Introduction to Data Science — Assignment 7

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh**

TA(s): Bardia Khalafi, Javad Seraj

Deadline: Tuesday, Khordad 15th, 11:59 PM

Introduction

You have been employed at Amazon as a data scientist for its movie analysis department. In this assignment, you are given a dataset and asked to train a model to classify IMDb review comments automatically. First, you need to use different methods to expand your labeled data for training, extract features from sentences, and then train and evaluate your model.

Dataset

In the previous assignment, you gained experience handling unsupervised data. In real-world scenarios, the majority of data remains unlabeled, and labeling all of it can be prohibitively expensive. Instead, we label a portion of the data and attempt to extend these labels to other samples, constituting a semi-supervised task.

You are given a dataset containing information about movies and their user reviews on Amazon's movie platform.

Dataset Description

The dataset comprises movie reviews submitted by users on IMDb for sentiment analysis tasks. Sentiment in movie comments refers to the overall emotional tone or attitude expressed towards a movie in a comment or review, which could be positive or negative. Positive sentiment might indicate enjoyment, appreciation, or excitement about the movie, while negative sentiment could suggest disappointment, dislike, or criticism.

Format

This dataset is provided in a structured JSONL format. Each entry usually contains:

- The text of a review.
- Its corresponding sentiment label (1 for positive, 0 for negative).
- A feature vector (embedding) for each comment, which is used to train machine-learning models.

Dataset Segments

- Training dataset: Use this portion of the data to train your models.
- Test dataset: Use this portion of the data for validation.
- Augmentation dataset: This unlabeled portion of the data should be labeled using existing models (machine learning models or large language models) to augment your training dataset.

Bit of Advice

We advise you to create and run your notebooks on the Google Colab platform. It provides free GPU time, making your computations faster. Google Colab also handles all the necessary installations, so you don't have to worry about dependencies. Plus, you can easily share your work and access your notebooks from any device with an internet connection, making it a convenient tool for this assignment.

In this assignment, your report and analysis are more valuable than just the code, so make sure to do your research well.

Task

1. EDA

Before we start building models, we need to take a close look at the data. This means examining information about movies and user reviews on Amazon. We want to understand what people are saying about the movies and what features they talk about the most. This helps us decide how to work with the data to build accurate models later on. You should also research how you can do EDA, on text data.

2. Feature Engineering

Explore methods for extracting features from text data and provide a brief explanation of each. We have already added these embedding features to each sentence in the dataset to simplify the complexity of NLP tasks, allowing you to focus on other aspects of your work. It's up to you how you want to use these features along the way.

3. Semi-Supervised

You have learned that supervised learning involves training a model on a labeled dataset, where each example is paired with an output label, while unsupervised learning deals with unlabeled data and aims to find hidden patterns or structures. Semi-supervised learning is a hybrid approach that combines both methods. It uses a small amount of labeled data along with a large amount of unlabeled data to improve learning accuracy. This approach leverages the labeled data to guide the learning process and the unlabeled data to capture the underlying data distribution, making it particularly useful when labeling data is expensive or time-consuming.

Using Traditional Methods (Label Propagation)

Research label propagation techniques and their traditional methods. Utilize one of these methods, such as KMeans, to propagate labels for unlabeled data. Next, train a model using the extracted features and propagate labels as desired (whether using only hand-labeled data or a combination of labeled and propagated labels is up to you).

Using LLMs

For working with LLMs, we suggest you look at the documentation on the Hugging Face website. Hugging Face is a leading platform for natural language processing, offering high-level libraries like Transformers. The Transformers library provides pre-trained models and tools that simplify implementing and fine-tuning state-of-the-art language models for various NLP tasks, such as text classification, translation, and question-answering. This library makes it easy to leverage powerful models without needing deep expertise in machine learning, accelerating your development process. For more information, we suggest you check this tutorial on how to use LLMs.

Now, let's integrate an LLM (Large Language Model) for label generation. Research how an LLM can be applied to a specific task like classification and explain the circumstances under which each method should be used.

Beware that using an LLM can be time consuming, so make sure to save your results. Just like the previous step, train a model using the extracted features and new labeled dataset (you can even use the LLM as your classifier!).

For this task, you are going to use prompt engineering methods to generate labels with the Phi-3 model. We have already implemented a function to load the model for you, so you don't need to worry about that.

Evaluate both models on the test set and compare their results. For evaluation metrics, you can use the ones you have learned in class. To find the best model in each method, you'll need to explore different methods for labeling, models, datasets, features (including text features), and more. Take your time and use everything you've learned so far in this assignment 😉.

Questions

1. Research semi-supervised problems and explain how to select a proportion of data for manual labeling. Does the choice of data matter?
2. For label propagation, how many data points did you label using your manually labeled data? Explain the trade-off between quality and quantity of your labeled data.
3. Research the limitations of label propagation methods. Can these limitations be overcome using a large language model (LLM)? If yes, explain how and why using LLMs is more effective for these tasks.
4. Research the history of language modeling, their evolution, and explain how they work. Discuss the advantages and limitations of language models, especially large language models (LLMs), in various tasks.
5. For each method (LLM/Label-Propagation), explain its advantages and disadvantages.
6. Certainly, you have worked with large language models (LLMs) since they were made public, and you might even be using one right now to help you with your assignment! How has your experience been while interacting with one of them? How do other LLMs differ from each other? Name some of these LLMs that you have worked with and share your opinion about them.

Notes

- Upload your work as a zip file in this format on the website: DS_CA7_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.