Aykut Aykut
0034039
October 17th, 2018
Comp421 Homework01

In this homework, I was asked to implement a naive Bayes' classifier which can clasify 20*16 pixel images to 5 distinct classes, A, B, C, D, E. In order to do that , I followed these 14 steps:

1) I read the x_data_set which contains all 320 (20*16) features of 195 images into memory and y_data_set which contains the true class labels of 195 images. In both files, each row corresponds to a image.
2) I changed the class labels with the following mapping in order to obtain numerical class labels for the naive Bayes' classifier:
    a) A -> 1
    b) B -> 2
    c) C -> 3
    d) D -> 4
    e) E -> 5
3) Train-test split: I splitted the x_data_set and y_data_set into 2 groups, training and test. First 25 images of each class in x_data_set and y_data_set are in training set and remaning 14 images of each class in x_data_set and y_data_set are in test set.
4) I merged training sets of each class into x_training_data_set with *rbind* function and reseted their index. After that in order to apply matrix multiplication, I converted it into data matrix. I applied the same procedure for x_test_data_set, y_training labels and y_test_labels.
5) I removed unuseful variables which I used on the way preparing training and test data matrices.
6) I calculated class prior estimates with the following formula:
$$\widehat{p}(y = c) = \frac{\textit{total number of images in class c}}{\textit{total number of images}}$$ where c=1, 2, 3, 4, 5
7) I calculated pcd parameter estimates with the following formula:
$$\widehat{p}_{cd} = \frac{\sum_{i=1}^{N} x_{id} * [1(y_i == c)]}{\sum_{i=1}^{N} [1(y_i == c)]}$$ where N=125, c=1, 2, 3, 4, 5 and d=1, 2, …, 320
8) I defined a *safelog* function which is the same with the *safelog* function the instructor asked us to use.
9) I defined a *get_label* function which takes a vector and returns the column index of the maximum element in input vector.
10) I defined *predict_labels* function which takes a matrix *m*, applies the *get_label* function to each row of *m* and returns the resulting matrix.
11) I calculated the score values of each 5 class for each training image using the following formula:

$g_c(x) = \sum_{d=1}^{D} [x_d * log(\widehat{p}_{cd}) + (1-x) * log(1-\widehat{p}_{cd})] + log(\widehat{p}(y = c))$ where D=320 and c=1, 2, 3, 4, 5

12) I column combined score values of 5 classes of each training image with *cbind*, and then predicted a label for each training image using *predict_labels* function. I calculated the training confusion matrix and it is as follows:

```
> print(train_confusion_matrix)
            y_train_truth
y_train_hat  1  2  3  4  5
           1 25  0  0  0  0
           2  0 24  1  0  1
           3  0  0 24  0  0
           4  0  1  0 25  0
           5  0  0  0  0 24
```

13) I calculated the score values of each 5 class for each test image using the same formula at article 11. I also column combined score values of 5 classes of each test image with *cbind*, and then predicted a label for each test image using *predict_labels* function. I calculated the test confusion matrix and it is as follows:

```
> print(test_confusion_matrix)
            y_test_truth
y_test_hat  1  2  3  4  5
           1  7  0  0  0  0
           2  0 11  3  2  4
           3  0  0  7  0  0
           4  7  3  3 12  0
           5  0  0  1  0 10
```

14) I obtained the same results, pcd vectors for each 5 class, training confusion matrix and test confusion matrix with the results given in the homework description.