CSE455/555 - Intro to Pattern Recognition
Problem Set 1: Bayesian Decision Theory

Due Date: Friday, February 26, 2021 11:59PM

In this problem you will apply discriminant analysis to recognize the digits in the MNIST data set `http://yann.lecun.com/exdb/mnist/`. As a bonus problem you will construct "Fisher digits". You will train your model using the training data sets ("train-images-idx3-ubyte.gz" and "train-labels-idx1-ubyte.gz") and test the performance using the test data set ("t10k-images-idx3-ubyte.gz" and "t10k-labels-idx1-ubyte.gz").

1 Task-1

The images are 28 x 28 pixels in gray-scale. The categories are 0, 1, ... 9. Concatenate the image rows into a 28 x 28 vector and treat this as your feature, and assume the feature vectors in each category in the training data ("train-images-idx3-ubyte.gz") have Gaussian distribution. Draw the mean and standard deivation of those features for the 10 categories as 28 x 28 images using the training images ("train-images-idx3-ubyte.gz"). There should be 2 images for each of the 10 digits, one for mean and one for standard deviation. We call those "mean digits" and "standard deviation digits".

2 Task-2

Classify the images in the testing data set ("t10k-images-idx3-ubyte.gz") using 0-1 loss function and Bayesian Decision Rule and report the performance.
Before coding the discriminant functions, review Section 2.6 in textbook "Pattern Classification by Richard O. Duda, Peter E. Hart and David G. Stork".
Answer the question: Why it doesn't perform as good as many other methods on LeCun's web page?

3 Task-3[Optional]

Construct the "Fisher digits" from the MNIST data set according to Sections 3.8.2 and 3.8.3 in textbook "Pattern Classification by Richard O. Duda, Peter E. Hart and David G. Stork".
These two web pages on Fisher Faces might be helpful:
`http://www.scholarpedia.org/article/Fisherfaces`
`https://www.bytefish.de/blog/fisherfaces/`.
Answer the following two questions about these sections:

(a) Why should the vector w minimizing Eq. (103) satisfy Eq. (104)?
(b) Why should the between-class scatter matrix in Eq. (115) is $\frac{n_1 \times n_2}{n}$ times the one in Eq. (102) in two-class case (i.e., c=2)?
In addition, convince ourselves that Eq. (125) is the quotient between two "volumes" by referring the Wikipedia page on determinant `https://en.wikipedia.org/wiki/Determinant`.

4 Submission

Submit your solutions as one ipynb file through UBlearn. You can use Google Colab:
`https://colab.research.google.com/notebooks/intro.ipynb`
`https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c`.
The ipynb file should include your code, execution results, any explanations and answers to the questions. Use text cells to answer questions and add explanations.
Markdown guide for text cells:
`https://colab.research.google.com/notebooks/markdown_guide.ipynb#scrollTo=Lhfnlq1Surtk`
`https://colab.research.google.com/notebooks/basic_features_overview.ipynb#scrollTo=4hfV37gxpP_c`
You can also add math to text cells using LaTeX. Just place the statement within a pair of $ signs. Please typeset your mathematics. Do not upload pictures of handwriting math formulas. Math typesetting help: `https://www.codecogs.com/latex/eqneditor.php`

5 Libraries

Basic libraries are allowed, such as gzip, pickle, math, numpy, matplotlib, etc.
Do not use any Python libraries/toolboxes, built-in functions, or external tools/libraries that directly perform classification. You have to code the discriminant functions by yourself. Do not use libraries like sklearn to perform the discriminant analysis.

6 Rubric

Task-1:
50 points: 30 points code, 20 points output images(1 point each).
Task-2:
50 points: 30 points code,10 points performance report, 10 points question answers.
Task-3:
20 bonus points: 10 points code and result, 10 points question answers.

7 Acknowledgement

policies; (3) that your use of SafeAssign will be without recourse against Blackboard Inc. and its affiliates.

8 Academic Integrity

Academic integrity is a fundamental university value. Any violation will be reported to the University and will result in penalties in grades.
Do not share your answers with other students. This is an individual assignment. You are not allowed to work in groups. Working in groups and submitting similar answers is considered a violation of academic integrity.
Do not plagiarise someone else's words, ideas, or data you find online. Always cite your sources.