**4.1 Using knn (k=2), EXCEL and the following training and test datasets, predict the species for the test datasets. What is the error rate? DO NOT normalize the data. You can copy and paste the tables into excel.**

Test Data Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 5.7 | 2.8 | 4.5 | 1.3 | virginica |
| 7.6 | 3 | 6.6 | 2.1 | virginica |

Training Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sepal.Length** | **Sepal.Width** | **Petal.Length** | **Petal.Width** | **Species** |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 3.6 | 1.4 | 0.2 | setosa |
| 5 | 3.3 | 1.4 | 0.2 | setosa |
| 7 | 3.2 | 4.7 | 1.4 | versicolor |
| 6.4 | 3.2 | 4.5 | 1.5 | versicolor |
| 6.9 | 3.1 | 4.9 | 1.5 | versicolor |
| 5.5 | 2.3 | 4 | 1.3 | versicolor |
| 6.5 | 2.8 | 4.6 | 1.5 | versicolor |
| 5.7 | 2.8 | 4.1 | 1.3 | versicolor |
| 6.3 | 3.3 | 6 | 2.5 | virginica |
| 5.8 | 2.7 | 5.1 | 1.9 | virginica |
| 7.1 | 3 | 5.9 | 2.1 | virginica |
| 6.3 | 2.9 | 5.6 | 1.8 | virginica |
| 6.5 | 3 | 5.8 | 2.2 | virginica |

**4.2 Load the “breast-cancer-wisconsin.data.csv” into R/Python from CANVAS (see the description bellow)**

1. **Remove the rows with missing values**
2. **Store every fifth record in a “test” dataset starting with the first record**
3. **Store the rest in the “training” dataset**
4. **Use knn with k=1 and classify the test dataset**
5. **Measure the performance of knn**
6. **Repeat the above steps with k=2, k=5, k=10.**

Results of Breast Cancer test

**Data Set Information:**

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at [[Web Link]](http://www.cs.wisc.edu/~street/images/)   
  
Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.   
  
The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].   
  
This database is also available through the UW CS ftp server:   
ftp ftp.cs.wisc.edu   
cd math-prog/cpo-dataset/machine-learn/WDBC/

Source: UCI