CS 292F.300 Final project proposal

Michael Saxon

TOTAL POINTS

1 / 1

QUESTION 1

1 Proposal submitted 1 / 1

✓ **- 0 pts** **Correct**

This sounds great. I look forward to seeing the results.

CS 292F Final Project Proposal:
Spectral Clustering for Bias Detection in Natural Language datasets
Michael Saxon and Xinyi Wang

1. Introduction

Spectral clustering is a clustering technique based on the idea of finding the 'best cut'/partition of a graph using the eigenvectors of the graph Laplacians. It can be easily extended to other forms of data beyond graph data, by regarding the data points as nodes in a graph and properly defining a similarity matrix/ adjacency matrix. By utilizing the graph structure, spectral clustering can detect complex cluster boundaries to better fit the data distribution.

In natural language processing, datasets for a lot of different tasks are known to be 'biased'. That is, there are unwanted artifacts in the observational data distribution that we do not want the machine learning models to learn. For example, natural language inference (NLI) is the task of predicting the relationship (entailment, neutral, contradiction) between a given premise and a given hypothesis. However, researchers have found that machine learning models can obtain non-trivial accuracy by only using the hypothesis to predict the relationship. This indicates that there are spurious patterns in the dataset that are highly correlated with the label. In this project, we hope to use the spectral clustering technique to try to detect these spurious patterns by analyzing the label distribution of each cluster: if there is a highly unbalanced distribution of the label in a cluster, then we would take a closer look at the examples in the cluster to see what is their common feature. In this case, this common feature (e.g. the existence of negation) would be the spurious pattern we want to detect. Another way of evaluating the soundness of the detected biased clusters would be trying to train a new classifier on the data without the biased clusters and then test its out-of-domain generalization performance.

2. Method

We consider building the graph structure for the text data using the following two methods: build a K-nearest neighbors (KNN) graph using the RoBERTa embedding of the sentences; or construct a similarity matrix using the BERTScore or other self-defined frequency based similarity scores. In the first method, we basically try to use KNN to construct a unweighted undirected graph by connecting each data point with its K nearest neighbor in the representation space. In the second method, we try to construct a weighted undirected graph with the similarity score as the edge weight between each pair of data points. Then we calculate the eigenvalues and eigenvectors of the Laplacians to find the possible partitions of the resulting graph.

We primarily consider using the spectral clustering method to analyze the NLI datasets (e,g, MNLI, SNLI), but we would definitely try to apply this analysis to other NLP tasks (e.g. Quora Question Pair matching) if we get promising results and if the time permits.

3. Outlook and Justification

As the first-year members of the NLP Lab, we have both spent time in the last year thinking about and working on the problem of biased datasets in NLI. Xinyi is submitting a manuscript to NeurIPS 2021 on using causal methods to ameliorate these bias patterns during training on NLI and image classification tasks, and Michael has a manuscript in progress on using traditional unsupervised clustering classifiers to find biased distributions of hypotheses in learned embedding spaces.

In addition to fitting with the topic of this course, this project intersects with a lot of research ideas we have had bouncing around already, and fits in with the shared elements of both of our research directions. We anticipate the outlook for including results from this work in a future publication at NLP venu

1 Proposal submitted 1 / 1

✓ **- 0 pts Correct**

This sounds great. I look forward to seeing the results.