CSE440: Natural Language Processing II

Lab Assignment 3

Problem 1 [OPTIONAL]

- Load the Wine dataset from sklearn.datasets.load_wine()
- Split the data into 75 % training and 25 % test sets.
- Define a feed-forward neural network with:
  - 1 hidden layer
  - 32 hidden units
  - ReLU activation
  - Output layer: 3 units with softmax activation
  - Optimizer: Adam
  - Loss: Categorical cross-entropy
  - Epochs: 150
- Train the network on the training set.
- Print the network's summary and the accuracy on the test set.
- Plot the training and validation loss (or accuracy) curves across epochs.

Problem 2

- Use keras.datasets.imdb to load the IMDb movie-review dataset. Limit the vocabulary to the top 15,000 most frequent words.
- Take the original training set and split it further into 75 % training and 25 % validation.
- Define a neural network with:

  - Use an embedding layer (15,000 input dim, 64 output dim), then a dense layer with 64 ReLU units, and a sigmoid output layer with 1 unit.

  - Loss: Binary cross-entropy.
  - Optimizer: Adam.
  - Epochs: 10–15

- Train the model on the training set..
- Evaluate the model on the official **test data** provided in keras.datasets.imdb.
  - Print the summary of your network.
  - Report the test accuracy.
  - Plot the training and validation loss and accuracy curves.
  - Draw the ROC curve and calculate the AUC (area under the curve) to evaluate the model’s performance beyond just accuracy.

Problem 3

- Reuse the preprocessed data from Question 2.
- Take the original training set and split it further into 80 % training and 20 % validation.
- Use an embedding layer (output dimension = 64) followed by three hidden dense layers:
  - Hidden Layer 1: 256 units, ReLU, Dropout = 0.4.
  - Hidden Layer 2: 128 units, ReLU, Dropout = 0.4.
  - Hidden Layer 3: 64 units, ReLU.
  - Output layer: 1 unit, sigmoid activation.
  - Loss: Binary cross-entropy.
  - Optimizer: Adam.
  - Epochs: 15–20.

- Train the model on the training set, then compare the shallow model from Question 2 with the deep model in terms of training time, test accuracy, overfitting (difference between training and validation performance), and confusion matrices. Summarize which model performs better overall and justify your conclusion.

Problem 4

- You are tasked with building an autoencoder model to reconstruct text data from the IMDb movie review dataset. Start by using keras.datasets.imdb to load the dataset.
- Autoencoder Architecture:
  - Encoder
    - Input layer: 15,000 units
    - Dense layer: 128 units, ReLU
    - Dense layer: 64 units, ReLU
  - Decoder
    - Dense layer: 128 units, ReLU
    - Output layer: 15,000 units, sigmoid or linear
- Loss: Binary cross-entropy.
- Optimizer: Adam
- Epochs: ~20–30
- Batch size: ~64

- After training the model, plot the training and validation loss curves to visualize the performance during training. Then, calculate the reconstruction error on the test set . Finally, compute the cosine similarity between the original and reconstructed vectors to evaluate how similar the reconstructed text is to the original.