**5.4 Performance Comparison**

1. **KNN Classifier – Euclidian Distance**

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **Training Accuracy** | **Validation Accuracy** | **Testing Accuracy** |
| **1**  **3**  **5**  **7**  **9**  **11**  **13**  **15**  **17**  **19**  **21**  **23** | 77.7895  85.7895  88.9474  90.5263  90.3158  90.3158  89.4737  89.2632  87.8947  86.7368  86.3158  85.6842 | 78.6632  78.6632  81.7481  82.2622  83.0334  83.5476  84.0617  83.8046  83.2905  82.7763  82.2622  82.2622 | 75.5784  83.8046  85.8612  86.8895  87.4036  88.9460  87.4036  85.6041  85.6041  85.6041  84.3188  83.8046 |

**KNN Classifier – Hamming Distance**

|  |  |  |  |
| --- | --- | --- | --- |
| **K** | **Training Accuracy** | **Validation Accuracy** | **Testing Accuracy** |
| **1**  **3**  **5**  **7**  **9**  **11**  **13**  **15**  **17**  **19**  **21**  **23** | 82.9474  88.1053  88.9474  90.4211  90.6316  89.7895  88.9474  87.4737  86.4211  84.9474  84.7368  85.0526 | 82.2622  85.8612  86.6324  87.6607  87.4036  86.1183  85.3470  85.3470  84.3188  83.5476  83.2905  83.5476 | 85.0900  89.7172  88.9460  89.7172  91.5167  89.9743  89.4602  87.9177  86.6324  85.8612  84.8329  84.3188 |

1. **Decision Tree Classifier**

|  |  |  |  |
| --- | --- | --- | --- |
| **SN** | **Training Accuracy** | **Validation Accuracy** | **Testing Accuracy** |
| **1**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  **10**  **11**  **12**  **13**  **14**  **15**  **16**  **17**  **18**  **19**  **20** | 0.9674  0.9674  0.9653  0.9653  0.9632  0.9621  0.9600  0.9589  0.9526  0.9474  0.9379  0.9379  0.9379  0.9263  0.9263  0.9179  0.9126  0.9084  0.9084  0.8968 | 0.9332  0.9332  0.9383  0.9383  0.9409  0.9460  0.9486  0.9512  0.9434  0.9357  0.9332  0.9332  0.9332  0.9126  0.9126  0.9075  0.9023  0.8895  0.8895  0.8843 | 0.9434  0.9434  0.9434  0.9434  0.9460  0.9434  0.9460  0.9512  0.9383  0.9383  0.9434  0.9434  0.9434  0.9254  0.9254  0.9177  0.8895  0.8869  0.8869  0.8740 |

1. **Naïve Bayes Classifier : Uses log of probabilities**

Training Data Accuracy = 87.0526%

Validation Data Accuracy = 83.8046%

Testing Data Accuracy = 83.8046%

1. **Logistic Regression Classifier: multi-class logistic regression**

Training Accuracy = 94.53%

Validation Data Accuracy = 91.5167%

Testing Accuracy = 91.52%

**5.5 Decision Boundary**

Number of randomly generated points = 10,000

**Plot Analysis:**

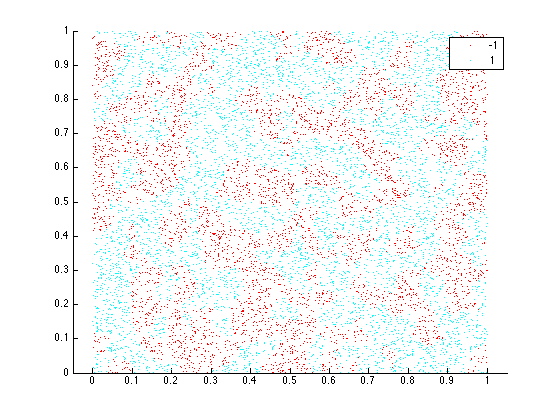
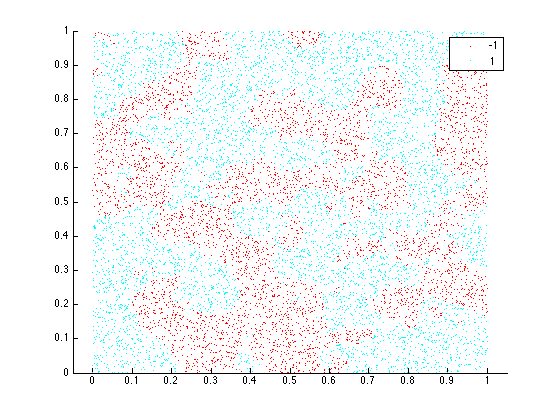
 

Figure 1: K = 1 Figure 2: K = 5

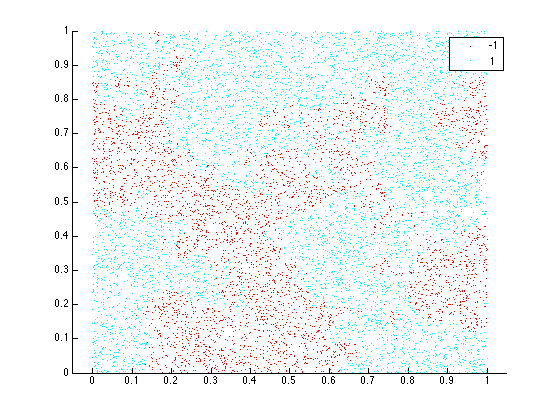
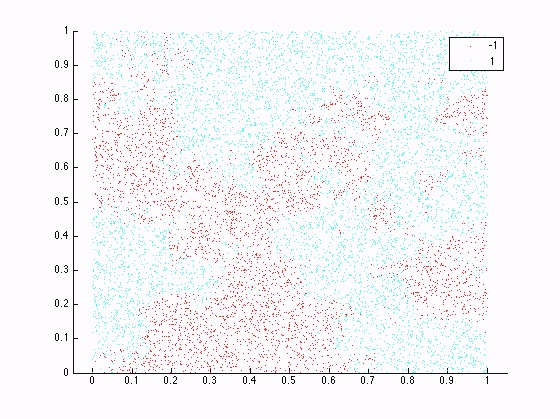
 

Figure 3: K=15 Figure 4: K=20

It can be seen from the analysis of the plots that:

Initially when K = 1, the decision boundary is pretty rough. Points from both classes are scattered over the plot.

As K increases gradually, the decision boundary becomes smoother and smoother. For K = 5, the blue class (1) is taking over the majority and splitting of points is becoming smooth and large spots are coming up.

For K=15, we can see that the decision boundary is very smooth, and large areas of classes have come up.

For K=20, we can see that the classification has pretty much converged and by further increasing the K we would be extremely overfitting, and the whole graph may turn blue (single class).

So it can be seen that as K increases the decision boundary becomes smoother and smoother.