Machine Learning

Olivetti Face Data – Bernoulli Naive Bayes Classifier

Bernoulli Naive Bayes classifier
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.

Convert the data into binary form by replacing any feature value below 0.5 by a 0 and value greater than or equal to 0.5 by a 1.

Use the Bernoulli Naive Bayes classifier to classify the test data and report the results

CODE:

Please find the code committed for Bernoulli Naïve Bayes Classifier as *NaiveBayesClassifier_Bernoulli_OlivettiFaceData_Impl.py*

- **BernoulliNB** from sklearn is used to solve this subtask.
- The dataset is converted into binary format by initially changing the feature values to either 0 or 1 as defined in problem statement above.
- **Bernoulli Naïve Bayes classifier** is applied over this modified dataset and average accuracy is found for 10 iterations.

RESULT:

Average Accuracy of 10 iteration is ~92.3 % for Bernoulli NBC.

accuracy_score for iteration 1 0.7875
accuracy_score for iteration 2 0.9
accuracy_score for iteration 3 0.9166666666666666
accuracy_score for iteration 4 0.934375
accuracy_score for iteration 5 0.95
accuracy_score for iteration 6 0.9520833333333333
accuracy_score for iteration 7 0.9535714285714286
accuracy_score for iteration 8 0.9484375
accuracy_score for iteration 9 0.9458333333333333
accuracy_score for iteration 10 0.9475
Avearge accuracy = 0.9235967261904762

INFERENCE/ANALYSIS:

- **Bernoulli NBC is** one of the classifiers which is a part of the family of Naïve Bayes.
- It considers only binary values as input to the classifier and hence in this problem, we did pre processing of the data into binary format based on the value of each feature.
- Given that the input is binary, this greatly helps in reducing the computational complexity of this algorithm.
- Similar to Gaussian NBC it requires independent features in the dataset.
- **Compared to Gaussian NBC, the accuracy is slightly lower for Bernoulli NBC as it has smoothened the feature values into binary and some data would hence be lost in decision making.**

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