Implementation of Active Learning Methods on Poisson Learning Framework

Faculty Mentor: Professor Jeff Calder

Student researcher: Jason Setiadi

May 13, 2021

In this era, data is continuously evolving and being generated in various areas of our lives. Therefore, it is important for us to learn from it and turn it to useful insights to help us make better decisions. Machine learning is one field of computer science that gives machines access to data to process and turn it into knowledge. It learns and analyzes underlying patterns in the data as a basis for making future predictions. One application of machine learning from [1] was how banks calculate customers' risk given the amount of credit and their information, also known as credit scoring. Banks can utilize information such as income, savings, profession, age, past financial history, etc, to determine the customer's risk and use it to accept or reject his or her future applications. This is also an example of classification in machine learning where we could generate a rule to classify the customers into low or high-risk from the given data.

In classification problems, labels are very important since they are the output of a dataset that is used to make predictions, such as low or high-risk from the example above. However, in the real world, there are massive amounts of unlabeled data whereas labeled data is hard to obtain, time consuming, and expensive. Hence it is more practical to perform semi-supervised learning (SSL), which is an approach in machine learning that uses a large amount of unlabeled data with only a small set of labeled data to make predictions. Moreover, it is also important to choose the most informative set of labeled data so that we could achieve great accuracy while minimizing labeling cost. This is where it makes sense to combine SSL with active learning.

Active learning is a sub field of machine learning that allows the learning algorithm to choose a subset of data to label. The key idea behind this is, rather than choosing the data at random, there are many sampling strategies that can be used to get the most informative data that yields greater

accuracy. One method is to label the data that the machine learning model is least certain of how to label, which is called uncertainty sampling. It was shown in [6] that uncertainty sampling provides higher accuracy than random sampling in various amounts of data being sampled on a text classification problem. This UROP project will be focusing on implementing three different active learning methods on the new Poisson learning framework. Poisson learning is a new framework for graph based semi-supervised learning at very low label rates. This framework is motivated by the need to address the degeneracy of Laplacian learning, which is a widely used method for graph based semi-supervised learning, that does not perform well at very low label rates. Poisson learning solves the problems in Laplacian learning (mostly due to large constant bias in the Laplace equation solution) by replacing the assignment of labeled values at training points with placement of sources and sinks and solving the resulting Poisson equation on the graph. A numerical experiment from [2] showed that the Poisson learning framework has significantly higher accuracy rates compared to Laplacian learning and several other semi-supervised learning methods on MNIST, FashionMNIST, and Cifar-10 datasets at very low label rates.

There are three active learning methods that would be applied on the Poisson learning framework, which are Variance Minimization Criterion from [5], Error Bound Minimization from [4], and Sampling Theory for Graph Signals from [3]. The idea of Variance Minimization Criterion from [5] is to analyze the probability distribution of the unlabeled points conditioned on the label information, which is a multivariate normal with harmonic solution over the field as the mean. Then, we select the node to label such that the total variance of the distribution on the unlabeled data and the expected prediction error are minimized. Furthermore, the Error Bound Minimization from [4] tries to select a subset of nodes on a graph such that the empirical transductive Rademacher complexity of a graph-based learning method called learning with local and global consistency (LLGC) is minimized. Lastly, the Sampling Theory for Graph Signals from [3] identifies the class of graph signals that can be reconstructed from their values on a subset of vertices that allows us to define a criterion for active learning based on sampling set selection which aims at maximizing the frequency of the signals that can be reconstructed from their samples on the set.

Based on the papers [5] [4] [3], the three active learning methods have shown to improve accuracy rates by utilizing unlabeled points on graphs. We propose to apply these three acitve learning methods in the Poisson learning framework to see whether any further improvements can be achieved. The first part of this project will be focused on reading and understanding the background papers of the algorithms going to be used. Then we will implement the active learning methods on the Poisson learning framework through Python. Finally, we will compare and evaluate the accuracy rates of the three active learning methods with sampling at random and some other methods, implemented on the Poisson learning framework in MNIST, FashionMNIST, and Cifar-10 datasets. We will use the Python package GraphLearning https://github.com/jwcalder/GraphLearning, developed by Dr. Calder, which includes methods for solving the Poisson learning problem.

Through this UROP project, the implementation of active learning methods on the Poisson learning framework could produce an improved graph based semi-supervised learning algorithm that can attempt to solve modern machine learning problems, which are ubiquitous (diagnosing medical images, self-driving cars, robotic navigation, etc.). Improvements in active learning methods can also lessen the burden on experts by requiring fewer labels while still achieving high accuracy. Moreover, research in this area also have the potential to have broad societal impacts in many areas of science and engineering.

REFERENCES

[1] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2014.

[2] J. Calder, B. Cook, M. Thorpe, and D. Slepčev. Poisson Learning: Graph Based semi-supervised learning at very low label rates. *Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119*, 2020.

[3] A. Gadde, A. Anis, and A. Ortega. Active semi-supervised learning using sampling theory for graph signals. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 492–501, New York, NY, USA, 2014. Association for Computing Machinery.

[4] Q. Gu and J. Han. Towards active learning on graphs: An error bound minimization approach. In *2012 IEEE 12th International Conference on Data Mining*, pages 882–887, 2012.

[5] M. Ji and J. Han. A variance minimization criterion to active learning on graphs. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 556–564, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[6] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.