EMPIRICAL METHODS FOR MINIMIZING STRUCTURAL RISK

SUMIT SOMAN

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

EMPIRICAL METHODS FOR MINIMIZING STRUCTURAL RISK

by

SUMIT SOMAN

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

Certificate

This is to certify that the thesis entitled **"Empirical Methods for Minimizing Structural Risk"**, being submitted by **Sumit Soman** for the award of the degree of **Doctor of Philosophy** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, is a record of bonafide work done by him under my supervision and guidance. The matter embodied in this thesis has not been submitted to any other University or Institute for the award of any other degree or diploma.

Dr. Jayadeva

Professor

Department of Electrical Engineering,

Indian Institute of Technology Delhi,

Hauz Khas, New Delhi - 110016,

INDIA.

Acknowledgments

I would like to thank my supervisor Prof. Jayadeva, Research Committee members Professors Santanu Chaudhary, Amit Kumar and Shouri Chatterjee; faculty members with whom I worked and published - Professors Suresh Chandra, Amit Bhaya and Manan Suri; friends and colleagues Aashish Rajiv, Siddharth Srivastava, Prashant Gupta, Aashi Jindal, Himanshu Pant, Mayank Sharma, Udit Kumar, Pawas Gupta, Munish Jain, Shruti Sharma, P Govind Raj, Soumya Sen Gupta; department staff members Rakesh, Yatindra, Mukesh and Ritwick; my parents and brother for supporting me through this journey.

(Sumit Soman)

Abstract

The performance of a machine learning model on test data, or its generalization ability, is formally measured in terms of the probability of its making an erroneous decision, also termed as the risk. Risk is composed of the empirical risk, which measures how well the model has learnt from training data, and the structural risk, which measures the risk due to model selection. The structural risk is a function of the training set size and the Vapnik-Chervonenkis (VC) dimension, a concept introduced by Vladimir Vapnik and Alexey Chervonenkis. While there is some work on estimating the VC dimension of learning models, it is a difficult and not very tractable problem. Most results, therefore, focus on bounds. The difficulty these present is that they do not permit model complexity to be optimized in terms of the variables being minimized. The practitioner is usually looking for the simplest model that can explain the available data. Bounds on the VC dimension, on the other hand, typically measure the range of functions that can be represented by a particular kind of model. In a nutshell, minimizing the structural risk appears to be an ill-posed problem.

Two approaches can be pursued for minimizing the structural risk. These are developing learning models with a small VC dimension, and methods to increase the number of samples available for training a model, by augmenting the training dataset.

The recently proposed Minimal Complexity Machine (MCM) has shown that it is possible to develop a hyperplane classifier that can minimize an approximate bound on the VC dimension. We initially develop a hybrid model of the MCM within the Extreme Learning Machine (ELM) framework that offers the benefits of scalability and low model complexity.

Next, we shift focus to developing a variant of the MCM that minimizes a tighter bound on the VC dimension. This approach is termed as the Quadratic MCM (QMCM). In the sequel, we show that the QMCM minimizes a tighter bound on the VC dimension, in comparison to the MCM. Extensions of the QMCM for learning settings other than classification, such as regression and feature selection, are also presented. A scalable version of the QMCM for large datasets via the stochastic sub-gradient minimization

approach is also described. Another approach for developing classifiers for learning when few training data are available, by introducing constraints based on computing the derivative of the discriminant function, is discussed subsequently. This approach is developed for both the Support Vector Machine (called SVM-Derivative) and the MCM (MCM-Derivative).

The last part of the thesis focuses on methods for risk minimization by augmenting the training dataset. We propose *EigenSample*, a novel method for generating additional samples for training a classifier. The new samples are added in a lower-dimensional space spanning the principal components of the original dataset, and a novel inverse projection problem is formulated and solved. This tends to least perturb the eigenstructure of the original dataset. *EigenSample* is evaluated on various datasets using multiple classifiers and results show improved generalization when the training samples are augmented using *EigenSample*, compared to competing methods for augmenting datasets. Extensions of *EigenSample* for regression datasets, as well as a non-iterative least-squares variant are also discussed.

सार

परीक्षण डेटा पर मशीन सीखने के मॉडल का प्रदर्शन, या इसकी सामान्यीकरण क्षमता, औपचारिक रूप से एक गलत निर्णय लेने की संभावना के संदर्भ में मापा जाता है, जिसे जोखिम भी कहा जाता है। जोखिम प्रयोगसिद्ध जोखिम से बना है, जो यह मापता है कि प्रशिक्षण डेटा से मॉडल ने कितना अच्छा सीखा है, और संरचनात्मक जोखिम, जो मॉडल चयन के कारण जोखिम को मापता है। संरचनात्मक जोखिम प्रशिक्षण सेट आकार और वेपनिक-चेरोवेनेकिस (वीसी) आयाम का एक कार्य है, जो कि व्लादिमीर वाप्निक और एलेक्सी चेरोनेंकिस द्वारा शुरू की गई अवधारणा है। जबकि सीखने के मॉडल के वीसी आयाम का आकलन करने पर कुछ काम है, यह एक कठिन और बहुत ही समस्याजनक समस्या नहीं है। अधिकांश परिणाम, इसलिए, सीमा पर ध्यान केंद्रित करते हैं। वर्तमान में जो कठिनाई है, वह यह है कि वे मॉडल जटिलता को वैरिएबल के रूप में अनुकूलित करने की अनुमति नहीं देते हैं। चिकित्सक आमतौर पर सबसे सरल मॉडल की तलाश में है जो उपलब्ध आंकड़ों की व्याख्या कर सकता है। दूसरी ओर, वीसी आयाम पर सीमाएं, आमतौर पर उन कार्यों की श्रेणी को मापती हैं जिन्हें एक विशेष प्रकार के मॉडल द्वारा दर्शाया जा सकता है। संक्षेप में, संरचनात्मक जोखिम को कम करना सही प्रकार से परिभाषित प्रश्न नहीं है ।

संरचनात्मक जोखिम को कम करने के लिए दो दृष्टिकोण अपनाए जा सकते हैं। एक छोटे वीसी आयाम के साथ सीखने के मॉडल विकसित कर सकते हैं , और प्रशिक्षण डाटासेट को बढ़ाने के द्वारा एक मॉडल को प्रशिक्षित करने के लिए उपलब्ध नमूनों की संख्या बढ़ा सकते हैं ।

हाल ही में प्रस्तावित मिनिमल कॉम्प्लेक्सिटी मशीन (एमसीएम) ने दिखाया है कि हाइपरप्लेन क्लासिफायर को विकसित करना संभव है जो वीसी आयाम पर एक अनुमानित सीमा को कम कर सकता है। हम शुरू में चरम सीखने की मशीन (ईएलएम) ढांचे के भीतर एमसीएम का एक हाइब्रिड मॉडल विकसित करते हैं जो स्केलेबिलिटी और कम मॉडल जटिलता का लाभ प्रदान करता है।

अगला, हम एमसीएम के एक संस्करण को विकसित करने पर ध्यान केंद्रित करते हैं जो वीसी आयाम पर बंधे हुए एक तंग को कम करता है। इस दृष्टिकोण को द्विघात MCM (QMCM) कहा जाता है। अगली कड़ी में, हम दिखाते हैं कि QMCM, MCM की तुलना में VC आयाम पर बंधे हुए एक टीयर को छोटा करता है। वर्गीकरण के अलावा सीखने की सेटिंग्स के लिए QMCM के विस्तार, जैसे प्रतिगमन और सुविधा चयन भी प्रस्तुत किए जाते हैं। स्टोचैस्टिक सब-ग्रेडिएंट मिनिमाइजेशन दृष्टिकोण के माध्यम से बड़े डेटासेट के लिए QMCM का एक स्केलेबल संस्करण भी वर्णित है। कम प्रशिक्षण डेटा उपलब्ध होने पर सीखने के लिए सहपाठियों के विकास के लिए एक और दृष्टिकोण, विभेदक फ़ंक्शन के व्युत्पन्न की गणना के आधार पर बाधाओं का परिचय देकर, बाद में चर्चा की जाती है। यह दृष्टिकोण सपोर्ट वेक्टर मशीन (जिसे SVM-Derivative कहा जाता है) और MCM (MCM-Derivative) के लिए विकसित किया गया है।

थीसिस का अंतिम भाग प्रशिक्षण डाटासेट को बढ़ाने के द्वारा जोखिम को कम करने के तरीकों पर केंद्रित है। हम EigenSample का प्रस्ताव करते हैं, एक क्लासिफायरियर के प्रशिक्षण के लिए अतिरिक्त नमूने उत्पन्न करने के लिए एक उपन्यास विधि। नए नमूने मूल डेटासेट के प्रमुख घटकों के बीच एक कम-आयामी स्थान में जोड़े जाते हैं, और एक उपन्यास उलटा प्रक्षेपण समस्या तैयार और हल की जाती है। यह मूल डेटासेट के स्वदेशीकरण को कम से कम प्रभावित करता है। EigenSample का मूल्यांकन कई डेटासेट पर किया जाता है, जिसमें कई क्लासिफायर का उपयोग किया जाता है और परिणाम बेहतर सामान्यीकरण दिखाते हैं, जब डेटा नमूने बढ़ाने के लिए प्रतिस्पर्धा के तरीकों की तुलना में EigenSample का उपयोग करके प्रशिक्षण नमूने संवर्धित किए जाते हैं। प्रतिगमन डेटासेट के लिए EigenSample के विस्तार, साथ ही एक गैर-पुनरावृत्त न्यूनतम-वर्ग संस्करण पर भी चर्चा की गई है।

Contents

List of Figures

List of Tables