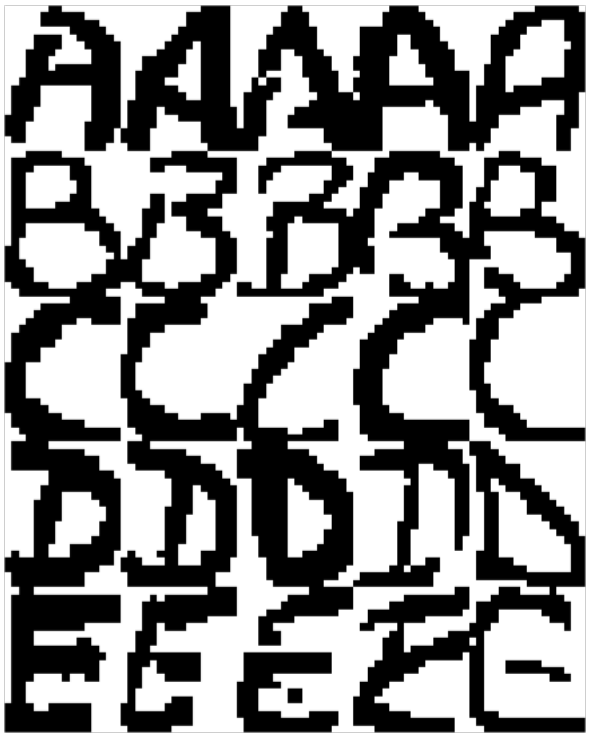
ENGR 421 / DASC 521: Introduction to Machine Learning
Homework 02: Naïve Bayes' Classifier

Deadline: March 21, 2022, 11:59 PM

In this homework, you will implement a naïve Bayes' classifier using Python. Here are the steps you need to follow:

1. Read Section 5.7 from the textbook.

2. You are given a multivariate classification data set, which contains 195 handwritten letters of size 20 pixels × 16 pixels (i.e., 320 pixels). These images are from five distinct classes, namely, A, B, C, D, and E, where we have 39 data points from each class. The figure below shows five sample figures from each class. You are given two data files:
   a. `hw02_data_set_images.csv`: letter images,
   b. `hw02_data_set_labels.csv`: corresponding class labels.



3. Divide the data set into two parts by assigning the first 25 images from each class to the training set and the remaining 14 images to the test set. (10 points)

4. Estimate the parameters $\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{p}}_2, \hat{\boldsymbol{p}}_3, \hat{\boldsymbol{p}}_4, \hat{\boldsymbol{p}}_5, \hat{P}(y = 1), \hat{P}(y = 2), \hat{P}(y = 3), \hat{P}(y = 4)$, and $\hat{P}(y = 5)$ using the data points you assigned to the training set in the previous step. Your parameter estimations should be like the following figures. (30 points)

```
print(pcd)
[[0.   0.   0.   ... 0.68 0.68 0.64]
 [0.04 0.24 0.24 ... 0.8  0.56 0.24]
 [0.   0.   0.   ... 0.6  0.88 0.8 ]
 [0.12 0.44 0.4  ... 0.28 0.16 0.08]
 [0.   0.12 0.12 ... 0.32 0.48 0.44]]

print(class_priors)
[0.2 0.2 0.2 0.2 0.2]
```

5. Calculate the confusion matrix for the data points in your training set using the parametric classification rule you will develop using the estimated parameters. Your confusion matrix should be like the following matrix. (30 points)

```
print(confusion_matrix)
y_truth   1   2   3   4   5
y_pred
1        25   0   0   0   0
2         0  24   1   0   1
3         0   0  24   0   0
4         0   0   0  25   0
5         0   0   0   0  24
```

6. Calculate the confusion matrix for the data points in your test set using the parametric classification rule you will develop using the estimated parameters. Your confusion matrix should be like the following matrix. (30 points)

```
print(confusion_matrix)
y_truth  1   2  3   4   5
y_pred
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
```

What to submit: You need to submit your source code in a single file (.py file) named as ***STUDENTID.py***, where ***STUDENTID*** should be replaced with your 7-digit student number.

How to submit: Submit the file you created to Blackboard. Please follow the exact style mentioned and do not send a file named as ***STUDENTID.py***. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.