UROP Final Report

Jason Marcell Setiadi

14 December 2021

Firstly, I want to introduce some terminology about my UROP project to get a better understanding about it. Graphs are a unique structure of a dataset that keeps the relationship between each data point. For example, we can create a graph of research papers and keep the citations among each other as the edge/link of the graph to preserve the relationship between the papers. Training data is the data we use to train our machine learning model, and contains labels, which are the output/outcome of our prediction. For example if we have an image and we want to know whether it is a dog or not, the data would be the image and the label would be a yes or no whether it is a dog or not. Semi-supervised learning is a machine learning term that uses a small set of training data to predict the rest of data points in the dataset. So, graph-based semi-supervised learning framework is just a semi-supervised learning method for graphs.

In real life, it is often hard to get labeled data points especially in certain fields like medical or government. Therefore, it is important in semi-supervised learning to choose the most optimal set of training data to still get a high accuracy for our prediction with just a small amount of training data. This is where we introduce active learning to solve the labelling problem. Active learning is a field that tries to choose the most optimal set of data points as our training data to gain high accuracy but minimize the amount of points being used. The objective of the project is to apply active learning methods on graph-based semi-supervised learning frameworks, particularly Poisson Learning, which is a framework proposed by my faculty mentor, Dr. Calder. We will compare the active learning results with random sampling results (Figure 1) to determine how much improvement the active learning methods produce.

We research three different active learning methods namely, V-Opt, Bounds, and Signals for simplicity. We discovered that the Bounds method had mathematical derivation errors in the paper which leads to a similar method as V-Opt so we decided to not continue with it. We performed our tests by creating a toy dataset consisting of 8 gaussian clusters formed into a circle. Our instinct of choosing the most optimal points for this dataset would be to choose 1 point on each of the 8 clusters, preferably the centers, then start exploring to more ambiguous regions like the intersection between clusters. Therefore that is our expectation as to how the active learning methods would perform in this dataset. Our results show that V-Opt (Figure 2) and Signals (Figure 3) are indeed successful active learning methods for minimizing training data and choosing optimal points to label. Both of them performed as desired but maybe the signals method choose the outer points rather than the centers of the clusters.

We then calculate the accuracy after choosing each point to label using both Laplace and Poisson Learning. Although Laplace Learning works really well on this toy dataset, we found out that Poisson Learning doesn't work so well in this toy dataset (Figure 4 and some other datasets which is a problem we want to address in the future. Furthermore, the Signals method is not scalable to real-world datasets because it requires expensive computation to choose each point to label as such datasets contain a huge amount of data points. Nevertheless, we can still analyze our results using Laplace Learning (Figure 4) to see how the active learning methods compare with random sampling. As our future work, we would like to explore other active learning methods especially one that works well with Poisson Learning as well as address why Poisson Learning performs poorly in some datasets like our toy dataset.

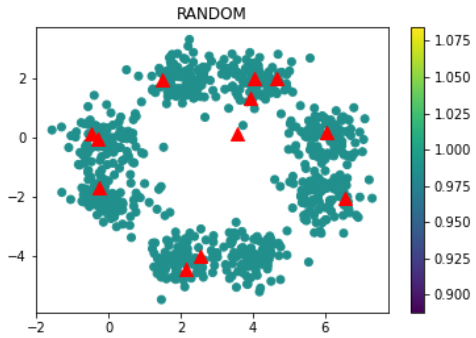


Figure 1: Random Sampling Result on Toy Dataset

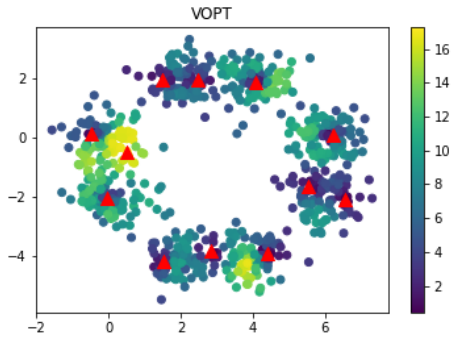


Figure 2: V-Opt Result on Toy Dataset

Overall, we can say that we have successfully implemented the active learning methods in Laplace Learning but there are still some problems using Poisson Learning which we hope to research more in the future. Active Learning certainly is a promising field to address the labelling problem and should be beneficial for experts in fields where labels can be hard or expensive to obtain. We certainly encourage more people to research on this topic since it is a relatively new but promising field.

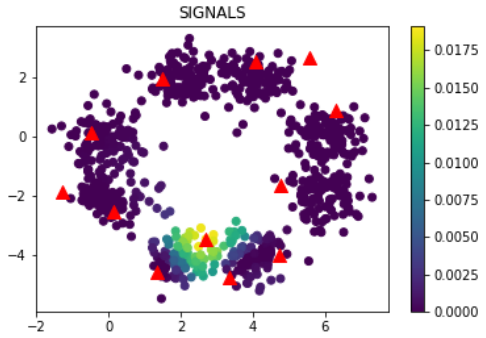
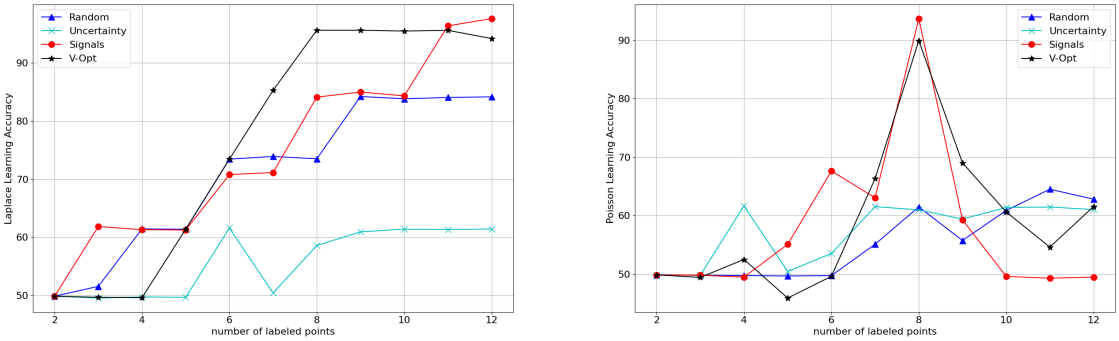


Figure 3: Signals Result on Toy Dataset



(a) Laplace Learning

(b) Poisson Learning

Figure 4: Toy Dataset Accuracy Results

I am really grateful to have the opportunity to work with my mentor in this UROP project. However, in my opinion, the time to do a UROP project each semester can be quite challenging especially while taking courses, which is why I applied again for next semester's UROP to continue research about this topic. UROP certainly is very helpful for students to gain experience especially in research. Personally, I learn a lot from the research I have done so far including reading research papers which is very overwhelming at first, learn to code complicated algorithms, write research papers, and improve my communication skills both written and orally. This program will be helpful for anyone who would like to gain their first real-world experience outside of coursework and serves as an alternative to doing internships.