Machine Learning

Olivetti Face Data – Gaussian Naive Bayes Classifier (NBC)

Gaussian Naive Bayes classifier (NBC)
Download the Olivetti faces dataset.
Visit https://scikit-learn.org/0.19/datasets/olivetti_faces.html
There are 40 classes (corresponding to 40 people), each class having 10 faces of the individual; so there are a total of 400 images.
Here each face is viewed as an image of size 64 × 64 (= 4096) pixels; each pixel having values 0 to 255 which are ultimately converted into floating numbers in the range [0,1].

Split the dataset into train and test parts such that there are 320 images, 8 images per person (8 X 40) for training and 80 images, 2 images per person, (2 X 40) for testing.

Repeat the experiment using 10 different random splits having 320 training faces and 80 test faces as specified above and report the average accuracy

Use the Gaussian Naive Bayes classifier (NBC) to classify the test data and report the results

CODE:

Please find the code committed for Gaussian Naïve Bayes Classifier as *NaiveBayesClassifier_OlivettiFaceData_Impl.py*

- The dataset is first split into 8 imager per person (8*40) for training and 2 images per person(2*40) for testing. To achieve this array manipulation of Olivetti dataset is done.
- **GaussianNB** from sklearn is used on a train size of 320 samples (80%) of the olivetti dataset that contains a total of 400 samples.
- Accuracy is found for 10 iterations where in each iteration , the train and test samples are randomly selected.

RESULT:

Average Accuracy of 10 iteration is ~97.8% for Gaussian NBC.

accuracy_score for iteration 1 0.9125
accuracy_score for iteration 2 0.95
accuracy_score for iteration 3 0.9708333333333333
accuracy_score for iteration 4 0.9875
accuracy_score for iteration 5 0.995
accuracy_score for iteration 6 0.9895833333333334
accuracy_score for iteration 7 0.9946428571428572
accuracy_score for iteration 8 0.9890625
accuracy_score for iteration 9 0.9958333333333333
accuracy_score for iteration 10 0.99625
Avearge accuracy = 0.9781205357142856

INFERENCE/ANALYSIS:

- **Gaussian NBC** works well and provides a high accuracy of ~97% for the Olivetti dataset.
- **In Olivetti dataset** there are 4096 features which forms X and the class variable y represents each individual person among 40 persons.
- **Gaussian NBC** provides high accuracy for continuous independent features and the general assumption is that these features are distributed according to normal gaussian distribution.

As Gaussian or Bernoulli NBC is applied over Olivetti dataset completely, the average accuracy of classification is very high. However, when clustering is applied, whether its single link, complete link or K-means++, the feature set is reduced and the representation is modified. Complete link-based clustering provides better accuracy than K means++ and Single link accuracy. This impacts the overall avg accuracy of the Gaussian NBC model.

RESOURCES USED FOR THE ASSIGNMENT:

- **Environment:**
  Anaconda, Jupyter notebook
- **Software :**
  Python
  Python libraries/modules: Pandas, Numpy, SkLearn etc