STATISTICAL LEARNING ALGORITHMS: MULTI-CLASS CLASSIFICATION AND REGRESSION WITH NON-I.I.D. SAMPLING

PAN ZHIWEI

DOCTOR OF PHILOSOPHY

CITY UNIVERSITY OF HONG KONG

JUNE 2009

CITY UNIVERSITY OF HONG KONG

香港城市大學

Statistical Learning Algorithms: Multi-class Classification and Regression with Non-i.i.d. Sampling

統計學習算法：多分類及非獨立同分佈抽樣下的回歸

Submitted to
Department of Mathematics

數學系

in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

哲學博士學位

by

PAN Zhiwei
潘志偉

June 2009

二零零九年六月

Abstract

Learning theory is an inter-disciplinary research field involving applied mathematics, statistics, computer science, computational biology and data mining. It aims at learning function features (such as function value and variables) or data structures from samples by learning algorithms. Main research topics include designing efficient algorithms for various purposes and theoretical analysis of learning methods.

In this thesis, we consider two research problems. The first is to propose a new learning algorithm for multi-class classification by Parzen windows and to conduct both theoretical understanding and application for this algorithm. This Parzen windows classifier is better than the usual way of designing multi-class classifiers by combining binary classifiers in various ways, which is often complex and has the problems of overlapping. We give the convergence rates of the excess misclassification error, under some regularity conditions on the conditional probability distributions and some decay conditions on the marginal distribution near the boundary of the input space.

In the literature of Parzen windows for density estimation and regression, the approximation error is estimated locally at points which are in the interior of the input space X away from the boundary. Our key contribution for the mathematical analysis is to show how the decay of marginal distributions near the boundary yields satisfactory bounds for errors in terms of L^1 or $C(X)$ norms taken globally on the whole input space.

The second research problem considered in this thesis is the study of learning algorithms with non-i.i.d. sampling. The algorithms include least square regularized regression and binary classification. In the last few years, there have been significant developments in theoretical understanding of learning algorithms with i.i.d. sampling.

But either independence or identical sampling is a rather restrictive assumption in real data analysis, such as Shannon sampling, randomized sampling and weakly dependent sampling. Our setting does not require independence or identity. Under the conditions that the sequence of marginal distributions for sampling converges exponentially fast in the dual of a Hölder space and the sampling process satisfies a polynomial strong mixing condition, we derive capacity indepedent learning rates. Our convergence rate is consistent to that of the i.i.d. setting when the mixing condition parameter tends to zero.

For a binary classification learning algorithm with non-identical sampling, we also derive satisfactory capacity dependent estimates for the excess misclassification error.

Contents

Abstract **i**

Acknowledgement **iii**

1 Introduction **1**

1.1 Binary Classification Problems 1

1.2 Empirical Risk Minimization Principle 4

1.3 Strictly Separable Case by Hyperplanes 8

1.4 Classical Support Vector Machine and Kernels 12

1.5 Regularization Schemes in Reproducing Kernel Hilbert Spaces 14

1.6 Outline of this Thesis 15

2 Some Recent Non-traditional Developments in Learning Theory **17**

2.1 Online Learning 17

2.2 Semi-supervised Learning 18

2.3 Multi-task Learning 19

2.4 Main Contributions of this Thesis 20

3 Parzen Windows for Multi-class Classification **22**

3.1 A New Setting for Multi-class Classification 22

3.2 A New Comparison Theorem for Multi-class Classification 24

3.3 Basic Window Functions and Parzen Windows 26

3.4 Error Analysis and Learning Rates 28

3.4.1 Basic Assumptions and Learning Rates 28

3.4.2 Sample Error Estimate 31
3.4.3 Approximation Error Estimate 33
3.5 Conclusions 37

4 Learning with Non-i.i.d. Sampling 39
4.1 Least-square Regularized Regression Algorithm 39
4.2 Setting for Non-i.i.d. Sampling 40
4.2.1 Setting for Dependent Sampling 41
4.2.2 Setting for Non-identical Sampling 42
4.3 Basic Assumptions and Learning Rates 42
4.4 Error Analysis 44
4.4.1 Error Decomposition and Approximation Error Estimate . . . 45
4.4.2 Drift Error Estimate 46
4.4.3 Sample Error Estimate 47
4.4.4 Deriving Learning Rates 52
4.5 Learning from Non-identical Sampling for Classification 55
4.6 Conclusions 57

5 Further Discussion 59
5.1 Manifold Learning 59
5.2 Window Width 60
5.3 Dimension Reduction 61
5.3.1 Graph-Laplacian 62
5.3.2 Local ISOMAP 63
5.3.3 Local Linear Embedding 64

Bibliography 65

Publications 76