Lab Course Machine Learning

Exercise Sheet 11

Prof. Dr. Dr. Lars Schmidt-Thieme
Kiran Madhusudhanan

HiWi: Basharat Mubashir Ahmed
Sophia Damilola Lawal

Submission deadline : February 5, 2023

General Instructions

1. Perform a data analysis, deal with missing values and any outliers.
2. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
3. Data should be normalized.
4. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use one-hot encoding.

1 Naive Bayes. (5 points)

In this part of the assignment, you will be using the processed dataset from previous Lab Question 1. Using the BOW and Tf-Idf representation, implement a Naive-Bayes classifier for the data. Use Laplace smoothing for the implementation. Compare your implemenation with the sklearn implementation.

2 N-gram Language Model. (10 points)

You won't believe what happened ??? !

Is the word "next" on the tip of your tongue? Although there are other possibilities, that is undoubtedly the most likely one. Other options are "after", "after that", and "to them". Our intuition tells us that some sentence endings are more plausible than others, especially when we take into account the previous information, the location of the phrase, and the speaker or author.

N-gram language models simply formalize that intuition. An n-gram model gives each possibility a probability score by solely taking into account the words that came before it. The probability of the word "next" in our example may be 80%, whereas the probabilities of the words "after" and "then" might be 10%, 5%, and 5%, respectively.

By leveraging these statistics, n-grams fuel the development of language models, which in turn contribute to an overall speech recognition system.

Task

In this assignment you are tasked with coding a N-gram language model on the dataset `https://www.kaggle.com/datasets/nltkdata/europarl`. Use the english language for the task.

Evaluate your model based on perplexity and generate sentences using n-grams with n=2,3,4,5.

Reading Material: `https://web.stanford.edu/~jurafsky/slp3/3.pdf`

3 Latent Dirichlet Allocation (5 points)

We'll utilize the dataset of papers from the NeurIPS (NIPS) conference `https://www.kaggle.com/datasets/benhamner/nips-papers`, one of the most esteemed annual gatherings in the machine learning field, for this assignment. The CSV data file includes details on the many NeurIPS articles that have been published between

1987 and 2016 (that's 29 years!). These articles cover a wide range of machine learning issues, including neural networks, optimization techniques, and many more.

You are free to use sklearn/gensim/nltk or any library for this task. Perform topic modelling using LDA and present a comprehensive analysis of the task. You can use word clouds and present analysis using the LDA visualization library pyLDAvis.

Points will be awarded based on the depth of results!