CSE 4392 Special Topic: Natural Language Processing

Homework 8 - Spring 2024

Due Date: Apr 2nd, 2024, 11:59 p.m. Central Standard Time

Welcome to this week's assignment for the Natural Language Processing (NLP) course focusing on Expectation Maximization (EM) algorithms. In this assignment, you will dive into the application of EM for part-of-speech (POS) tagging using the Wall Street Journal (WSJ) POS tagging dataset.

Problem 1 - 100%

Your task is to implement the Expectation Maximization algorithm for POS tagging using the WSJ POS tagging dataset. EM is a powerful iterative algorithm used to estimate parameters of probabilistic models when there are hidden variables involved. In the context of POS tagging, EM helps to infer the most likely POS tags for words in a corpus based on observed data.

You will be provided with the Wall Street Journal (WSJ) POS tagging dataset(**same as last week homework**), which contains a collection of annotated sentences where each word is tagged with its corresponding POS tag.

Instrustions:

1. **Data Preprocessing:** Begin by preprocessing the dataset to extract the necessary features, such as word frequencies and transition probabilities between POS tags.
2. **Initialization:** Initialize the parameters of the model, including initial state probabilities, transition probabilities, and emission probabilities.
3. **Expectation Step:** Use the forward-backward algorithm to compute the expected counts of the hidden states (POS tags) given the observed data (words).
4. **Maximization Step:** Update the model parameters based on the expected counts obtained in the previous step.
5. **Convergence:** Repeat the expectation-maximization steps until the model converges, i.e., the log-likelihood of the data no longer increases significantly.
6. **Evaluation:** Evaluate the performance of the model by comparing the predicted POS tags with the ground truth labels.

Deliverables:

1. Python code implementing the Expectation Maximization algorithm for POS tagging.
2. Report documenting your approach, including details of preprocessing, initialization, EM algorithm implementation, evaluation results, and analysis of the model's performance.