Machine Learning - CSL7620

Assignment 2 (50 marks): Weightage **15%**

6th March 2025

Date of Submission: 13th April 2025; 23:59

Instructions (Non-conformance will result in penalties):

- A typed report must be written for the assignment (no handwritten reports will be accepted). Please ensure the report is neat, properly formatted, and contains all relevant information to the tasks in the assignment.
- Grading of the assignment shall be done based on both the codes and the report. The report shall reflect your understanding regarding the concepts, methods that you have used, and the analysis of the results. Please keep the content of your reports crisp and precise. **Max. 12 page report.**
- Submission should contain the following files - **one** .ipynb file to **LMS** and report (.pdf). to **Turnitin** (link will be shared later). Please name each of these as roll_number(s).ipynb and roll_number(s).pdf.
- Within the code notebook, please make proper sections for each question as discussed in the demonstration session.
- Any deviation from the above-mentioned submission format or deadline breach will result in a penalty of marks.
- Even though intellectual discussion is encouraged, any plagiarism within the report is completely unacceptable and will result in the nullification of marks and a possible action for academic misconduct (according to the relevant institute regulations).
- **You can do this in groups of either 2 or 3. Mention the contributions of all members.**

Part 1: Implementation Questions

Q1: Download the following dataset diabetes (1).csv - Google Drive. **(10 Marks)**

a) Find the optimum number of principal components for the features in the above-mentioned data

b) Use any two regression models of your choice and find the prediction accuracy and error between the reduced data (with an optimum number of principal components) and the complete data.

Q2: We will use the fashion-MNIST dataset for this question (you can download it from any other source also including libraries). Flatten and preprocess the data (if required) before starting the tasks. It will become a 784-dimensional data with 10 classes, more details are available in the link. **(20 Marks)**

a) Train the k-means model on f-MNIST data with k = 10 and 10 random 784-dimensional points (in input range) as initializations. Report the number of points in each cluster.

b) Visualize the cluster centers of each cluster as 2D images of all clusters.
c) Visualize 10 images corresponding to each cluster.
d) Train another k-means model with 10 images from each class as initializations , report the number of points in each cluster, and visualize the cluster centers.
e) Visualize 10 images corresponding to each cluster.
f) Evaluate Clusters of part a and part d with Sum of Squared Error (SSE) method. Report the scores and comment on which case is a better clustering.

Q3: Implementation of Neural Networks from Scratch Using NumPy and Comparison with Sklearn **(20 marks)**

a) Load and preprocess the MNIST Digits Dataset. (3 marks)
b) Implement a neural network with one input layer, one hidden layer, and one output layer using NumPy. (5 marks)
c) Train the neural network with various hyperparameters (e.g., learning rate, number of hidden nodes). (3 marks)
d) Evaluate the performance of the neural network on the testing set. (2 marks)
e) Implement the same neural network using sklearn and compare the results with the NumPy implementation. (4 marks)
f) Plot the training and validation loss/accuracy curves (for both experiments). (3 marks)

Part 2: Project Part

Q4: Select a project from the list provided in the document and implement a complete end-to-end machine learning pipeline for the same. Also, prepare a demo using gradio/streamlit for evaluation. For gradio sample app ref: https://colab.research.google.com/github/kirenz/lab-huggingface/blob/main/code/gradio.ipynb
(50 marks)
List of projects:
https://docs.google.com/document/d/1HYz5TA1QBhhutvWdKIKvTr4v6SzebjrKxZbklh8zsZo/edit?usp=sharing

---

Note: The report should contain detailed explanations and analysis for your observations. Just reiterating the code will not fetch you any marks.