Lab6: Data classification using K-Nearest Neighbor and Bayes Classifier and Effect of Dimension Reduction in Classification

You are given the **Steel Plates Faults Data Set** as a csv file (`SteelPlateFaults-2class.csv`). This dataset contains features extracted from the steel plates of types A300 and A400 to predict whether an image contains two types of faults such as Z_Scratch and K-Scratch. It consists 581 tuples each having 28 attributes. The last attribute for every tuple signifies the class label (0 for K_Scratch fault and 1 for Z_Scratch fault). It is a two class problem. Other attributes are input features. For more information refer [1, 2].

1) Show the performance of **K-nearest neighbor (KNN) classifier** for different values of **K (1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21)**

   A. Find **confusion matrix (**use '*confusion_matrix'***)** for each K.

   B. Find the **classification accuracy** (You can use '*accuracy_score'*) for each K. Note the value of K for which the accuracy is high.

2) Show the performance of **Bayes classifier** with Gaussian distribution as class conditional density for each class. Consider the parameters (mean vector and covariance matrix) of Gaussian distribution estimated using maximum likelihood method as sample mean vector and sample covariance matrix.

   A. Find **confusion matrix (**use '*confusion_matrix'***)**.
   B. Find the **classification accuracy** (You can use '*accuracy_score'*).

3) Reduce the multidimensional data into *l* dimensions using **principle component analysis (PCA).** Now repeat Part 1 and 2 using reduced dimensional representation of each samples. Show the results for different values of *l* (1, 2, …, *d*). Here *d* is the actual dimension of the data.

Observation:

I. Comments on the accuracy for each classifiers.

II. Is there any significant reduction in the accuracy of classification after dimensionality reduction?

Notes:

a) Normalize the data using z**ero mean and unit standard deviation**

b) 70% of data from each class should be used for training and remaining for testing.

1. Results should be shown using confusion matrix and classification accuracy for all the assignment. (use inbuilt function '*confusion_matrix')*

Reference:

[1] M Buscema, S Terzi, W Tastle, A New Meta-Classifier,in NAFIPS 2010, Toronto (CANADA),26-28 July 2010.

[2] M Buscema, MetaNet: The Theory of Independent Judges, in Substance Use & Misuse, 33(2), 439-461,1998