

---

**МОНОГРАФИИ НГТУ**

---

**Т.А. ГУЛЬТЯЕВА, А.А. ПОПОВ,  
А.С. САУТИН**

**МЕТОДЫ  
СТАТИСТИЧЕСКОГО ОБУЧЕНИЯ  
В ЗАДАЧАХ РЕГРЕССИИ  
И КЛАССИФИКАЦИИ**



**Т. А. ГУЛЬТЯЕВА, А. А. ПОПОВ,  
А. С. САУТИН**

**МЕТОДЫ  
СТАТИСТИЧЕСКОГО ОБУЧЕНИЯ  
В ЗАДАЧАХ РЕГРЕССИИ  
И КЛАССИФИКАЦИИ**



**НОВОСИБИРСК  
2 0 1 6**

УДК 519.217+519.233.5

Г 944

Рецензенты:

д-р техн. наук, доцент *В.Б. Бериков*  
д-р техн. наук, профессор *Б.Ю. Лемешко*

**Гульятеева Т.А.**

Г 944 Методы статистического обучения в задачах регрессии и классификации: монография / Т.А. Гульятеева, А.А. Попов, А.С. Саутин. – Новосибирск : Изд-во НГТУ, 2016. – 323 с. (серия «Монографии НГТУ»).

ISBN 978-5-7782-2817-7

В монографии рассматриваются вопросы статистического обучения в задачах построения регрессии по методу опорных векторов и в задачах классификации с использованием скрытых марковских моделей (СММ). Для решения задачи устойчивого оценивания модели регрессии по методу опорных векторов (SVM) в условиях зашумленных данных с помехой, имеющей распределение с большим эксцессом или имеющей характер асимметричного засорения, предложено использовать адаптивные и асимметричные функции потерь. Приводятся формулировки двойственных задач квадратичного программирования для этих случаев. Описывается метод квантильной регрессии на основе SVM для произвольной функции потерь. На его основе рассмотрен метод построения доверительных интервалов для отклика, а также непараметрический метод оценки неизвестной дисперсии ошибок наблюдений. Для построения компактной модели регрессии в условиях работы с выборками большого объема предлагаются алгоритмы построения разреженных решений в SVM. Показывается их эффективность в сравнении с классическим методом построения разреженных решений на основе функции нечувствительности Вапника. Описывается модификация SVM, позволяющая строить разреженные решения в условиях гетероскедастичности ошибок наблюдений. Приводятся результаты экспериментальных исследований по построению регрессионных моделей с использованием SVM при мультиколлинеарности данных, автокорреляции и гетероскедастичности ошибок наблюдений.

Приводятся результаты исследования подхода к решению задачи классификации наблюдаемых последовательностей, представленных скрытыми марковскими моделями, с использованием инициализированных этими моделями признаков. С использованием метода статистического моделирования рассматривается поведение нескольких классификаторов, когда наблюдаемые последовательности подвергались искажению действием на них различных помех. Также проанализированы случаи, когда нарушены одни из априорных представлений либо о наблюдаемых последовательностях, либо о структуре скрытых марковских моделей, описывающих эти последовательности.

Книга будет полезна научным сотрудникам и специалистам, сталкивающимся в своей деятельности с необходимостью решения задач построения зависимостей и классификации последовательностей, а также студентам и аспирантам.

УДК 519.217+519.233.5

ISBN 978-5-7782-2817-7

© Гульятеева Т.А., Попов А.А.,  
Саутин А.С., 2016

© Новосибирский государственный  
технический университет, 2016

**T. A. GULTYAEVA, A. A. POPOV,  
A. S. SAUTIN**

**METHODS OF STATISTICAL LEARNING  
IN SOLVING REGRESSION  
AND CLASSIFICATION PROBLEMS**



**NOVOSIBIRSK  
2 0 1 6**

УДК 519.217+519.233.5

Г 944

Reviewers:

Prof *V.B. Berikov*, D. Sc. (Eng.),  
Prof *B.Yu. Lemeshko*, D. Sc. (Eng.)

**Gulyaeva T.A.**

Г 944 Methods of statistical learning in solving regression and classification problems: Monograph / T.A. Gulyaeva, A.A. Popov, A.S. Sautin. – Novosibirsk: NSTU Publisher, 2016. – 323 pp. ("NSTU Monographs" series).

ISBN 978-5-7782-2817-7

The book deals with statistical learning in the construction of Models of regression and support vector classification problems with the use of Hidden Markov Models (HMM). To solve the problem of estimating the regression model by using the sustainable Support Vector Machines (SVM) technology under noisy data with an interference having a distribution with large kurtosis or having an asymmetric clogging character, the use of adaptive and asymmetric loss functions is proposed. The formulations of the dual quadratic programming problems for these cases are given. The quantile regression method based on SVM for arbitrary loss functions is described. Based on it the method of constructing confidence intervals for the response as well as a non-parametric method for estimating the unknown variance of observation errors are considered. To construct a compact regression model to deal with a large volume of samples the algorithms for constructing sparse solutions in SVM are proposed. Their effectiveness in comparison with the classical method of constructing sparse solutions based on the function of the Vapnik insensitivity is shown. The modification of SVM which gives possibilities to build sparse solutions under heteroscedasticity of observation errors described. The results of experimental studies on the construction of regression models using SVM with data multicollinearity, autocorrelation and heteroscedasticity of observation errors are given.

The results of studying the approach to solving the problem of classification of the observed sequences represented by HMM using features initiated by these models are presented. Using the method of statistical modeling the behavior of classifiers is considered when observed sequences are subject to distortion caused by various disturbances. Cases are also analyzed when some of the a priori representations of either any observed sequence or the structure of the HMM describing these sequences are violated.

The book will be of interest to researchers and specialists who have to solve problems of constructing dependences and sequence classifications as well as to undergraduate, graduate and postgraduate students.

УДК 519.217+519.233.5

ISBN 978-5-7782-2817-7

© T.A. Gulyaeva, A.A. Popov,  
A.S. Sautin, 2016  
© Novosibirsk State  
Technical University, 2016

## ОГЛАВЛЕНИЕ

Предисловие .....	7
Глава 1. СТАТИСТИЧЕСКОЕ ОБУЧЕНИЕ ПРИ ПОСТРОЕНИИ РЕГРЕССИИ МЕТОДОМ ОПОРНЫХ ВЕКТОРОВ .....	13
1.1. Основы теории машинного обучения .....	14
1.2. Алгоритм опорных векторов как метод построения непараметрической регрессии .....	17
1.2.1. Алгоритм опорных векторов .....	17
1.2.2. Двойственная задача .....	19
1.2.3. Вычисление параметра смещения $b$ .....	21
1.2.4. Разреженность решения .....	22
1.3. Нелинейная регрессия на основе SVM .....	23
1.4. Обзор подходов к решению оптимизационной задачи .....	26
Глава 2. ПОСТРОЕНИЕ РОБАСТНЫХ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОПОРНЫХ ВЕКТОРОВ С АДАПТИВНЫМИ ФУНКЦИЯМИ ПОТЕРЬ .....	29
2.1. Построение квазиоптимальных функций потерь .....	30
2.1.1. Функции потерь .....	30
2.1.2. Функционал риска .....	32
2.1.3. Метод максимального правдоподобия и модели плотностей .....	33
2.2. Робастные функции потерь в условиях асимметричных засорений .....	37
2.3. Конструирование двойственной задачи SVM .....	42
2.3.1. Двойственная задача для классических функций потерь .....	42
2.3.2. Двойственная задача для адаптивных функций потерь .....	47
2.3.3. Решение двойственной задачи с динамическими ограничениями .....	50
2.4. Исследования .....	52



Глава 3. ПОСТРОЕНИЕ РОБАСТНЫХ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОПОРНЫХ ВЕКТОРОВ С АСИММЕТРИЧНЫМИ ФУНКЦИЯМИ ПОТЕРЬ .....	63
3.1. Скошенные распределения и их моделирование .....	63
3.2. Конструирование квазиоптимальных функций потерь на основе линейно-квадратичных аппроксимаций для использования их в SVM .....	65
3.3. Конструирование двойственной задачи для случая асимметричных функций потерь в SVM .....	67
3.4. Исследования .....	71
3.5. Оценка параметра скошенности распределения .....	75
3.6. Квантильная регрессия на основе SVM .....	79
3.7. Построение доверительных интервалов для отклика .....	85
3.8. Оценка неизвестной дисперсии ошибок наблюдений .....	86
Глава 4. ПОСТРОЕНИЕ РАЗРЕЖЕННЫХ РЕШЕНИЙ .....	91
4.1. Задача построения компактной модели регрессии .....	92
4.2. Функция $\epsilon$ -нечувствительности Вапника и разреженные решения .....	93
4.3. Использование адаптивных функций потерь для получения разреженных решений .....	96
4.4. Метод «решето» Лапласа .....	98
4.5. Двухшаговый метод аппроксимации .....	101
4.6. Разреженность в условиях гетероскедастичности ошибок наблюдений .....	103
4.7. Исследования .....	106
Глава 5. ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ МАШИНЫ ОПОРНЫХ ВЕКТОРОВ В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ .....	111
5.1. Построение параметрических моделей на основе SVM .....	111
5.2. Построение полупараметрических моделей на основе SVM .....	114
5.3. Построение регрессии в условиях гетероскедастичности ошибок наблюдений .....	116
5.4. Построение регрессии в условиях мультиколлинеарности данных .....	121
5.5. Построение регрессии в условиях автокорреляции ошибок наблюдений .....	123
5.6. Выбор параметров алгоритма SVM .....	129
5.7. Применение метода SVM в прикладных задачах .....	135



5.7.1. Анализ выборки LIDAR .....	135
5.7.2. Анализ выборки «Motorcycle» .....	138
5.7.3. Анализ выборки «Boston Housing» .....	139
<b>Глава 6. СТАТИСТИЧЕСКОЕ ОБУЧЕНИЕ В ЗАДАЧАХ КЛАССИФИКАЦИИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ.....</b>	<b>141</b>
6.1. Введение в теорию СММ .....	141
6.2. Обучение СММ .....	145
6.2.1. Алгоритм прямого-обратного прохода .....	148
6.2.2. Алгоритм Баума–Велша .....	151
6.2.3. Масштабирование .....	156
6.3. Моделирование последовательностей.....	160
6.4. Классификация последовательностей .....	163
6.5. Историческая справка.....	165
<b>Глава 7. ПОСТРОЕНИЕ КЛАССИФИКАТОРОВ В ПРОСТРАНСТВЕ ПЕРВЫХ ПРОИЗВОДНЫХ ОТ ФУНКЦИИ ПРАВДОПОДОБИЯ .....</b>	<b>171</b>
7.1. Постановка задачи классификации последовательностей .....	171
7.2. Классификаторы.....	174
7.2.1. Метод ближайших соседей .....	174
7.2.2. Машины опорных векторов .....	175
7.2.3. Многоклассовая классификация.....	183
7.3. Вычисление первых производных от логарифма функции правдоподобия для СММ .....	184
7.4. Исследование возможности проведения классификации в пространстве первых производных .....	189
<b>Глава 8. КЛАССИФИКАЦИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В УСЛОВИЯХ ДЕЙСТВИЯ ПОМЕХ РАЗЛИЧНОЙ ИНТЕНСИВНОСТИ И ПРИРОДЫ.....</b>	<b>199</b>
8.1. Используемые модели помех, искажающие наблюдаемые последовательности .....	199
8.2. Исследования для моделей, в которых вероятности появления наблюдаемых символов описываются одним распределением .....	201
8.2.1. Аддитивная помеха с постоянными параметрами распределения.....	202
8.2.2. Аддитивная помеха с изменяемыми параметрами распределения .....	206
8.2.3. Замещающая вероятностная помеха с постоянными параметрами распределения .....	210





8.2.4. Замещающая вероятностная помеха с изменяемыми параметрами распределения .....	212
8.3. Исследования для моделей, в которых вероятности появления наблюдаемых символов описываются смесями распределений .....	215
8.3.1. Аддитивная помеха с постоянными параметрами распределения .....	216
8.3.2. Аддитивная помеха с изменяемыми параметрами .....	223
8.3.3. Замещающая вероятностная помеха с постоянными параметрами распределения .....	228
8.3.4. Замещающая вероятностная помеха с изменяемыми параметрами распределения .....	232
Глава 9. ПОВЕДЕНИЕ КЛАССИФИКАТОРОВ В УСЛОВИЯХ ОТКЛОНЕНИЯ ОТ ПРЕДПОЛОЖЕНИЙ .....	237
9.1. Различные законы распределений появления наблюдаемых символов, отличные от нормального .....	237
9.2. Процесс или объект, порождающий наблюдаемые последовательности и не имеющий скрытых состояний .....	246
9.3. Поведение классификаторов в условиях структурной неопределенности .....	252
9.3.1. Исследование на двухклассовой задаче классификации .....	252
9.3.2. Исследование на многоклассовой задаче классификации .....	260
Послесловие .....	265
Библиографический список .....	268
Приложения .....	293
Приложение А .....	293
Приложение Б .....	296
Приложение В .....	299

## CONTENTS

Preface .....	7
Chapter 1. STATISTICAL LEARNING IN CONSTRUCTING THE REGRESSION SUPPORT VECTOR MACHINES.....	13
1.1. Fundamentals of machine learning.....*	14
1.2. A support vector machine algorithm as a method for constructing a nonparametric regression .....	17
1.2.1. Support vector machines .....	17
1.2.2. The dual problem .....	19
1.2.3. Calculation of the parameter $b$ .....	21
1.2.4. The sparseness of the solution.....	22
1.3. A nonlinear regression based on SVM.....	23
1.4. Overview of approaches to the solution of the optimization problem.....	26
Chapter 2. CONSTRUCTION OF A ROBUST REGRESSION USING SUPPORT VECTOR MACHINES WITH ADAPTIVE LOSS FUNCTIONS.....	29
2.1. Construction of a quasi-optimal loss function.....	30
2.1.1. Loss functions .....	30
2.1.2. Risk functional .....	32
2.1.3. The maximum likelihood method and density models.....	33
2.2. A robust loss function under asymmetric noises .....	37
2.3. The design of the SVM dual optimization problem .....	42
2.3.1. The dual problem for classical loss functions .....	42
2.3.2. The dual problem for adaptive loss functions .....	47
2.3.3. The solution of the dual optimization problem with dynamic constraints.....	50
2.4. Research .....	52



---

Chapter 3. CONSTRUCTION OF A ROBUST REGRESSION USING SUPPORT VECTOR MACHINES WITH ASYMMETRIC LOSS FUNCTIONS .....	63
3.1. Skewed distributions and their modeling .....	63
3.2. Construction of quasi-optimal loss functions based on linear-quadratic approximations for use in SVM.....	65
3.3. The construction of the dual problem for the case of symmetric loss functions in SVM .....	67
3.4. Research .....	71
3.5. Parameter estimation of distribution skewness .....	75
3.6. Quantile regressions based on SVM .....	79
3.7. Confidence intervals for the response .....	85
3.8. The estimation of the unknown variance of the observation errors.....	86
Chapter 4. BUILDING SPARSITY SOLUTIONS.....	91
4.1. The task of building a compact model regression .....	92
4.2. The function $\varepsilon$ -insensitivity of Vapnik and sparse solutions.....	93
4.3. The use of loss functions to obtain sparse solutions.....	96
4.4. The Laplace sieve method.....	98
4.5. A two-step approximation method .....	101
4.6. Sparseness under heteroscedasticity of observation errors.....	103
4.7. Research.....	106
Chapter 5. EXAMPLES OF THE USE OF SUPPORT VECTOR MACHINES TO BUILD DEPENDENCIES .....	111
5.1. Construction of parametric models based on SVM.....	111
5.2. Building semi-parametric models based on SVM.....	114
5.3. Building a regression under observation error heteroscedasticity .....	116
5.4. Building a regression under data multicollinearity.....	121
5.5. Building a regression under observation error autocorrelation .....	123
5.6. Selection of SVM algorithm parameters .....	129
5.7. The use of the SVM method in applied problems .....	135
5.7.1. Analysis of the LIDAR sample .....	135
5.7.2. Analysis of the Motorcycle sample.....	138
5.7.3. Analysis of the Boston Housing sample.....	139



---

Chapter 6. STATISTICAL LEARNING IN SEQUENCE CLASSIFICATION PROBLEMS USING HIDDEN MARKOV MODELS .....	141
6.1. Introduction to the theory of HMM.....	141
6.2. Training HMM .....	145
6.2.1. The forward-backward algorithm.....	148
6.2.2. The Baum–Welch algorithm .....	151
6.2.3. Scaling .....	156
6.3. Sequence modeling .....	160
6.4. Classification of sequences .....	163
6.5. Historical background .....	165
Chapter 7. CONSTRUCTION OF CLASSIFIERS IN THE SPACE OF THE FIRST DERIVATIVES OF THE LIKELIHOOD FUNCTION.....	171
7.1. Statement of the problem of sequence classification.....	171
7.2. Classifiers.....	174
7.2.1. The k nearest neighbor.....	174
7.2.2. Support vector machines.....	175
7.2.3. Multi-class classification.....	183
7.3. Calculation of the first derivatives of the likelihood function logarithm for HMM .....	184
7.4. Studying the possibility of classification in the first derivative space.....	189
Chapter 8. CLASSIFICATION OF SEQUENCES UNDER CONDITIONS WITH NOISE OF VARYING INTENSITY AND NATURE .....	199
8.1. The interference models skewing the observed sequences.....	199
8.2. Studying models in which the probability of the observed symbol occurrence is described by one distribution .....	201
8.2.1. An additive noise with constant distribution parameters.....	202
8.2.2. An additive noise with variable distribution parameters .....	206
8.2.3. A substituting probabilistic noise with constant distribution parameters .....	210
8.2.4. A substituting probabilistic noise with variable distribution parameters .....	212
8.3. Studying models in which the probability of observed symbol occurrence is described by a mixture of distributions .....	215
8.3.1. An additive noise with constant distribution parameters.....	216
8.3.2. An additive noise with variable distribution parameters .....	223
8.3.3. A substituting probabilistic noise with constant distribution parameters.....	228
8.3.4. A substituting probabilistic noise with variable distribution parameters .....	232



---

Chapter 9. THE BEHAVIOR OF CLASSIFIERS UNDER DEVIATIONS FROM ASSUMPTIONS .....	237
9.1. Various distribution laws of the occurrence of observed symbols different from normal .....	237
9.2. A process or an object generating the observed sequence with no hidden states .....	246
9.3. Behavior of classifiers under structural uncertainty .....	252
9.3.1. Studying the two-class classification problem .....	252
9.3.2. Studying the multi-class classification problem .....	260
Afterword.....	265
References.....	268
Appendices .....	293
App A.....	293
App Б .....	296
App В .....	299