ECE 457B - Fundamentals of Computational Intelligence
Instructor: Dr. Amir-Hossein Karimi
Teaching Assistant: Kaixiang Zheng
Assignment #: 3
Department: Electrical and Computer Engineering — UWaterloo
Term: Winter 2024
Student Name:
UID:

NOTE: You should submit a pdf for Crowdmark and a zip file with your solutions & code on Learn. For your pdf submission on Crowdmark:

1. Please include your **name** and **student UID** at the start of your submission (for Exercise 1 at least).
2. Do **NOT** include links to your code in your Crowdmark submission. Instead, you **SHOULD** include the pdf/screenshots of your .ipynb or .py files.

Assignment 3

Exercise 1: MLE and Naïve Bayes (20 marks)

1. Given a Laplace distribution $f(x;\mu,b) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$, what are the maximum likelihood estimates of μ and b? (10 marks)
   Hint: $\arg\min_m \sum_{i=1}^{n} |x_i - m| = \text{median}\{x_1, x_2, \ldots, x_n\}$.
2. Given the following training set comprised of 8 samples, predict the label for a test sample $(X_1, X_2, X_3) = (-1, 1, 0)$ using Naïve Bayes (with Laplace smoothing). (10 marks)

	X_1	X_2	X_3	Y
①	0	0	0	1
②	0	0	1	0
③	0	1	1	0
④	0	1	1	0
⑤	0	0	1	1
⑥	1	0	1	1
⑦	-1	0	1	0
⑧	-1	0	1	0

   Note that X_1, X_2 and X_3 are three features of the data, with alphabets $\{-1, 0, 1\}$, $\{0, 1\}$ and $\{0, 1\}$ respectively. Y denotes the label with the alphabet $\{0, 1\}$, i.e., binary classification.

Exercise 2: K-Means Clustering (20 marks)

1. Given the following tabular dataset consisting of 8 samples (each with 2 features X_1 and X_2), conduct k-Means clustering with initial centroids $(0, 0)$ and $(2, 1)$. Write down the complete calculation process. (10 marks)

	X_1	X_2
①	0	1
②	3	3
③	1	1
④	2	3
⑤	1	0
⑥	0	0
⑦	3	2
⑧	2	2

2. A 2D toy dataset is given in `Q2.2_skeleton.py`. Use k-Means to cluster the provided data, and select the optimal hyperparameter k with the elbow method. Visualize the clustering result of the k you choose, and explain why you choose that k value. (10 marks)

Exercise 3: PCA (20 marks)

1. Given the following tabular dataset consisting of 8 samples (each with 2 features X_1 and X_2), conduct PCA to reduce the dimensionality from 2 to 1, and show the reconstruction error (difference between the original data and the reconstructed data) in an 8×2 matrix. Write down the complete calculation process. (10 marks)

   Note: To simplify calculation, please use the formula for sample covariance with the denominator n instead of $n-1$ for this question.

	X_1	X_2
①	1	1
②	1	1
③	4	0
④	0	0
⑤	0	0
⑥	1	1
⑦	1	1
⑧	0	4

2. Reduce the dimensionality of MNIST dataset to 2 using PCA, and visualize the data distribution of digits '0', '1' and '3' in a 2D scatter plot. Based on the observation, try to generate an image of digit '3' using 2D representations of digits '0' and '1'. Briefly write down your idea and show your generated image. (10 marks)

   Please build your code based on `Q3.2_skeleton.py`.

Exercise 4: Kaggle Competition (30 marks)

Kaggle is an online data science platform that's primarily known for its ML competitions where participants can pit their models against other on a (typically) ranked leaderboard which culminates in a podium finish. Kaggle is also known for its extensive dataset repository, you used the Titanic dataset from there in Assignment 1, and a discussion board for helping data science enthusiasts. For this exercise, you will have two parts to your submission.

1. A Jupyter/Colab notebook outlining the following: (23 marks)

(a) Data exploration: Explore the data and comment on the distribution of the dataset. Use graphs where needed to illustrate your exploration. (3 marks)

(b) Data pre-processing: Load your data into an appropriate data loader for your model. (3 marks)

(c) Setting up a Neural Network Model: Build a simple neural network model (no CNN, RNN, LSTMs, GRU, etc) that takes a value for the number of hidden layers, the number of nodes each hidden layer should have, a learning rate and the number of epochs. Use cross-entropy loss as your loss criterion, and use an adam optimizer for optimization. For a learning rate of 0.001 and 100 epochs of training, Compute the accuracy_scores, class-wise accuracy, and f1 scores for your model with the following structures. Plot the train vs validation error over the 100 epochs of training for each structure. Compare the three structures and comment on what you see from their performances.

   i. For a model with 1 hidden layer, and 64 hidden nodes. (2 marks)

   ii. For a model with 2 hidden layers, and 32 hidden nodes for each hidden layer. (2 marks)

   iii. For a model with 2 hidden layers, and 64 hidden nodes each. (2 marks)

(d) You are now free to customize your model as you see fit. You can change your model's architecture, loss criterion, optimizer, and learning rate, and even use other layers such as CNN, RNN, LSTMs, GRU, transformers, etc. as you see fit. Train your model for up to 100 epochs. For your final model, compute the accuracy_scores, class-wise accuracy, and f1 scores for your model. Plot the train vs validation error as well. (Note: You cannot use or fine-tune an existing pre-trained model. This has to be a model you've built. You can, however, take inspiration from existing architectures.) (8 marks)

(e) Plot (or draw) a diagram showing the structure of your final model. Explain briefly your thoughts on the decisions you took for its design. (3 marks)

2. Submit your model on Kaggle for this competition for a spot on the leaderboard. There is a daily submission limit of 5 but you're allowed to update your submission as many times as you want before the deadline for the assignment.

   (a) Accuracy scores above 60% will get 2 marks

   (b) Accuracy scores above 70% receive an additional 3 marks

   (c) An explanation of what you would do to improve your model if you had more time/money/resources (2 mark)