Intermediate Progress Report: "Sentiment Analysis on IMDb Movie Reviews"

Course: CS454

Professor: Ahmet Ibrahim Ethem Alpaydın

Student Names: Roman Mordovtsev, Mehmet Fatih Karaosman

Project Description

This project aims to evaluate the effectiveness of different algorithms in sentiment classification of IMDb movie reviews. The objective is to determine whether a review is positive or negative using three different approaches:
1. K-Means Clustering (unsupervised)
2. Linear Perceptron (shallow supervised model)
3. Multi-Layer Perceptron (MLP) (deep supervised model)

Progress Overview

The following steps have been successfully completed:

- Data Download and Preparation:
  - IMDb dataset was downloaded from [Maas et al., 2011].
  - 10,000 reviews (5,000 positive, 5,000 negative) were sampled as per requirements.
  - Data was split into training (70%), validation (15%), and test (15%) sets.

- Text Processing:
  - TF-IDF vectorization with max_features=5000 and English stopword removal was applied.

- Model Implementation:
  - K-Means Clustering: Implemented to cluster data into 2 classes. Since K-Means is unsupervised, label mapping between clusters and sentiments was done manually.
  - Linear Perceptron: Trained on TF-IDF features, tested for baseline performance.
  - MLP Classifier: Configured with hidden_layer_sizes=(100,) and trained to capture non-linear decision boundaries.

- Evaluation Metrics:
  - accuracy_score and classification_report are used to compare performance across models.

Remaining Tasks

- Hyperparameter Tuning:
  - Each model will be tested under different configurations (e.g., number of clusters for K-Means, learning rate for Perceptron, and number of layers/neurons for MLP).

- Validation Analysis:
  - Use validation set to fine-tune model performance before final testing.

- Visualization:
  - Generate confusion matrices and training performance plots.

- Final Comparison:
  - Compare models based on accuracy, precision, recall, and F1-score to determine the most effective approach.

- Report Writing & Presentation:
  - Write final report including methodology, results, and conclusion.
  - Prepare slides for in-class presentation.

References

Maas, A. L., et al. (2011). "Learning Word Vectors for Sentiment Analysis."
https://ai.stanford.edu/~amaas/data/sentiment/