УДК 519:234:24:85

В.И. НОРКИН, М.А. КАЙЗЕР

ОБ ЭФФЕКТИВНОСТИ МЕТОДОВ КЛАССИФИКАЦИИ, ОСНОВАННЫХ НА МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА

Ключевые слова: *машинное обучение, классификация, распознавание, минимизация эмпирического риска, метод опорных векторов (SVM), состоятельность, скорость сходимости.*

ВВЕДЕНИЕ

В настоящей работе обсуждается теоретическая эффективность некоторых методов (бинарной) классификации, в частности метода опорных векторов (Support Vector Machine/Method — SVM) [1]. Задача классификации рассматривается в стандартной для статистической теории обучения модели «обучения с учителем». Предполагается, что имеется обучающая выборка парных наблюдений $\{(y_i, x_i), i = 1,\ldots, m\}$ размера m, где x_i — вектор признаков объекта i со значениями в множестве X, y_i — метка класса из дискретного множества Y, которому принадлежит объект i. В статистической теории обучения считается, что пары (y_i, x_i) являются независимыми случайными векторами с общим неизвестным вероятностным распределением P на множестве $Y \times X$. Под задачей классификации понимается построение на основе обучающей выборки отображения (классификатора) из X в Y. В качестве меры эффективности классификатора используется средняя вероятность ошибочной классификации как функция объема обучающей выборки и других параметров модели. Эта величина называется усредненным байесовским риском (в узком смысле), и для него существует теоретический минимум. Для рационального метода классификации риск ошибочной классификации должен стремиться к теоретическому минимуму с ростом объема обучающей выборки, в этом случае говорим о сходимости (по вероятности или почти наверное) метода классификации. Такие методы классификации называются состоятельными, однако состоятельность может иметь место только для определенных классов распределений обучающей выборки.

Одна из проблем статистической теории классификации заключается в том, что теоретическое распределение элементов обучающей выборки неизвестно, поэтому нельзя формально проверить, принадлежит ли распределение данной обучающей выборки к тому или иному классу. Некоторым разрешением этой проблемы могли бы быть методы классификации, состоятельные на любом распределении обучающих данных. Такие методы естественно называть универсально состоятельными [2]. Долгое время не было известно, существуют ли универсально состоятельные методы классификации. Только в 1977 году было показано [3], что этим свойством обладает известный с 1951 года метод k-ближайших соседей. Однако выяснилось [2], что универсально состоятельные методы могут сходиться (снижать риск ошибочной классификации с ростом обучающей выборки) как угодно плохо на некоторых распределениях обучающих данных и, следовательно, не существует универсально наилучшего (оптимального) метода классификации. Таким образом, утверждения об оценках скорости сходимости риска ошибочной классификации к неустранимому минимуму или об оптимальности некоторого метода классификации справедливы только для определенного класса распределений обучающих данных.

Этот вывод относится и к методам минимизации эмпирического риска, в частности к методу опорных векторов [1]. Его линейный вариант (метод оптимальных разделяющих плоскостей) детально исследован в [4, 5], а нелинейный (метод потенциальных функций) — в [6], новейшие версии (методы опорных векторов — SVM) описаны в [1, 7, 8]. В настоящее время SVM успешно конкурируют

с наиболее развитыми системами машинной классификации, поэтому он продолжает оставаться объектом интенсивного теоретического анализа [8, 9]. Классическое обоснование метода базируется на равномерном функциональном законе больших чисел, а полученные оценки скорости сходимости зависят от так называемой VC-емкости (Вапника–Червоненкиса) класса решающих функций [1, 4, 5]. Однако оценка VC-емкости в общем случае представляет непростую проблему, и более того, далеко не всегда класс допустимых функций имеет конечную VC-емкость. Хотя некоторые часто используемые минимизируемые (квадратичные, абсолютного отклонения) функционалы эмпирического риска отражают качество классифицирующего правила, их связь с вероятностью безошибочной классификации не очевидна. Вид имеющихся оценок скорости сходимости в терминах доверительных границ для риска не позволяет сравнивать данный метод с другими, для которых эти оценки получены в терминах сходимости среднего риска.

В настоящей работе исследуется метод опорных векторов для решения задач бинарной классификации с позиций теории некорректных задач и устанавливаются оценки скорости сходимости метода при довольно общих предположениях о распределении обучающих данных. Эти предположения состоят в том, что некоторые характеристики распределения данных (условные медианы и средние) принадлежат определенному функциональному гильбертову пространству (с воспроизводящим ядром). В статье уточняется связь между используемыми функционалами риска и вероятностями ошибочной классификации. Получены оценки скорости сходимости вероятности ошибочной классификации к минимуму, зависящие от распределения данных, но не зависящие от VC-емкости функционального пространства. При этом не используется равномерный функциональный закон больших чисел. Эти оценки содержат неизвестные константы, поэтому непригодны для количественных выводов, однако показывают характер стремления к теоретическому минимуму средней ошибки данного классификатора. Как правило, скорость сходимости имеет порядок $\text{const} / \sqrt[4]{m}$, где m — число элементов в обучающей выборке.

Изложение построено следующим образом. В первом разделе обсуждаются методы классификации, основанные на аппроксимации точного решения задачи минимизации риска классификации. Во втором разделе рассматривается альтернативный подход к классификации, а именно, показано, как задача минимизации вероятности ошибочной классификации может быть сведена к задаче минимизации выпуклого функционала риска. В третьем разделе описывается метод регуляризации для минимизации выпуклых функционалов эмпирического риска, а в четвертом — исследуется его сходимость при увеличении числа обучающих примеров. В пятом разделе эти результаты интерпретируются для задач классификации. В заключении обсуждаются основные особенности метода опорных векторов в свете полученных в статье результатов.

1. БАЙЕСОВСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

Пусть данные наблюдений представляют собой случайные пары (y, x) с распределением P, причем скалярная величина $y \in Y$ может принимать только дискретные значения (метки классов), например $y \in Y = \{0, 1\}$, а компоненты n-мерного вектора $x \in X$ (признаки) могут быть как дискретными, так и непрерывными. Задача с s классами стандартным образом сводится к решению s задач бинарной классификации, в которых один класс — это один из исходных классов, а второй — все остальные. Для любой измеримой функции $f(x): X \to R^1$ бинарное классифицирующее правило определяется по формуле

$$I_{1/2}(f(x)) = \begin{cases} 1, & f(x) > 1/2, \\ 0 & \text{в противном случае.} \end{cases} \qquad (1)$$

Качество классифицирующего правила $I_{1/2}(f(\cdot))$ измеряется байесовским риском, т.е. вероятностью $P\{I_{1/2}(f(x)) \neq y\}$ ошибочной классификации, где $y \in \{0,1\}$. Напомним [2, с. 10], что байесовский риск достигает минимального значения P^* на решающем правиле, задаваемом функцией условной вероятности $p_1(x) = P\{y = 1 | x\}$, но она не известна. В случае многих классов, когда $y \in Y = \{0, 1, 2, \dots\}$, оптимальная байесовская стратегия классификации состоит в максимизации по $l \in \{0, 1, 2, \dots\}$ условного распределения вероятностей $p_l(x) = P\{y = l | x\}$ [10, с. 22], которое, однако, тоже не известно.

Таким образом, один возможный путь построения оптимальных классификаторов состоит в аппроксимации условной вероятности $p_1(x) = P\{y = 1 | x\}$ в бинарном случае или распределения $p_l(x) = P\{y = l | x\}$, $l = 0, 1, \dots$, в общем случае. Например, в методе классификации по k-ближайшим соседям [1, разд. 5] отбирается k наблюдений $\{x_i, i \in I_k(x)\}$, ближайших к вектору признаков x, строится их распределение по классам и вектор x относится к классу с максимальной частотой. Обозначим такой классификатор $g_k(x)$, его качество измеряется величиной вероятности ошибочной классификации $L_k(m) = E_{\{(y_1, x_1), \dots, (y_m, x_m)\}} P\{g_k(x) \neq y\}$, а асимптотическое качество — величиной $L_k^* = \lim_{m \to \infty} L_k(m)$. Известно [1, разд. 5], что $P^* \leq L_k^* \leq P^*(1 + 1/\sqrt{ke})$ для всех распределений и четных k, где e — основание натуральных логарифмов. Кроме того, этот классификатор является универсально состоятельным, т.е. $L_{k(m)}(m) \to P^*$ при $m \to \infty$ и $k(m)/m \to 0$ независимо от вероятностного распределения элементов выборки, хотя скорость сходимости $L_{k(m)}(m)$ к P^* может быть медленной. Интересно отметить, что простейший классификатор $g_1(x)$ (классифицирующий по одному ближайшему соседу) может быть в среднем лучше на некоторых распределениях данных, чем более сложные классификаторы $g_k(x)$ с $k > 1$. В [1] показано, что нельзя построить универсально состоятельный классификатор с фиксированной скоростью сходимости вероятности ошибочной классификации к теоретическому минимуму P^*. Для любого классификатора скорость сходимости может оказаться как угодно медленной при соответствующем выборе распределения исходных данных. Поэтому оценки скорости сходимости могут быть получены только при дополнительных предположениях о распределении наблюдений.

Заметим, что в бинарном случае $p_1(x) = P\{y = 1 | x\}$ является функцией условного среднего (регрессии), поэтому для ее оценки можно применять стандартные подходы регрессионного анализа, в частности непараметрические методы [11]. Пусть $\{(y_i, x_i), i = 1, \dots, m\}$ — обучающая выборка, $\rho(\cdot, \cdot)$ — некоторая функция расстояния между точками в пространстве признаков X, $k(\cdot)$ — некоторая одномерная симметричная плотность вероятностей, θ_m — положительные числа. Тогда ядерная оценка Надараи–Ватсона [11, разд. 5] функции регрессии $p_1(x)$ в данном случае имеет вид

$$\tilde{p}_1(x) = \sum_{i:\, y_i = 1} k\left(\frac{\rho(x, x_i)}{\theta_m}\right) \Big/ \sum_{i=1}^{m} k\left(\frac{\rho(x, x_i)}{\theta_m}\right),$$

а соответствующий бинарный классификатор задается формулой (1) с $f(x) = \tilde{p}_1(x)$.

В работах [12–14] неизвестное условное распределение вероятностей $\{p_l(x), l = 0, 1, \dots\}$ аппроксимируется байесовской оценкой $\{\tilde{p}_l(x), l = 0, 1, \dots\}$ при (сильном) предположении условной независимости признаков (компонент случайного вектора x для объектов из фиксированного класса l). Для такого классификатора в [12–14] получены оценки скорости сходимости вида

$$\mathrm{B(m)} = \mathrm{E}_{\{(y_1, x_1), \dots, (y_m, x_m)\}} \mathrm{P}\{\arg\max_l \tilde{p}_l(x) \neq y\} \leq P^* + C/\sqrt{m},$$

где C — универсальная константа, не зависящая от распределения данных, и доказана их неулучшаемость при сделанных предположениях по характеру зависи-

мости от размера обучающей выборки m. Существенное для этой оценки предположение о независимости признаков детально обсуждается в [10, разд. 3.3].

В бинарном случае хорошо известно (см. [2, с. 16] и ссылки в этой работе), что байесовская ошибка классификации выражается через ошибку $\widetilde{p}_1(x) - p_1(x)$ аппроксимации условной вероятности $p_1(x) = P\{y = 1 | x\}$ следующим образом:

$$P\{I_{1/2}(\widetilde{p}_1(x)) \neq y\} - P^* \leq 2E | \widetilde{p}_1(x) - p_1(x) |. \tag{2}$$

Здесь символ E обозначает математическое ожидание по мере P. Эта оценка дает статистическое обоснование методам классификации, основанным на аппроксимации условных вероятностей $p_l(x) = P\{y = l | x\}$, $l = 0, 1$.

2. СВЯЗЬ ЗАДАЧИ БИНАРНОЙ КЛАССИФИКАЦИИ С ОПТИМИЗАЦИЕЙ ВЫПУКЛЫХ ФУНКЦИОНАЛОВ РИСКА

Другой подход к построению методов классификации состоит в сведении задачи классификации к выпуклой задаче оптимизации функционала риска [9, разд. 4.2]. Далее рассмотрим случаи, не представленные в обзоре [9]. Например, известно [2, с. 11] , что $p_1(x) = \arg\min_f E(y - f(x))^2$. Если $f(x)$ — некоторое приближенное решение задачи минимизации квадратичного риска, то соответствующее решающее правило определяется по формуле (1), а оценка качества классификации — по формуле (2). Этот подход к бинарной классификации подробно обсуждается в [15]. Кроме того, в статистической теории классификации и обучения используются функционалы риска вида

$$R_\varepsilon(f) = E \max\{0, | y - f(x) | - \varepsilon\}, \ \varepsilon \geq 0,$$

и, в частности, $R_0(f) = E | y - f(x) | = L_1(f)$ [1]. Их применение в какой-то мере обосновано оценкой [2, с. 20]

$$P\{I_{1/2}(f(x)) \neq y\} - \min_f P\{I_{1/2}(f(x)) \neq y\} \leq$$
$$\leq 2(E | y - f(x) | - \min_f E | y - f(x) |), \tag{3}$$

где минимумы берутся по множеству борелевских функций на X.

Следующая теорема дает оценку качества классификатора, минимизирующего квадратичный функционал риска $L_2(f) = E(y - f(x))^2$, отличную от (2).

Теорема 1. Пусть F — множество борелевских функций на $x \in X$ такое, что $p_1(x) = P\{y = 1 | x\} \in F$. Тогда для любой функции $f(\cdot) \in F$ имеет место оценка

$$P\{I_{1/2}(f(x)) \neq y\} - \min_{f\text{-измерима}} P\{I_{1/2}(f(x)) \neq y\} \leq$$
$$\leq 2\sqrt{L_2(f) - \min_{f \in F} L_2(f)}. \tag{4}$$

Доказательство. Представим

$$P\{I_{1/2}(f(x)) \neq y\} = E_x\{P\{I_{1/2}(f(x)) \neq y | x\}\},$$
$$E(f(x) - y)^2 = E_x\{E\{(f(x) - y)^2 | x\}\},$$

где $P\{\cdot | x\}$ и $E\{\cdot | x\}$ — условная вероятность и условное математическое ожидание при фиксированной компоненте x случайного вектора (y, x); E_x — математическое ожидание по распределению компоненты x. Рассмотрим функции $p_1(x) = P\{y = 1 | x\}$, $p_0(x) = P\{y = 0 | x\} = 1 - p_1(x)$ и $e(h, x) = E\{(h - y)^2 | x\}$.

Справедливы соотношения:

$$r(h, x) = P\{I_{1/2}(h) \neq y | x\} = \begin{cases} p_0(x) = 1 - p_1(x), & h > 1/2, \\ p_1(x), & h \leq 1/2; \end{cases}$$

$$e(h, x) = p_1(x)(h - 1)^2 + (1 - p_1(x))h^2 = h^2 - 2h p_1(x) + p_1(x) =$$

$$= (h - p_1(x))^2 + p_1(x)(1 - p_1(x)) = (h - p_1(x))^2 + e^*(x),$$

$$e^*(x) = p_0(x)p_1(x).$$

Отсюда следует, что для любого $h \in R^1$ выполнено

$$r(h,x) \geq r(p_1(x),x) = \min\{p_0(x), p_1(x)\} = r^*(x), \tag{5}$$

$$e(h,x) \geq e(p_1(x),x) = p_0(x)p_1(x) = e^*(x). \tag{6}$$

Если $p_1(x) \leq 1/2$, то

$$r(h,x) - r^*(x) = \begin{cases} 0, & h \leq 1/2, \\ p_0(x) - p_1(x), & h > 1/2. \end{cases}$$

Если $p_1(x) > 1/2$, то

$$r(h,x) - r^*(x) = \begin{cases} p_1(x) - p_0(x), & h \leq 1/2, \\ 0, & h > 1/2. \end{cases}$$

Пусть $p_1(x) \leq 1/2$. При $h \leq 1/2$ выполнено $r(h,x) - r^*(x) = 0 \leq$ $\leq 2(e(h,x) - e^*(x))^{1/2}$. При $h > 1/2$ имеет место

$$e(h,x) - e^*(x) = (h - p_1(x))^2 \geq (1/4)(1 - 2p_1(x))^2 =$$

$$= (1/4)(p_0(x) - p_1(x))^2 = (1/4)(r(h,x) - r^*(x))^2.$$

Таким образом, при $p_1(x) \leq 1/2$ и всех h выполнено

$$r(h,x) - r^*(x) \leq 2\sqrt{e(h,x) - e^*(x)}. \tag{7}$$

Доказательство этого неравенства для случая $p_1(x) > 1/2$ проводится аналогично. Подставляя в (5), (6), (7) значение $h = f(x)$ и взяв математическое ожидание по x, для любой измеримой функции $f(x)$ имеем

$$P\{I_{1/2}(f(x)) \neq y\} \geq P\{I_{1/2}(p_1(x)) \neq y\} = P^*,$$

$$E\{(f(x) - y)^2\} \geq E\{(p_1(x) - y)^2\},$$

$$P\{I_{1/2}(f(x)) \neq y\} - P\{I_{1/2}(p_1(x)) \neq y\} \leq 2\sqrt{E\{(f(x) - y)^2\} - E\{(p_1(x) - y)^2\}}.$$

При получении последнего неравенства использовалось неравенство Иенсена для вогнутой функции $\sqrt{(\cdot)}$. Поэтому если $p_1(\cdot) \in F$, то для любой функции $f(\cdot) \in F$ выполнено

$$P\{I_{1/2}(f(x)) \neq y\} - P^* \leq 2\sqrt{E\{(f(x) - y)^2\} - \min_{f(\cdot) \in F} E\{(f(x) - y)^2\}},$$

что и требовалось доказать.

Рассмотрим задачу [2, с. 20]

$$L_1(f) = E|f(x) - y| \to \inf_{f \in F}, \tag{8}$$

где F — множество борелевских функций на $x \in X$ такое, что $g_1(\cdot) \in F$, где

$$g_1(x) = \begin{cases} 1, & p_1(x) > 1/2, \\ 0, & p_1(x) \leq 1/2, \end{cases} \qquad p_1(x) = P\{y = 1 | x\}.$$

Следующая теорема обобщает оценку (3) и устанавливает связь между байесовским риском и выпуклым функционалом $L_1(f) = E|f(x) - y|$.

Теорема 2. Пусть F — множество борелевских функций на $x \in X$ такое, что $g_1(\cdot) \in F$ или $\mu(\cdot) \in F$, где $\mu(\cdot)$ — любая условная медиана распределения $P\{\cdot | x\}$ при фиксированном x. Тогда для любой функции $f(\cdot) \in F$ имеет место оценка

$$\mathrm{P}\{\mathrm{I}_{1/2}(\mathrm{f}(\mathrm{x})) \neq \mathrm{y}\} - \min_{\mathrm{f-измерима}} \mathrm{P}\{\mathrm{I}_{1/2}(\mathrm{f}(\mathrm{x})) \neq \mathrm{y}\} \leq$$
$$\leq 2(\mathrm{R}(\mathrm{f}) - \min_{\mathrm{f} \in \mathrm{F}} \mathrm{R}(\mathrm{f})), \qquad (9)$$

где $R(f) = L_1(f) = E|f(x) - y|$.

Доказательство. Для случая, когда F — множество всех измеримых функций на $x \in X$, утверждение теоремы имеется в [2, с. 20] (без множителя 2 в правой части (9)). Представим

$$P\{I_{1/2}(f(x)) \neq y\} = E_x\{P\{I_{1/2}(f(x)) \neq y | x\}\},$$
$$E|f(x) - y| = E_x\{E\{|f(x) - y| | x\}\}.$$

Рассмотрим функции $p_1(x) = P\{y = 1 | x\}$ и $a(h, x) = E\{|h - y| | x\}$. Справедливы представления:

$$r(h, x) = P\{I_{1/2}(h) \neq y | x\} = \begin{cases} 1 - p_1(x), & h > 1/2, \\ p_1(x), & h \leq 1/2, \end{cases}$$

$$a(h, x) = E\{|h - y| | x\} = p_1(x)|h - 1| + (1 - p_1(x))|h|.$$

Обозначим $r^*(x) = \min\{p_1(x), 1 - p_1(x)\}$. Для любой условной медианы $\mu(\cdot)$ имеет место

$$\mu(x) \in \begin{cases} 1, & p_1(x) > 1/2, \\ [0, 1], & p_1(x) = 1/2, \\ 0, & p_1(x) \leq 1/2, \end{cases}$$

и, в частности, $g_1(x)$ является условной медианой распределения P при фиксированном x. Отсюда следует, что для любого $h \in R^1$ выполнено

$$r(h, x) \geq r(p_1(x), x) = r^*(x), \qquad (10)$$

$$\mathrm{a}(\mathrm{h}, \mathrm{x}) \geq \mathrm{a}(\mu(\mathrm{x}), \mathrm{x}) = \mathrm{r}^*(\mathrm{x}). \qquad (11)$$

Докажем неравенство

$$r(h, x) - r^*(x) \leq 2(a(h, x) - r^*(x)). \qquad (12)$$

Рассмотрим функции

$$\varphi(p, h) = \begin{cases} 1 - p, & h > 1/2, \\ p, & h \leq 1/2; \end{cases}$$

$$\psi(p, h) = p|h - 1| + (1 - p)|h| = \begin{cases} p - h, & h \leq 0, \\ p + h - 2ph, & 0 \leq h \leq 1, \\ h - p, & h \geq 1. \end{cases}$$

Покажем, что при $p \leq 1/2 \leq 1 - p$ выполнено $\varphi(p, h) - p \leq 2(\psi(p, h) - p)$. Действительно,

$$\varphi(p, h) - p = 0 \leq 2(\psi(p, h) - p) = -2h \text{ при } h \leq 0;$$

$$\varphi(p, h) - p = 0 \leq 2(\psi(p, h) - p) = 2h(1 - 2p) \text{ при } 0 \leq h \leq 1/2;$$

$$\varphi(p, h) - p = 1 - 2p \leq 2(\psi(p, h) - p) = 2h(1 - 2p) \text{ при } 1/2 < h \leq 1;$$

$$\varphi(p, h) - p = 1 - 2p \leq 2(\psi(p, h) - p) = 2(h - 2p) \text{ при } 1 < h.$$

Аналогично проверяется, что при $1 - p \leq 1/2 \leq p$ выполнено $\varphi(p, h) - (1 - p) \leq$ $\leq 2(\psi(p, h) - (1 - p))$. Таким образом, неравенство (12) доказано.

Подставляя в (10), (11), (12) значение $h = f(x)$ и взяв математическое ожидание по x, для любой борелевской функции $f(x)$ получаем

$$P\{I_{1/2}(f(x)) \neq y\} \geq P\{I_{1/2}(p_1(x)) \neq y\} = P^*,$$

$$E|f(x)-y| \geq E|\mu(x)-y|,$$

$$P\{I_{1/2}(f(x)) \neq y\} - P^* \leq 2(E|f(x)-y| - E|\mu(x)-y|). \qquad (13)$$

Из (13) следует требуемое неравенство (9).

Теорема доказана.

Таким образом, минимизация функционала $L_1(f) = E|f(x)-y|$ по множеству борелевских функций F такому, что $g_1(\cdot) \in F$ или $\mu(\cdot) \in F$, в силу (9) автоматически ведет к минимизации функционала байесовского риска.

Как известно, минимум квадратичного функционала риска $L_2(f)$ достигается на функции условного среднего $m(x) = \int_{R^1} yP(dy|x)$ распределения P. Для неквадратичных функционалов риска соответствие их минимумов каким-либо характеристикам распределения менее очевидно, но в случае функционала среднего абсолютного отклонения, часто используемого в теории статистического обучения, такое соответствие может быть установлено.

Теорема 3. В задаче минимизации функционала риска

$$R(f) = E_{(x,\,y)} \max\{(1-\delta)(f(x)-y), \delta(y-f(x))\}$$

по всем измеримым функциям $f(x)$ минимум достигается на условных δ-квантилях распределения P, т.е. на функциях $q(x)$ таких, что $P\{y \leq q(x)|x\} \geq \delta$. В частности, при $\delta = 0{,}5$ функционал риска имеет вид $R(f) = (1/2)E|f(x)-y|$ и его минимум достигается на условных медианах $\mu(x)$ распределения $P\{\cdot|x\}$.

Данное утверждение получено в [16, 17]; в контексте стохастических минимаксных задач этот факт был установлен в работах [18, 19]; он детально обсуждается в [20]. Отметим, что δ-квантиль и медиана распределения, в общем случае, могут быть не единственными.

Если есть априорные основания полагать, что условные медианы распределения $P(\cdot)$ принадлежат некоторому классу функций, например некоторому гильбертову пространству H, то в (8) можно положить $F = H$. В этом случае говорят об отсутствии ошибки аппроксимации (медиан) функциями из H. В общем случае ошибка аппроксимации существует, ее оценки имеются в [2, 8, 11, 21].

В [1] при решении задач классификации часто используются ε-нечувствительные функционалы риска вида

$$R_\varepsilon(f) = E\max\{0, |f(x)-y| - \varepsilon\}.$$

Легко видеть, что функционал $L_1(f) = E|f(x)-y|$ связан с $R_\varepsilon(f)$ соотношением $L_1(f) - \varepsilon \leq R_\varepsilon(f) \leq L_1(f)$ равномерно по всем борелевским функциям f, поэтому в условиях теоремы 2 из (9) следует соотношение

$$P\{I_{1/2}(f(x)) \neq y\} - \min_{f\in F} P\{I_{1/2}(f(x)) \neq y\} \leq 2(R_\varepsilon(f) - \min_{f\in F} R_\varepsilon(f)) + 2\varepsilon.$$

Использование ε-нечувствительных функционалов риска позволяет упростить классификатор [1], хотя и ухудшает точность классификации на 2ε.

В задачах классификации часто используются функционалы вида $R(f) = E\varphi(-yf(x))$ [9], где метки классов $y \in \{\pm 1\}$, $\varphi(\cdot)$ — некоторая неотрицательная выпуклая неубывающая функция потерь такая, что $\lim_{t\to-\infty} \varphi(t) = 0$ и $\varphi(0) = 1$. С их помощью также установлены оценки риска безошибочной классификации, аналогичные оценкам (4), (9).

Отметим, что в задачах классификации признаковое пространство X часто является дискретным, например, оно может состоять из вершин единичного куба [6, гл. III, §1.3]. В этом случае функция $f(x)$, $x \in X$, задается конечным, возможно очень большим, числом значений, т.е. является вектором большой размерности.

3. ОПТИМИЗАЦИЯ РЕГУЛЯРИЗОВАННЫХ ФУНКЦИОНАЛОВ ЭМПИРИЧЕСКОГО РИСКА И МЕТОД ОПОРНЫХ ВЕКТОРОВ

В разд. 2 показано, что задача бинарной классификации может быть сведена к минимизации выпуклого функционала риска. В общем случае она имеет вид

$$R(f) = Ec(y, f(x)) \to \min_{f \in F}, \tag{14}$$

где $c(y, f(x))$ — некоторая функция потерь, например, $c(y, f(x)) = (y - f(x))^2$, $c(y, f(x)) = |y - f(x)|$, $c(y, f(x)) = \max\{0, 1 - yf(x)\}$; F — допустимый класс функций. Обозначим F^* множество решений задачи (14). В предыдущем разделе также показано, что минимум в таких задачах может достигаться на некоторой характеристике распределения случайного вектора наблюдений $z = (y, x)$, например функции условного среднего $p_1(x)$ или условной медиане $\mu(x)$. Если есть основания полагать, что эти характеристики принадлежат некоторому классу функций F, например подмножеству некоторого гильбертова пространства функций H, то в (14) можно считать $F \subset H$. В статистической теории обучения используются разнообразные классы функций (классические гильбертовы пространства с заданным базисом, нейросетевые суперпозиции, деревья и другие [2]) и, в частности, так называемые репродуктивные гильбертовы пространства функций H_k, порожденные ядром k.

Определение 1 (репродуктивное гильбертово пространство). Гильбертово пространство $H_k(X)$ функций, определенных на замкнутом множестве $X \subset R^n$, называется репродуктивным гильбертовым пространством (РГП), если существует функция двух векторных переменных $k(\cdot,\cdot)$, определенная на декартовом произведении $X \times X$, обладающая следующими свойствами:

а) $k(\cdot, x) \in H_k(X)\ \forall x \in X$;

б) $f(x) = \langle f, k(\cdot, x)\rangle_k\ \forall f \in H_k(X)$, $\forall x \in X$ (репродуктивное свойство ядра).

Теория РГП изложена в работах [7, 21, 22, 23]. В частности, известно, что множество функций $\left\{ f(x) = \sum_s \alpha_s k(\bar{x}_s, x) \right\}$ из РГП $H_k = H_k(X)$, где $\{\bar{x}_s\}$ — произвольный конечный набор точек из X, $\{\alpha_s\}$ — произвольный конечный набор чисел, является плотным в $H_k(X)$.

В задачах классификации распределение $P(\cdot)$ наблюдений обычно не известно полностью, а имеется набор независимых наблюдений $\{z_i = (y_i, x_i), i = 1, \ldots, m\}$ векторной случайной величины $z = (y, x)$ с распределением $P(\cdot)$, который в статистической теории обучения называется обучающей выборкой. Это позволяет аппроксимировать неизвестное распределение $P(\cdot)$ эмпирическим распределением $P_m(\cdot)$, а функционал риска $R(f) = Ec(y, f(x))$ с функцией потерь $c(y, f)$ — эмпирическим средним (эмпирическим риском) $\widetilde{R}_m(f) = (1/m)\sum_{i=1}^{m} c(z_i, f(x_i))$.

Задача минимизации функционала риска (14), вообще говоря, может быть некорректной, т.е. иметь неоднозначные решения, быть неустойчивой по отношению к возмущениям функционала. В статистической теории обучения классификации исходный функционал риска $R(f)$ заменяется случайным приближением $\widetilde{R}_m(f)$, т.е. рассматривается его стохастическое возмущение вида $R(f) + \delta_m(f)$, где $\delta_m(f) = \widetilde{R}_m(f) - R(f)$. Поэтому для нахождения приближенных решений применяется метод регуляризации Тихонова в функциональном (гильбертовом) пространстве H [24, 25]. Рассмотрим метод регуляризации в РГП при определенных (эмпирических) случайных возмущениях функционала и для общих выпуклых (не только квадратичных) функционалов риска, который сводится к решению семейства задач минимизации регуляризованного эмпирического риска

$$\widetilde{\mathrm{R}}_{\mathrm{m}}(\mathrm{f}) + \lambda \,\|\, \mathrm{f} \,\|_{\mathrm{k}}^2 = \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{c}(\mathrm{y}_{\mathrm{i}}, \mathrm{f}(\mathrm{x}_{\mathrm{i}})) + \lambda \,\|\mathrm{f}\,\|_{\mathrm{k}}^2 \to \inf_{\mathrm{f} \in \mathrm{H}_{\mathrm{k}}}, \tag{15}$$

где H_k — некоторое РГП, порожденное ядром k. Оказывается, что решение регуляризованной задачи (15) в РГП сводится к задаче конечномерной оптимизации, а для кусочно-линейных функций потерь — к задаче квадратичной оптимизации при линейных ограничениях. В силу так называемой теоремы о представлении решения в РГП [7, Theorem 4.2, p. 90; 26] решение задачи (15) существует и может быть представлено в виде

$$f_m^\lambda(x) = \sum_{i=1}^m \alpha_i k(x_i, x), \tag{16}$$

где $\alpha^m = \{\alpha_i\}$ — некоторый неизвестный набор действительных чисел, $\{x_i\}$ — известный набор точек наблюдения. Подставляя выражение (16) в (15) и используя репродуктивное свойство ядра, приходим к следующей конечномерной задаче оптимизации:

$$R_m(\alpha^m) = \frac{1}{m}\sum_{i=1}^m c\left(y_i, \sum_{j=1}^m \alpha_j k(x_i, x_j)\right) + \lambda \sum_{i,j=1}^m \alpha_i \alpha_j k(x_i, x_j) \to \min_{\alpha^m}. \tag{17}$$

Если функция потерь $c(y, \cdot)$ выпукла и неотрицательна, а матрица $\{k(x_i, x_j)\}$ положительно определена, то эта задача имеет единственное решение f_m^λ. В решении задачи (17) в силу наличия квадратичного штрафа в целевой функции значительная часть коэффициентов разложения (16) может быть равна нулю. Векторы x_i, соответствующие ненулевым коэффициентам разложения (16), называются опорными векторами, а в целом метод классификации, основанный на решении задач (15)–(17), называется методом опорных векторов [1, 7]).

Отметим, что для негладких кусочно-линейных функций потерь, например, $c(y, f(x)) = |y - f(x)|$, $c(y, f(x)) = \max\{0, 1 - yf(x)\}$, задача (17) является выпуклой и негладкой, однако с помощью дополнительных переменных она легко сводится к задаче квадратичного программирования при линейных ограничениях. Детали численной реализации метода можно найти, например, в [7, 27].

4. СХОДИМОСТЬ МЕТОДА ОПОРНЫХ ВЕКТОРОВ ПРИ НЕОГРАНИЧЕННОМ РОСТЕ ЧИСЛА НАБЛЮДЕНИЙ

Рассмотрим асимптотические свойства при $m \to \infty$ и $\lambda \to 0$ решений $f_m^\lambda(x)$ задачи минимизации регуляризованного эмпирического риска (15). В работах [1, 4, 5] вопрос сходимости $R(f_m^\lambda) \to \inf_{f \in F} R(f)$ исследован в предположении ограниченной емкости класса функций F. Примененный подход основан на установлении условий равномерной по $f \in F$ сходимости эмпирических аппроксимаций функционала риска $R_m^\lambda(f) = \frac{1}{m}\sum_{i=1}^m c(z_i, f(x_i))$ к его истинному значению $R(f) = Ec(z, f(x))$, т.е. $\sup_{f \in F} |R_m^\lambda(f) - R(f)| \to 0$ при $m \to \infty$. Однако не всегда подходящий класс функций имеет конечную емкость (конечную размерность в смысле Вапника–Червоненкиса [4]). Более слабые требования для равномерной на классе функций сходимости эмпирических средних к функционалу риска можно сформулировать в терминах сложности класса по Радемахеру [9, разд. 3]. Заметим, что условие равномерной сходимости аппроксимаций $R_m^\lambda(f)$ к $R(f)$ не является необходимым для сходимости минимумов [28]. Поэтому следуем другому подходу, основанному на свойстве устойчивости регуляризованных решений $f_m^\lambda(x)$ по отношению к отдельным наблюдениям. Подобный подход использовался в [7, разд. 12.1; 29, 30, 31], где исследовалась сходимость оценок риска по вероятности. В отличие от этих работ в данной статье устанавливаются условия на $\lambda = \lambda(m)$, при которых оценки $f_m^{\lambda(m)}(x)$ равномерно по $x \in X$ сходятся с вероятностью единица к минимуму f^* функционала риска $R(f)$, имеющему минимальную норму. В этом смысле построенные классификаторы асимптотически устойчивы.

Предположение 1 (свойства функции потерь). Функция потерь $c(y,\cdot)$ неотрицательна, выпукла и липшицева по второму аргументу с константой L_y на множестве $\Phi = \{f(x) \mid f \in F, x \in X\}$.

Предположение 2 (свойства ядра). Порождающее ядро $k(\cdot,\cdot)$ удовлетворяет условию $\sup_{x\in X} |k(x,x)| = K^2 < +\infty$.

Очевидно, функции потерь $c(y,f) = |y-f|$, $c(y,f) = \max\{0, 1-yf\}$ удовлетворяют предположению 1 при любом множестве Φ, а функция $c(y,f) = (y-f)^2$ удовлетворяет этому предположению при ограниченном множестве Φ. Обозначим

$$L = \max_{y\in Y} L_y,\ C = \max_{y\in Y} c(y,0). \tag{18}$$

Следующая теорема дает оценку неоптимальности (в среднем) приближенных решений f_m^λ как функцию m и λ. Эти оценки являются случайными величинами со значениями в функциональном пространстве H_k и определены на счетном произведении исходного вероятностного пространства (X, B_X, P).

Теорема 4 [32, 33]. Пусть решение задачи (14) существует, функции f_m^λ являются решениями задачи (15). Тогда в сделанных предположениях для любого $\lambda > 0$ и m имеет место оценка

$$E_m R(f_m^\lambda) \le R(f^*) + 2\frac{2C + L\,\|f^*\|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{\lambda C})}{\lambda\sqrt{m}} + \lambda\,\|f^*\|_k^2, \tag{19}$$

где математическое ожидание E_m берется по всем выборкам $\{z_1,\dots,z_m\}$ с независимыми одинаково распределенными наблюдениями, f^* — любое решение задачи (14), $\|f^*\|_\infty = \sup_{x\in X} |f(x)|$, $\|f^*\|_k$ — норма функции f^* в пространстве H_k.

Теорема гарантирует сходимость в среднем величины $R(f_m^\lambda)$ к минимальному значению $R(f^*)$ при $\lambda(m) \to 0$ и $\sqrt{m}\lambda(m) \to 0$, когда $m \to \infty$.

Укажем условия сильной состоятельности оценок $f_m^\lambda(x)$, т.е. их равномерной по $x \in X$ сходимости к некоторому минимуму $f^*(x)$ функционала риска R при $\lambda = \lambda(m) \to 0$ и $m \to \infty$.

Определение 2 [24]. Решение $f^* \in F^*$ задачи называется нормальным, если оно имеет минимальную норму, $\|f^*\|_k = \min_{f\in F^*} \|f\|_k$.

Следующие две теоремы из [32, 33] дают достаточные условия равномерной сходимости с вероятностью единица приближенных решений $f_m^{\lambda(m)}$ к нормальному решению $f^* \in F^*$ задачи (14), т.е. $\lim_{m\to\infty} \sup_{x\in X} |f_m^{\lambda(m)}(x) - f^*(x)| = 0$.

Теорема 5 (достаточные условия сильной состоятельности метода опорных векторов). Пусть решение задачи (14) существует и выполнены предположения 1, 2. Рассмотрим семейство решений $f_m^{\lambda(m)}$ задачи (15), причем $\lim_{m\to\infty} \lambda(m) = 0$. Тогда если $\lim_{m\to\infty} m\lambda^2(m)/\ln m = \infty$, то $R(f_m^{\lambda(m)}) \to R(f^*)$. Если $\lim_{m\to\infty} m\lambda^4(m)/\ln m = \infty$, то $R(f_m^{\lambda(m)}) \to R(f^*)$ и решения $f_m^{\lambda(m)}$ задачи (15) равномерно по $x \in X$ сходятся к нормальному решению f^* задачи (14) с вероятностью единица при $m \to +\infty$.

Теорема 6 (оценка скорости сходимости метода опорных векторов). Пусть в условиях предыдущей теоремы $\lambda(m) = \Lambda(\ln m)^\varepsilon / m^{1/4}$, $\Lambda > 0$, $1/4 < \varepsilon \le 1$, тогда справедливы утверждения теоремы 5 и имеет место оценка

$$E_m R(f_m^{\lambda(m)}) - R(f^*) \le$$

$$\le 2\frac{2C + L\|f^*\|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{2\Lambda C})}{\Lambda(\ln m)^\varepsilon \sqrt[4]{m}} + \frac{\|f^*\|_k^2\,\Lambda(\ln m)^\varepsilon}{\sqrt[4]{m}}. \tag{20}$$

5. ЭФФЕКТИВНОСТЬ МЕТОДА ОПОРНЫХ ВЕКТОРОВ ПРИ РЕШЕНИИ ЗАДАЧ БИНАРНОЙ КЛАССИФИКАЦИИ

С помощью решения f_m^λ задачи (15) соответствующий бинарный классификатор строится следующим образом:

$$I_{1/2}(f_m^\lambda(x)) = \begin{cases} 1, & f_m^\lambda(x) > 1/2, \\ 0 & \text{в противном случае.} \end{cases} \tag{21}$$

Для заданной обучающей выборки эффективность классификатора измеряется величиной вероятности ошибки классификации

$$\Delta_m^\lambda = P\{I_{1/2}(f_m^\lambda(x)) \neq y\} - \min_{f \in F} P\{I_{1/2}(f(x)) \neq y\},$$

которая оценивается сверху через разности $[R(f_m^\lambda) - R(f^*)]$ согласно неравенствам (4), (9) из теорем 1, 2, при условии, что условные медианы и средние принадлежат допустимому множеству F задачи минимизации риска (14). Для получения средней вероятности ошибки классификации необходимо взять математическое ожидание $E_m \Delta_m^\lambda$ по всем независимым обучающим выборкам $\{(y_i, x_i)\}$ объема m. В свою очередь, среднее значение $[E_m R(f_m^\lambda) - R(f^*)]$ ошибки минимизации функционала риска по всем возможным обучающим выборкам оценивается неравенствами (19), (20) из теорем 4, 6. Таким образом, приходим к следующим результатам.

Теорема 7 (оценка эффективности метода опорных векторов при использовании негладкого функционала риска $L_1(f)$). Предположим, что условная медиана $f^*(x)$ вероятностного распределения P независимых элементов обучающей выборки $\{(y_i, x_i)\}$ принадлежит подмножеству F некоторого репродуктивного гильбертова пространства H_k с порождающим ядром k. Для бинарного классификатора (21), где функция $f_m^\lambda(x)$ является решением задачи (15) с функцией потерь $c(y, f) = |y - f|$ или $c(y, f) = \max\{0, 1 - yf\}$, средняя по всем обучающим выборкам $\{(y_i, x_i)\}$ объема m ошибка классификации оценивается следующим образом:

$$E_m \Delta_m^\lambda \le 4 \frac{2C + L\|f^*\|_\infty}{\sqrt{m}} + \frac{2LK(5LK + 2\sqrt{\lambda C})}{\lambda\sqrt{m}} + 2\lambda \|f^*\|_k^2.$$

Здесь константы L, C определены в (18), константа K определена в предположении 2. При $\lambda(m) = \Lambda \ln m / m^{1/4}$, $\Lambda > 0$, эта оценка принимает вид

$$E_m \Delta_m^\lambda \le 4 \frac{2C + L\|f^*\|_\infty}{\sqrt{m}} + \frac{2LK(5LK + 2\sqrt{2\Lambda C})}{\Lambda(\ln m)\sqrt[4]{m}} + \frac{2\|f^*\|_k^2 \Lambda(\ln m)}{\sqrt[4]{m}}.$$

Теорема 8 (оценка эффективности метода опорных векторов при использовании квадратичного функционала риска $L_2(f)$). Предположим, что условное среднее $p_1(x) = P\{y = 1 | x\} = E\{y | x\}$ вероятностного распределения P независимых элементов обучающей выборки $\{(y_i, x_i)\}$ принадлежит подмножеству F некоторого репродуктивного гильбертова пространства H_k с порождающим ядром k. Для бинарного классификатора (21), где функция $f_m^\lambda(x)$ является решением задачи (15) с квадратичной функцией потерь $c(y, f) = (y - f)^2$, средняя по всем обучающим выборкам объема m ошибка классификации оценивается следующим образом:

$$E_m \Delta_m^\lambda \le 2\left(2\frac{2C + L\|p_1\|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{\lambda C})}{\lambda\sqrt{m}} + \lambda\|p_1\|_k^2\right)^{1/2}.$$

Здесь константы L, C определены в (18), константа K определена в предположении 2. При $\lambda(m) = \Lambda \ln m / m^{1/4}$, $\Lambda > 0$, эта оценка принимает вид

$$E_m \Delta_m^{\lambda} \le \left(2\frac{2C + L\| p_1 \|_{\infty}}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{2\Lambda C})}{\Lambda(\ln m)\sqrt[4]{m}} + \frac{\| p_1 \|_k^2 \Lambda \ln m}{\sqrt[4]{m}} \right)^{1/2}.$$

ЗАКЛЮЧЕНИЕ

Из результатов настоящей статьи можно сделать несколько выводов, касающихся применения метода опорных векторов для решения задач бинарной классификации.

При использовании метода опорных векторов важно правильно определить класс функций F и пространство $H \supseteq F$, которым принадлежат условные медианы и условное среднее вероятностного распределения элементов обучающей выборки. В этом случае говорят об отсутствии ошибки аппроксимации медианы и среднего функциями из $F \subseteq H$. Поскольку теоретическое распределение обучающих данных неизвестно, а имеется только конечная выборка наблюдений с этим распределением, выбор пространства H и его подмножества F для конкретной реализации метода опорных векторов не является формализованным актом. Если $F = H = H_k$ — некоторое РГП функций, то построение классификатора сводится к решению задачи квадратичного программирования.

Метод опорных векторов является состоятельным (в случае отсутствия ошибки аппроксимации), а именно, при выборе параметра регуляризации $\lambda(m)$ согласно условиям $\lim_{m\to\infty} \lambda(m) = 0$ и $\lim_{m\to\infty} \lambda(m)\sqrt{m} = \infty$ вероятность ошибочной классификации стремится к теоретическому минимуму (в среднем и по вероятности) для любого распределения обучающих данных. Однако полученные оценки скорости сходимости средней ошибки классификации к минимуму содержат неизвестные константы ($\| f^* \|_{\infty}$, $\| f^* \|_k$, $\| p_1 \|_{\infty}$, $\| p_1 \|_k^2$), зависящие от вероятностного распределения элементов обучающей выборки.

Скорость сходимости к минимуму средней вероятности ошибочной классификации (при увеличении объема m обучающей выборки) методом опорных векторов при использовании функционала абсолютного отклонения $L_1(f)$ имеет порядок $\text{const} / \sqrt[4]{m}$, а квадратичного функционала риска $L_2(f)$ — порядок $\text{const} / \sqrt[8]{m}$. Оценки скорости сходимости не содержат в явном виде размерности признакового пространства (размерности вектора x), однако эта размерность может входить в оценки через константу K, характеризующую порождающее ядро k пространства H_k. Например, для полиномиального ядра вида $k(x, x') = (1 + \langle x, x' \rangle)^q$, $q \ge 1$, и n-мерного вектора x с бинарными компонентами соответствующая константа имеет вид $K = (1+n)^q$. В заключение заметим, что при более сильных предположениях на распределение P обучающих данных скорость сходимости метода опорных векторов может быть значительно выше, чем в теоремах 7, 8, например порядка const / m [9].

СПИСОК ЛИТЕРАТУРЫ

1. Vapnik V.N. Statistical learning theory. — New York: Wiley, 1998. — 736 p.
2. Devroye L., Györfi L., Lugosi G. A probabilistic theory of pattern recognition. — New York: Springer, 1996. — 634 p.
3. Stone C. Consistent nonparametric regression // Ann. Statistics. — 1977. — **5.** — P. 595–645.
4. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. Статистические проблемы обучения. — М.: Наука, 1974. — 416 с.
5. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 448 с.
6. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 384 с.

7. Schöelkopf B., Smola A.J. Learning with kernels. Support vector machines, regularization, optimization, and beyond. — Cambridge (MA): MIT Press, 2002. — 626 p.
8. Steinwart I., Christmann A. Support vector machines. — New York: Springer, 2008. — 602 p.
9. Boucheron S., Bousquet O., Lugosi G. Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — **9.** — P. 323–375.
10. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наук. думка, 2004. — 536 с.
11. Györfi L., Kohler M., Krzyzak A., Walk H. A distribution free theory of nonparametric regression. — New York; Berlin; Heidelberg: Springer, 2002. — 647 p.
12. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры классификации объектов // Кибернетика и системный анализ. — 1995. — № 4. — С. 76–89.
13. Сергиенко И.В., Гупал А.М. Оптимальные процедуры распознавания и их применение // Там же. — 2007. — № 6. — С. 41–54.
14. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. — Киев: Наук. думка, 2008. — 232 с.
15. Poggio T., Smale S. The mathematics of learning: Dealing with data // Notices Amer. Math. Soc. — 2003. — **50,** N 5. — P. 537–544.
16. Koenker R., Bassett G.W. Regression quantiles // Econometrica. — 1978. — **46.** — P. 33–50.
17. Koenker R. Quantile regression. — Cambridge; New York: Cambridge Univ. Press, 2005. — 366 p.
18. Ермольев Ю.М., Ястремский А.И. Стохастические модели и методы в экономическом планировании. — М.: Наука, 1979. — 254 с.
19. Ermoliev Y.M., Leonardi G. Some proposals for stochastic facility location models // Math. Modelling. — 1982. — **3.** — P. 407–420.
20. Ruszczynski A., Shapiro A. (Eds.) Stochastic programming // Handbooks in OR & MS. — Amsterdam: Elsevier, 2003. — **10.** — 682 p.
21. Cucker F., Smale S. On the mathematical foundations of learning // Bull. Amer. Math. Soc. — 2001. — **89,** N 1. — P. 1–49.
22. Ароншайн Н. Теория воспроизводящих ядер // Математика (Период. сб. перевод. иностр. статей). — М.: Изд-во иностр. лит., 1963. — **7,** № 2. — С. 67–130.
23. Berlinet A., Thomas-Agnan C. Reproducing kernel Hilbert spaces in probability and statistics. — Dordrecht; Boston; London: Kluwer Acad. Publ., 2004. — 355 p.
24. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. — Изд. 3-е, испр. — М.: Наука, 1986. — 288 с.
25. Васильев Ф.П. Методы решения экстремальных задач. Задачи минимизации в функциональных пространствах, регуляризация, аппроксимация. — М.: Наука, 1981. — 400 с.
26. Wahba G. Spline models for observational data // CBMS-NSF Reg. Conf. Series in Applied Mathematics. — Philadelphia (PA): SIAM, 1990. — **59.** — 169 p.
27. Keyzer M.A. Rule-based and support vector (SV-) regression/classification algorithms for joint processing of census, map, survey and district data: (Working Paper) / Centre for World Food Studies. — WP-05-01. — Amsterdam, 2005. — 88 p. (http://www.sow.vu.nl/pdf/wp05.01.pdf)
28. Rockafellar R.T., Wets R.J.-B. Variational analysis. — Berlin: Springer, 1998. — 733 p.
29. Bousquet O., Elisseeff A. Stability and generalization // J. Mach. Learn. Res. — 2002. — **2.** — P. 499–526.
30. Smale S., Zhou D.X. Shannon sampling. II: Connections to learning theory // Appl. Comput. Harmon. Anal. — 2005. — **19,** N 3. — P. 285–302.
31. De Vito E., Caponnetto A., Rosasco L. Model selection for regularized least-squares algorithm in learning theory // Found. Comput. Math. — 2005. — **5,** N 1. — P. 59–85.
32. Norkin V.I., Keyzer M.A. On convergence of kernel learning estimators // Proc. of 20th EURO Mini Conf. «Continuous Optimization and Knowledge-Based Technologies» (EUROPT-2008) / L. Sakalauskas, O.W. Weber and E.K. Zavadskas (Eds.). — Vilnius: Inst. of Math. and Inform., 2008. — P. 306–310.
33. Норкин В.И., Кайзер М.А. Об асимптотической эффективности ядерного метода опорных векторов (SVM) // Кибернетика и системный анализ. — 2009. — № 4. — С. 81–97.

Поступила 02.12.2008