6.867 Machine Learning Project Proposal

Rushi Ganmukhi

1 Introduction

For my final project in 6.867, I would like to explore semi-supervised learning methods for document classification. Currently for my research I am implementing a relevance feedback system for query generation. The idea is that an initial query is sent to a search engine(Google) and a set of documents is retrieved. These documents are labeled by a human as relevant or not to a query and then an algorithm is used to determine which words should be added to the query in order to produce documents with a higher chance of being relevant. The algorithms currently being used are [1] Rocchio's algorithm, maximum frequency and maximum tf-idf value. All three of these values look at the vector representation of the document and determine which word or words have the highest influence on whether the document is relevant or not. Labeling of the retrieved documents is costly, it takes a large amount of effort on the human user to perform. I would like to implement a semi-supervised method in which some only some of the documents are labeled by humans and we are tasked with producing the labels for the unlabeled documents.

For my project I would like to implement the spectral graph transduction algorithm for semi-supervised learning[2] . This model treats documents as nodes in a graph and edges as the cosine similarity between the word vector representation of the documents. The top k edges are selected for each node are used while the remainder are dropped. We then look to partition the graph by finding the minimum cut that partitions the graph into two sections relating the the positive and negatively labeled data and from here we can assign labels to the unlabeled documents. Solving this method consists of taking the eignvalue decomposition of the adjacency matrix and finding the optimal partition of the graph according to various constraints.

1.1 Evaluation Metrics

To evaluate these semi-supervised methods we can compare them to the k-nearest neighbor implementation on the graph to assign unlabeled data to the label that is maximum among its k-nearest neighbors. The data being used to test on will be documents generated from a set of queries on Google. Currently I have a small set of annotated data but this will need to be expanded upon. Additionally, I will integrate these semi-supervised methods into the relevance feedback model and evaluate their scores against the fully human annotated documents. The metric used here will be Mean Average Precision(MAP).

2 Timetable & Evaluation of Potential Risks

By November 8th - dataset completed
By November 15th - Spectral Graph transducer implemented
By November 22nd - KNN & comparisons done, extra time for run off
By November 29th - Integrate into relevance feedback system
By December 5th - Writeup

Potential risks stem from the fact that this method for semi-supervised learning is highly parameterized. Additionally we will only have access to around 25 documents for each run of the learning algorithm, possible making it difficult to tune in these parameters. The paper includes how they calculated the parameters and the intuition behind their decisions, which should help when I repeat the experiments.

References

[1] Okabe, Masayuki, Kyoji Umemura, and Seiji Yamada. ”Query expansion with the minimum user feedback by transductive learning.” Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005.

[2] Joachims, Thorsten. ”Transductive learning via spectral graph partitioning.” ICML. Vol. 3. 2003.

[3] Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty. ”Semi-supervised learning using gaussian fields and harmonic functions.” ICML. Vol. 3. 2003.