Quantifying Uncertainty of V-Information for Data Valuation
(Technical Paper)

Sources of Bias in Machine Learning Models and Methods to Mitigate Them

(STS Paper)

A **Thesis Prospectus Submitted to the**

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Srinivasa Josyula
Spring, 2023

Technical Project Team Members
Hanjie Chen

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Srinivasa Josyula
Yangfeng Ji, Department of Computer Science
Richard Jacques, Department of Engineering and Society

Introduction

In the last ten years, the world has seen an explosion of machine learning applications. Machine learning is found everywhere from voice assistants and search engines to self-driving cars and everything in between. At the core of every machine learning algorithm is a plethora of data from which the model and then is able to make predictions on previously unseen data. Moreover, due to the increase in computing power over the years there has been a focus on creating larger models and feeding these models with ever larger datasets. With this increasing complexity, it is often hard to see how the model variables or dataset set information is used in the final prediction. To explore these issues, I chose my technical project to be in the general area of information theory and how it relates to machine learning. I am interested in pursuing this topic because it goes against the prevailing philosophy in machine learning that more data is better. Employing more useful data can lead to speed up in training and lead to computational savings. In computer science, information has a precise mathematical definition, but for my use case it can be thought of as the contribution of a data point to the machine learning model's effectiveness. Recently, a new metric was introduced to measure the information in a dataset when trained on a specific model family. This metric, coined V-information, is an extension of previous works based in information theory. V-information also extends to pointwise V-information which is the information gained from each data point. In my technical project, I will be taking a closer look at this new metric and evaluating its robustness in various scenarios.

This project will be useful to explore because it provides a framework for evaluating different datasets and their effectiveness over certain model families. Until recently, the focus has been on finding larger datasets, however there should also be concern about the quality of the data and whether the data is providing useful information to a machine learning model. Having a

reliable way to measure information in a dataset will give individuals more insight into how the model is arriving at its prediction. Over the years we have also seen bias creep into many machine learning models, and it has been hard to track which dataset or decision causes the bias. Using this metric could help us exclude certain data points that contribute to bias in the dataset and help the model make more accurate predictions.

Technical Project

Measuring information in any piece of data is an ambiguous concept. However, in the field of computer science and mathematics there is a precise definition of information, known as Shannon entropy. But, this definition relies on knowing the assumption that the distribution of data is not known. For example, it takes one bit of information to transmit the outcome of a coin flip since we know the exact distribution of the coin flip. This does not hold for most datasets, therefore making it challenging to quantify the amount of information that the model is able to extract from the dataset.

A paper from ICLR 2020, named “A Theory of Usable Information Under Computational Constraints” sought to solve this problem by introducing a new metric called v-information. This metric computes the information in a dataset over a certain model family, or in simple terms a group of models that are built to achieve similar tasks. This metric is computed by measuring the information gain of the model from a null dataset to the real dataset. This works because in theory the null dataset does not give the model any useful information and serves as a benchmark for comparison. In most applications, this is equivalent to randomly guessing the result. Then using an algorithm proposed in the paper, we can compare a real dataset to the benchmark model in order to estimate the information in that dataset. In this manner, we can compare different dataset and the information they provide to the model. But, for this metric to be widely used it

needs to provide reliable results and in line with human intuition. This reliability or robustness of the metric has not been previously studied, and I plan to explore this in my technical project. The original paper tested the metric on some datasets and models, however it was not a comprehensive evaluation of the reliability of the metric. In my technical project, I will use the metric in various scenarios to see how it behaves in various situations and attempt to formulate bounds for where this metric works. Moreover, I want to explore if the metric follows human intuition. For example, if a model is trained on duplicate data the information gain should be the same as training the model after deleting the duplicate data. This follows human intuition because repeat data does not provide any additional information.

In order to explore this metric further, I utilized another paper released in ICML 2022 called “Understanding dataset difficulty using v-usable information”. This paper further extended work from the previous work and defined a closely related metric called pointwise v-information which measures information over each datapoint. The dataset used in this paper is called the Stanford Natural Language Interference (SNLI) corpus. In this dataset, the model is trained on a hypothesis and premise and has to decide whether the premise contradicts, entails or is neutral based on the hypothesis. In the paper, they used various configurations of the hypothesis and premise to determine characteristics of the metric. I intend to use this dataset and test more configurations to understand the robustness of v-information.

By using the metric in various configurations, I hope to understand where this metric works well and where its shortcomings are. My final goal is to create a mathematical formula that gives a confidence interval for the v-information and pointwise v-information. If this is not feasible, I want to create a framework or boundary in which v-information is reliable so future researchers can accurately use this metric depending on their use case.

STS Project

As the field of machine learning has grown over the last decade, there has been a growing sense of concern over these models stemming from various issues. A big issue that has surfaced in the last few years has been that machine learning models have given results that discriminate against certain groups. For example, in some instances facial recognition software has not worked for people of color or voice assistants only recognizing certain accents of English. There are many ways in which this bias can creep into the models. In some instances, the developers of the model might be unaware of the users who will use this model which might lead them to create a flawed application. In other cases, the data that is used to train the model could be biased which would result in the model being biased. I am interested in this second issue because it relates to my technical project on information theory and how machine learning models extract information from data. By understanding The question I want to explore is how has data affected the way in which machine learning models behave, and how can we make sure that the data we input into the model is not biased?

Research Question and Methods

I will explore this topic through the Actor-Network Theory (ANT) framework. As part of this theory, the relationship between everything in a network including humans, technology and inanimate objects is explored. In my case, the technology includes the various machine learning models and the datasets used to train them. The inanimate objects are the devices through which humans interact with these models, such as through an app or website. There needs to be a proper understanding of all the moving parts in order to make sure that the users in this network feel properly represented. I will use the reading and synthesis method to understand this topic by reading various texts on the issue of machine learning bias and understanding what

leads to biased datasets. I will also seek to understand if models or datasets are deliberately made to be biased in order for certain groups to gain an advantage.

Conclusion

In the growing world of machine learning, the data used to train the models is equally as important as the models themselves. However, there is far less research in the area of understanding the usefulness of datasets and how useful various data points are to the final model. In recent years, there has been new research in the field and I have chosen to focus on a metric introduced in 2020 called v-information. This metric extends the theoretical definition of information into the field of machine learning and aims to quantify information in datasets trained over a class of machine learning models. This metric, along with pointwise v-information which was introduced in a later paper have proven to be useful but their robustness or reliability has not been fully researched. In my technical project, I will do further research on this metric and determine an interval or bound for the values of this metric so it can be used more confidently. Additionally, in my STS project I will explore how biased datasets can have an impact on machine learning models, and how we can seek to better understand what makes these datasets biased in the first place. I will look at which groups might seek to benefit from these biased datasets and understand their motivations.