ENGR 421 / DASC 521: Introduction to Machine Learning
Homework 02: Naïve Bayes' Classifier
Deadline: October 30, 2021, 11:59 PM

In this homework, you will implement a naïve Bayes' classifier in 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 35000 clothing images of size 28 pixels × 28 pixels (i.e., 784 pixels). These images are from five distinct classes, namely, T-shirt, Dress, Coat, Shirt, and Bag. The figure below shows five sample clothing images from each class. You are given two data files:
   a. `hw02_images.csv`: clothing images,
   b. `hw02_labels.csv`: corresponding image labels (1: T-shirt, 2: Dress, 3: Coat, 4: Shirt, 5: Bag).

3. Divide the data set into two parts by assigning the first 30000 images to the training set and the remaining 5000 images to the test set.

4. Estimate the mean parameters $\hat{\mu}_{1,1}, \hat{\mu}_{1,2}, \dots, \hat{\mu}_{1,784}, \dots, \hat{\mu}_{5,1}, \hat{\mu}_{5,2}, \dots, \hat{\mu}_{5,784}$, the standard deviation parameters $\hat{\sigma}_{1,1}, \hat{\sigma}_{1,2}, \dots, \hat{\sigma}_{1,784}, \dots, \hat{\sigma}_{5,1}, \hat{\sigma}_{5,2}, \dots, \hat{\sigma}_{5,784}$, and the prior probabilities $\hat{P}(y = 1), \dots, \hat{P}(y = 5)$ using the data points you assigned to the training set in the previous step. Your parameter estimations should be similar to the following figures. Please note that, in Section 5.7, the naïve Bayes' classifier is derived for binary input features. However, in this homework, the input features are assumed to be continuous.

```
print(sample_means)
[[254.99866667 254.98416667 254.85616667 ... 254.679      254.87816667
  254.95933333]
 [254.99733333 254.99733333 254.9965     ... 254.96883333 254.99216667
  254.98866667]
 [254.99933333 254.99933333 254.99233333 ... 251.52483333 254.4725
  254.97483333]
 [254.99666667 254.98983333 254.91416667 ... 252.39516667 254.44166667
  254.93666667]
 [254.999      254.98433333 254.93783333 ... 250.673      253.23333333
  254.79083333]]

print(sample_deviations)
[[ 0.09127736  0.25609108  1.31090756 ...  5.29826629  3.9117332
   1.93959091]
 [ 0.2065419   0.2065419   0.2163818  ...  1.04076669  0.47057267
   0.70062226]
 [ 0.05163547  0.04081939  0.16002465 ... 18.43665868  6.7881694
   1.1061344 ]
 [ 0.18436076  0.21617116  1.81046936 ... 15.67799977  6.34549162
   1.79971911]
 [ 0.04471018  0.64582342  3.03248555 ... 23.62576428 13.9167006
   4.4727787 ]]

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 the following matrix.

```
y_truth     1     2     3     4     5
y_pred
1        3685    49     4   679     6
2        1430  5667  1140  1380   532
3         508   208  4670  2948   893
4         234    60   123   687   180
5         143    16    63   306  4389
```

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 the following matrix.

y_truth	1	2	3	4	5
y_pred					
1	597	6	0	114	1
2	237	955	188	267	81
3	92	25	785	462	167
4	34	11	16	109	29
5	40	3	11	48	722

What to submit: You need to submit your source code in a single file (.py file) and a short report explaining your approach (.doc, .docx, or .pdf file). You will put these two files in a single zip file named as ***STUDENTID.zip***, where ***STUDENTID*** should be replaced with your 7-digit student number.

How to submit: Submit the zip file you created to Blackboard. Please follow the exact style mentioned and do not send a zip file named as ***STUDENTID.zip***. 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.