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

You are given the **Diabetic Retinopathy Debrecen Data Set** as a csv file (`DiabeticRetinipathy.csv`). This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of a anatomical part or an image-level descriptor. It consists 1151 tuples each having 20 attributes. The last attribute for every tuple signifies the class label (0 no signs of diabetic retinopathy and 1 signs of diabetic retinopathy). It is a two class problem. Other attributes are input features. For more information refer [1].

Attribute Information:

Attribute1: The binary result of quality assessment. 0 = bad quality 1 = sufficient quality. Attribute2: The binary result of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack.
Attributes3-8: The results of MA detection. Each feature value stand for the number of MAs found at the confidence levels alpha = 0.5, . . . , 1, respectively.

Attributes9-16: Contain the same information as Attributes 3-8 for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructing the lesions, these features are normalized by dividing the number of lesions with the diameter of the ROI to compensate different image sizes.

Attribute17: The Euclidean distance of the centre of the macula and the centre of the optic disc to provide important information regarding the patient condition. This feature is also normalized with the diameter of the ROI.

Attribute18: The diameter of the optic disc.

Attribute19: The binary result of the AM/FM-based classification.

Attribute20: Class label. 1 = contains signs of diabetic retinopathy, 0 = no signs of diabetic retinopathy.

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.

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

Reference:

[1] Balint Antal, Andras Hajdu: An ensemble-based system for automatic screening of diabetic retinopathy, Knowledge-Based Systems 60 (April 2014), 20-27.