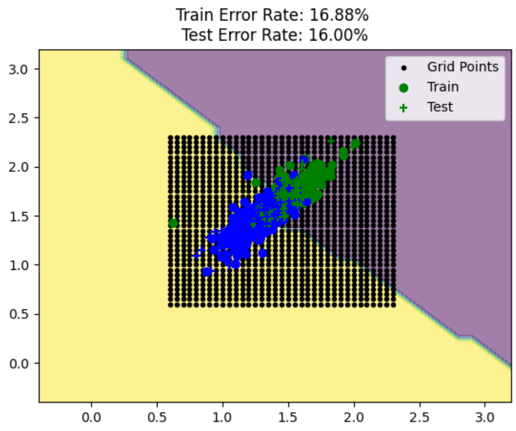
K=5



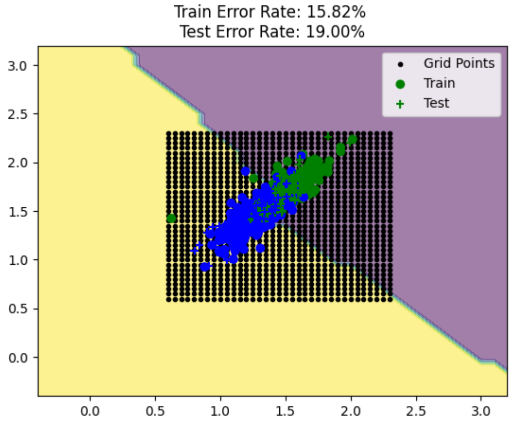
K=10



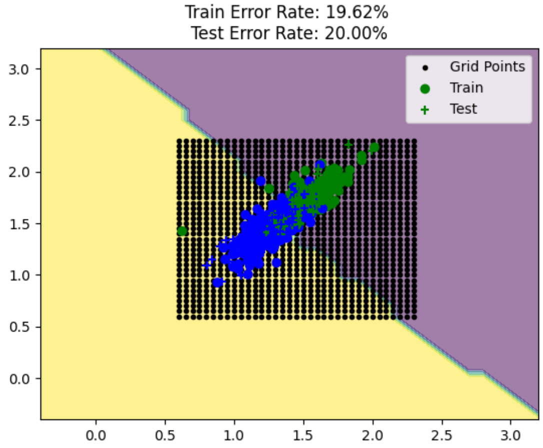
K=20



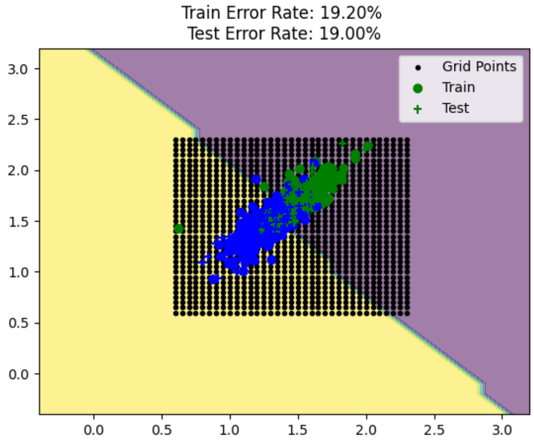
K=30



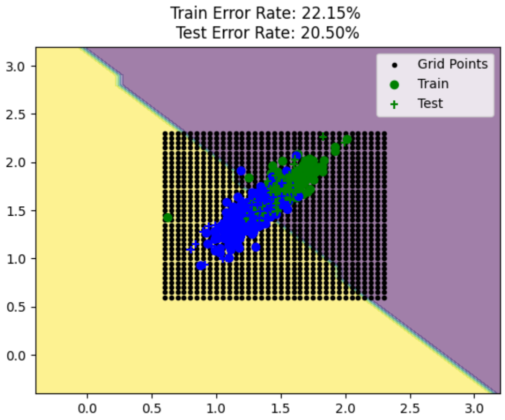
K=50



K=100



K=150



K=200

The decision boundary seems to get smoother and less complicated as the value of K rises, and the training error rate rises while the testing error rate initially reduces and subsequently climbs, according to the outputs produced.

Overfitting occurs when the decision boundary closely matches the training data for very low values of K (such as K=1). This is characterized by a low training error rate but a high testing error rate, demonstrating the model's poor generalization to unseen, new data.

The decision border smooths out as K grows, and the model generalizes to new data more well. Nevertheless, once K grows too big (for instance, K=200), the decision boundary smooths out and becomes extremely straightforward, which leads to underfitting. A low incidence of training and testing errors indicates that the model is too basic to accurately represent the complexity of the underlying data.

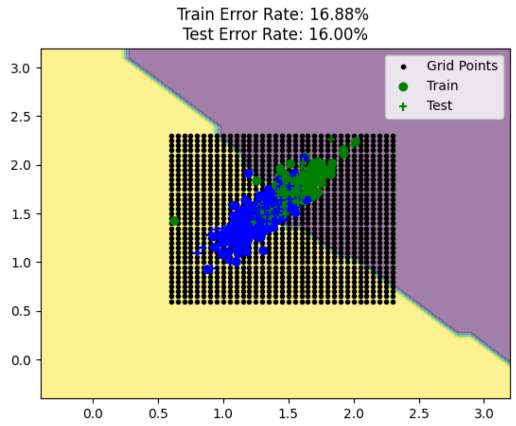
When a model makes unreliable assumptions about the data, it commits a systematic mistake known as bias. Contrarily, variation describes the degree to which the model's predictions change for various training sets. We can observe from the produced output photos that the model becomes less complicated and has lower variance as the value of k rises. This is due to the fact that as k rises, the model begins to depend less on the individual data points and more on its neighbors. A smoother decision boundary and less variation are produced as a result.

We also see that the model gets increasingly biased as k rises, which may lead to underfitting. This is due to the fact that as k rises, the model becomes less adept at capturing the intricate patterns seen in the data and is more likely to oversimplify the decision boundary.

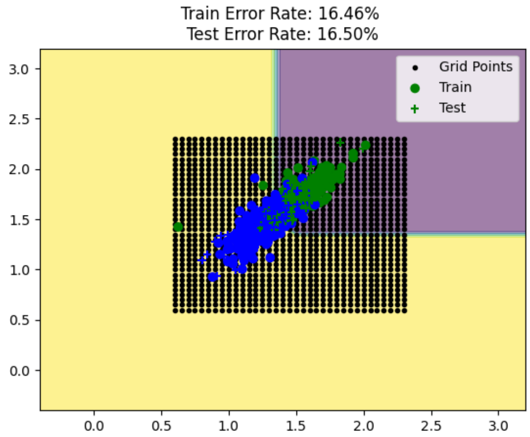
When choosing the value of k for a KNN model, bias and variance are generally trade-offs. A smaller k might lead to a simpler model with higher bias, whereas a bigger k could lead to a more complicated model with more variance. The ideal value of k will vary depending on the particular facts and the issue at hand.

In our case, K=30 gave the best results when building the classifier using Euclidean distance metric.

Question 2:



K=30 (Euclidean)



K=30 (Manhattan)

The performance of the classifier in the context of k-NN can be significantly impacted by the distance metric that is selected. By comparing the Manhattan distance metric to the Euclidean distance metric, we can see a clear difference in the classification borders between the two produced pictures.

By employing the Manhattan distance metric instead of the Euclidean distance metric, the classification border is less smooth and has a more step-like pattern. This is predicted since the Manhattan distance, which is calculated by adding the absolute differences between two points' attributes, frequently results in such borders. While Euclidean distance tends to generate more rounded borders, it is calculated by taking the square root of the sum of squared differences.

The bias and variance of the KNN classifier can be impacted by the distance metric selection. The classifier may match the training data better but may also be more sensitive to changes in the data if a more sophisticated distance measure, such as Euclidean distance, is used. The classifier may be less sensitive to changes in the data, but it may also underfit the training data, if you choose a simpler distance measure, such the Manhattan distance, which has a greater bias and a smaller variance.

In general, the particular situation and the properties of the data should be taken into consideration while selecting the distance measure. Using a simpler distance measure may be more acceptable in some situations while a more complicated distance metric may be more appropriate in others. It is also typical to experiment with several distance measurements and evaluate their effectiveness before choosing one.

Question 3:

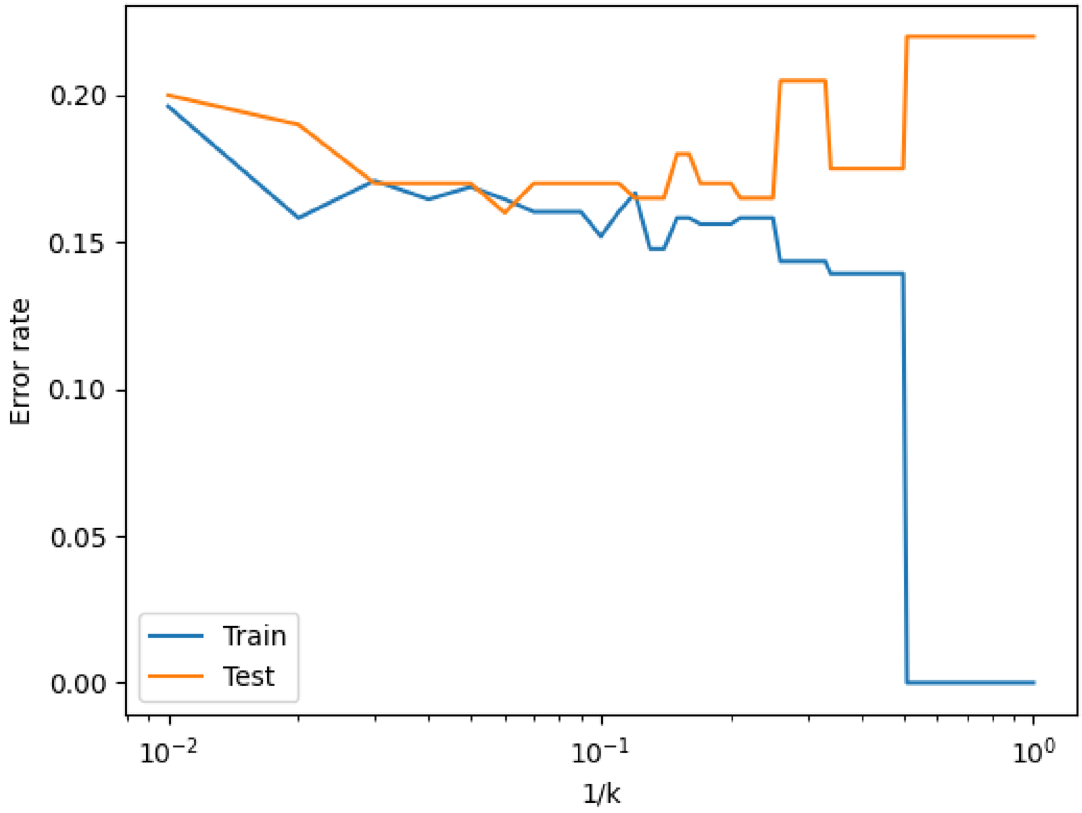


Fig: Error rate versus Model capacity plot

When the model capacity rises, or as the value of 1/k lowers, we can observe from the figure that the training and testing error rates both drop. This tendency is anticipated since a more complicated model that can better match the training data results from a larger model capacity.

When we get to the right side of the figure, the pace of decrease in the error rates begins to slow down, showing that, beyond a certain point, expanding the model capacity no longer significantly improves the model's performance. Alternatively said, the model begins to overfit the training set of data. The underfitting zone, where the model capacity is low and both the training and testing error rates are high, can be seen on the left side of the figure. This shows that the model is underperforming on both the training and testing datasets, indicating that it is too simplistic to capture the underlying patterns in the data.

On the right-hand side of the figure, when the model capacity is large, the training error rate is extremely low, but the testing error rate is high, is where the overfitting zone can be seen. This suggests that the model is overly complicated, which causes it to begin capturing noise and random changes in the training data, leading to subpar performance on the hidden testing data. In other words, rather than discovering underlying patterns, the model is just memorizing the training data.

The point at which the testing error rate is lowest—in this example, about 1/k = 0.3—is the sweet spot, or the ideal value of the model capacity. The model is sophisticated enough to capture the underlying patterns in the data but not so complicated that it starts to overfit the training data. This illustrates the ideal balance between bias and variance.