ML Learners' Space

Assignment - 1

Hello, we hope you enjoyed the introductory lectures and gained valuable insights. Let's apply your knowledge with the following problems!

Problem 1

Generate a 2-dimensional NumPy array of shape $(5, 4)$ filled with random integers between 1 and 50. Perform the following tasks:

- Print the array and extract the elements along the anti-diagonal (from top-right to bottom-left).
- Compute and print the maximum value in each row of the array.
- Create a new array containing only the elements less than or equal to the overall mean of the array.
- Write a Python function `numpy_boundary_traversal(matrix)` that takes a NumPy matrix and returns a list of elements visited along the boundary of the array in clockwise order, starting from the top-left corner.

Problem 2

Create a 1-dimensional NumPy array of 20 random floats between 0 and 10. Perform the following tasks:

- Print the array and round all elements to two decimal places.
- Calculate and print the minimum, maximum, and median of the array.
- Replace all elements less than 5 with their squares.
- Write a Python function `numpy_alternate_sort(array)` that takes a 1D NumPy array and returns a new array with elements sorted in an alternating pattern (smallest, largest, second smallest, second largest, etc.).

Problem 3

Create a Pandas DataFrame with 10 rows representing student records, with columns `Name`, `Subject`, `Score` (random integers between 50 and 100), and `Grade` (initially empty). Perform the following tasks:

- Assign grades based on scores: A (90–100), B (80–89), C (70–79), D (60–69), F (below 60).
- Print the DataFrame sorted by `Score` in descending order.
- Calculate and print the average score for each subject.
- Write a Python function `pandas_filter_pass(dataframe)` that takes a DataFrame and returns a new DataFrame containing only the records with grades A or B.

Problem 4

Create a synthetic dataset which is a list of 100 short movie reviews (50 positive, 50 negative) stored in a Pandas DataFrame with columns `Review` (text) and `Sentiment` (positive/negative). Perform the following tasks:

- Tokenize the reviews using `CountVectorizer` with a maximum of 500 features, removing stop words.
- Split the dataset into training (80%) and testing (20%) sets using `train_test_split`.
- Train a Multinomial Naive Bayes classifier on the tokenized training data and print the accuracy on the test set.
- Write a Python function `predict_review_sentiment(model, vectorizer, review)` that takes a trained model, the fitted `CountVectorizer`, and a single review (string), and returns the predicted sentiment.

Problem 5

Create a synthetic dataset of 100 short text samples (e.g., product feedback) with binary labels (good/bad) and apply an NLP pipeline using scikit-learn. Perform the following tasks:

- Preprocess the text using `TfidfVectorizer` with a maximum of 300 features, applying lowercasing and stop word removal.
- Split the dataset into training (75%) and testing (25%) sets.
- Train a Logistic Regression model on the vectorized training data and print the precision, recall, and F1-score for the test set.
- Write a Python function `text_preprocess_vectorize(texts, vectorizer)` that takes a list of text samples and a fitted `TfidfVectorizer`, and returns the vectorized feature matrix.