УДК 519:234:24:85

В.И. НОРКИН, М.А. КАЙЗЕР

ОБ АСИМПТОТИЧЕСКОЙ ЭФФЕКТИВНОСТИ ЯДЕРНОГО МЕТОДА ОПОРНЫХ ВЕКТОРОВ (SVM)

Ключевые слова: *машинное обучение, восстановление зависимостей, распознавание, ядерные оценки, метод опорных векторов (SVM), некорректные задачи, регуляризация, состоятельность, скорость сходимости*.

ВВЕДЕНИЕ

Важное место в современной теории распознавания образов занимает метод опорных векторов (Support Vector Method/Machine, — SVM), поскольку он успешно конкурирует с наиболее развитыми многоуровневыми нейросетевыми системами распознавания [1, 2]. Его нелинейная версия называется ядерным методом опорных векторов (kernel SVM). Данный метод является составной частью статистической теории обучения, основы которой были заложены в работах [3–5]. Суть теории состоит в следующем [1–11]. Пусть имеется задача обучения, т.е. восстановления зависимости (модели) по наблюдениям «вход-выход». В статистической теории обучения предполагается, что «вход» генерируется случайно, а «выход» наблюдается с ошибками. Таким образом, считается, что наблюдения независимо генерируются из некоторого совместного распределения вероятностей на пространстве «входов-выходов». Задачи распознавания и классификации являются важными специальными случаями задачи восстановления зависимости, когда в последней «выход» дискретен. В статистической теории обучения задачи восстановления зависимости, распознавания и классификации сводятся к двухкритериальным задачам оптимизации на некотором множестве допустимых зависимостей. При этом один критерий (функционал эмпирического риска) отвечает за соответствие модели наблюденным данным, а другой (сложность) — за ее обобщающую способность, т.е. за соответствие модели новым данным. В качестве меры сложности модели используется ее норма как элемента пространства моделей.

Основные задачи данного подхода заключаются в задании пространства и подмножества допустимых моделей, эффективном переборе моделей и нахождении компромисса между степенью соответствия модели данным и сложностью модели. В статистической теории обучения пространство моделей линейно и состоит из линейных или нелинейных (ядерных) функций вектора «входов», причем каждому наблюденному вектору «входов» соответствует некоторая модель из этого пространства. Метод опорных векторов предполагает поиск искомой модели как линейной комбинации только тех моделей пространства, которые соответствуют наблюденным «входам». Вычислительно он сводится к решению конечномерных задач квадратичной оптимизации с числом переменных, равным числу наблюдений.

Математической основой статистической теории обучения и метода опорных векторов являются теории репродуктивных гильбертовых пространств, выпуклых оптимизационных задач в гильбертовых пространствах, некорректных оптимизационных задач и обобщенные законы больших чисел, в частности обобщения теоремы Гливенко–Кантелли и теоремы об экспоненциальной концентрации меры. Метод опорных векторов, когда он применяется к оценке зависимостей, может рассматриваться как один из методов математической статистики: в его линейной версии он близок к методу гребневой регрессии [12], а в нелинейной — к методам непараметрического ядерного оценивания [13].

В настоящей работе изучаются асимптотические свойства SVM-оценок искомой зависимости, полученных методом опорных векторов при неограниченном увеличении числа наблюдений (обучающей выборки), так, как это делается в математической статистике. В литературе по статистической теории обучения в основном исследуется сходимость оценок только по функционалу [1–11, 14–16]. В случае квадратичных функционалов отсюда можно получить и сходимость оценок по вероятности к среднеквадратичной функции регрессии в той или иной норме [8, 13, 16, 17–19]. Отметим, что в методе опорных векторов, как правило, используются неквадратичные и даже негладкие функционалы качества. В данной статье дана оценка скорости сходимости в среднем (пропорциональная $1/\sqrt[4]{m}$, где m — число наблюдений) значений произвольного выпуклого функционала качества SVM-оценок к его теоретическому минимуму; подобная оценка скорости сходимости для доверительной границы квадратичного функционала риска приведена в [16, section 4]. В случае задач бинарной классификации полученные результаты дают оценку скорости сходимости байесовского риска (вероятности ошибочной классификации) к его теоретическому минимуму. Отметим, что при сильном предположении о независимости компонентов входного случайного вектора в [20, 21] для байесовского метода классификации найдена неулучшаемая оценка скорости сходимости, пропорциональная $1/\sqrt{m}$. Исследование сходимости по функционалу оправдано для задач классификации, но недостаточно для рассмотрения задач регрессии, в частности медианной и квантильной регрессии [15, 22, 23]. Поэтому в настоящей статье даются достаточные условия равномерной сходимости SVM-оценок регрессии к искомой зависимости с вероятностью единица, в частности устанавливается соответствующее правило изменения параметра регуляризации в методе опорных векторов с увеличением числа наблюдений. При этом не используются меры мощности класса моделей (типа VC-размерности [1, 4, 5]), которая в рассматриваемом случае может быть бесконечной), а учитывается свойство робастности метода опорных векторов по отношению к отдельным наблюдениям [14] и применяются теоремы об экспоненциальной концентрации распределения усредненных случайных величин вокруг их математического ожидания [24]. Для итеративных алгоритмов обучения результаты о сходимости имеются в [3, 25].

Альтернативный подход к доказательству сходимости с вероятностью единица оценок, полученных методом минимизации эмпирического риска, при условии компактности допустимой области и единственности решения представлен в работах [26, 27], где рассматриваются также случаи периодических, случайно распределенных и зависимых наблюдений. Заметим, что в задачах стохастического программирования и классификации решение, как правило, не единственное, поэтому предположение о единственности решения на практике означает, что применяется фиксированная (не исчезающая с ростом числа наблюдений) регуляризация задачи. Такая регуляризация приводит к асимптотически смещенным оценкам. Кроме того, она дает только слабую компактность множеств уровня целевого функционала. В отличие от [26, 27] в настоящей работе исследуется случай множественных решений и используется постепенно вырождающаяся регуляризация. Результаты данной работы частично представлены в [28, 29].

Изложение организовано следующим образом. В первом разделе дается краткая сводка результатов о репродуктивных гильбертовых пространствах. Во втором показана связь между задачей квантильной регрессии и бинарной классификации с минимизацией выпуклых функционалов риска. В третьем исследуется сходимость метода регуляризации Тихонова для минимизации интегральных функционалов риска в репродуктивных гильбертовых пространствах. В четвертом излагается вычислительная схема, а в пятом изучается сходимость метода опорных векторов. В заключении резюмируются основные результаты работы.

1. РЕПРОДУКТИВНЫЕ ГИЛЬБЕРТОВЫ ПРОСТРАНСТВА

В современной интерпретации метод опорных векторов базируется на теории репродуктивных гильбертовых пространств (Reproducing Kernel Hilbert Spaces, RKHS, — гильбертовы пространства с воспроизводящим ядром), которая возникла в начале XX столетия первоначально для потребностей теории интегральных уравнений и затем проникла во многие разделы математики [2, 8, 30]. Такие пространства являются частным случаем так называемых оснащенных гильбертовых пространств [31].

Определение 1. Гильбертово пространство $H_k(X)$ функций, определенных на замкнутом множестве $X \subseteq R^n$, 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) $f(x) = \langle f, k(\cdot, x)\rangle_k$ для любой функции $f \in H_k(X)$ и любого $x \in X$ (репродуктивное свойство ядра).

Свойство б) показывает, что в скалярных произведениях ядро $k(\cdot, x)$ действует подобно δ-функции Дирака. В статистической теории обучения значение $k(x, \bar{x})$ ядра интерпретируется как мера сходства объектов x и $\bar{x}$. РГП с ядром k обозначается $H_k(X)$, или H_k для краткости. Соответствующее скалярное произведение и норма в РГП обозначаются $\langle\cdot,\cdot\rangle_k$ и $\|\cdot\|_k = \langle\cdot,\cdot\rangle_k^{1/2}$; R^n — n-мерное векторное пространство.

Предложение 1. Множество функций $\left\{ f(x) = \sum_i \alpha_i k(x_i, x) \right\}$ из РГП $H_k(X)$, где $\{x_i\}$ — произвольный набор точек из X, $\{\alpha_i\}$ — произвольный набор чисел, является плотным в $H_k(X)$.

В силу репродуктивного свойства ядра скалярное произведение функций $f(x) = \sum_i \alpha_i k(x_i, x)$, $g(x) = \sum_j \beta_j k(x_i, x)$ из $H_k(X)$ представимо в виде $\langle f, g\rangle_k = \sum_{i,j} \alpha_i \beta_j k(x_i, x_j)$, в частности $\| f \|_k^2 = \langle f, f\rangle_k = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$.

Предложение 2. Гильбертово пространство $H(X)$ функций, определенных на множестве X, является РГП тогда и только тогда, когда для каждого $x \in X$ функционал f_x: $f \to f(x)$ непрерывен по норме $\| f \|_k = \langle f, f\rangle_k^{1/2}$, т.е. существует константа K_x такая, что $|f(x)| \le K_x \| f \|_k \ \ \forall f \in H(X)$.

Следующие утверждения устанавливают связь между нормами $\| f \|_k$ и $\| f \|_\infty = \sup_{x \in X} |f(x)|$.

Предложение 3. Если $\sup_{x \in X} |k(x, x)| \le K^2 < \infty$, то $\| f \|_\infty \le K \| f \|_k$ и, следовательно, сильная сходимость $f^m \to f$ (по H_k-норме) влечет равномерную сходимость $f^m \Rightarrow f$ на X, $m \to \infty$.

Предложение 4. Если ядро удовлетворяет условию $0 < \varepsilon \le k(x, \bar{x}) \le K^2 \ \ \forall x, \bar{x} \in X$ и функция $f = \lim_{n \to \infty} f_n$, где $f_n(x) = \sum_{i=1}^{I_n} \alpha_{ni} k(x, x^{ni})$, $\alpha_{ni} \ge 0$, то $\| f \|_k \le$ $\le (K / \varepsilon) \| f \|_\infty$.

Предложение 5. Если $\sup_{x \in X'} |k(x, x)| \le K < \infty$, то сильная сходимость $f^n \to f$ по норме в РГП $H_k(X)$ влечет равномерную сходимость на $X' \subseteq X$.

Предложение 6. Если ядро $k(x, \bar{x})$ непрерывно по $(x, \bar{x}) \in X \times X$, то соответствующее РГП состоит из непрерывных функций.

Определение 2. Функция $k(x, \bar{x})$ называется ядром Мерсера, если ядро $k(\cdot,\cdot)$ непрерывно и симметрично на компакте $X \times X$ и для любого конечного множества точек $\{x^i \in X\}$ матрица $\{k(x^i, x^j)\}$ является неотрицательно-определенной.

Например, функции $k(x,\bar{x})=\exp(-\|x-\bar{x}\|^2/\sigma^2)$, $k(x,\bar{x})=(\sigma^2+\|x-\bar{x}\|^2)^{-\alpha^2}$, $k(x,\bar{x})=(1+\langle x,\bar{x}\rangle)^l$, где l — целая положительная степень, являются ядрами Мерсера.

Предложение 7 (теорема Мерсера). Ядра Мерсера допускают разложение в равномерно сходящийся ряд $k(x,y)=\sum_{i=1}^{\infty}\lambda_i\varphi_i(x)\varphi_i(y)$ по собственным (непрерывным) функциям $\{\varphi_i(x)\}$ и значениям $\{\lambda_i>0\}$ интегрального оператора

$$A_k f(x)=\int_X f(\bar{x})k(x,\bar{x})d\bar{x}: L_2(X)\to L_2(X),$$

где $L_2(X)$ — гильбертово пространство интегрируемых с квадратом функций на X.

Предложение 8. Для любого ядра Мерсера $k(x,\bar{x})$, $x,\bar{x}\in X$, существует репродуктивное гильбертово пространство $H_k(X)$, задаваемое этим ядром согласно определению 1. По заданному ядру Мерсера $k(x,y)$, $x,y\in X$, соответствующее РГП можно построить двумя способами:

1) $H_k(X)=\mathrm{cl}\left\{f(x)=\sum_{i=1}^{N}\alpha_i k(x_i,x)\ \forall N\ge 1,\alpha_i\in R,x_i\in X\right\}$ со скалярным произведением $\langle f,g\rangle_k=\sum_{i,j=1}^{N}\alpha_i\beta_j k(x_i,x_j)$ функций $f(x)=\sum_{i=1}^{N}\alpha_i k(x,x_i)$ и $g(x)=\sum_{i=1}^{N}\beta_i k(x,x_i)$ из $H_k(X)$;

2) $H_k(X)=\left\{f(x)=\sum_{i=1}^{\infty}a_i\varphi_i(x):\sum_{i=1}^{\infty}a_i^2/\lambda_i<+\infty\right\}$ со скалярным произведением $\langle f,g\rangle_k=\sum_{i=1}^{\infty}a_i b_i/\lambda_i$ функций $f(x)=\sum_{i=1}^{\infty}a_i\varphi_i(x)$ и $g(x)=\sum_{i=1}^{\infty}b_i\varphi_i(x)$, где $\{\varphi_i(x)\}$, $\{\lambda_i\}$ — собственные векторы и значения интегрального оператора A_k.

В статистической теории обучения набор $\{\varphi_i(x)\}$ называется вектором производных характеристик объекта x.

Предложение 9. Конечномерное гильбертово пространство функций $H(X)$ с базисом $\{\varphi_i(x)\}_{i=1}^{N}$ и скалярным произведением $\langle\cdot,\cdot\rangle$, задаваемым матрицей Грамма $\Gamma=\{\gamma_{ij}=\langle\varphi_i,\varphi_j\rangle\}$, может рассматриваться как РГП с ядром $k(x,y)=\sum_{i,j=1}^{N}\delta_{ij}\varphi_i(x)\varphi_j(y)$, где $\{\delta_{ij}\}=\Gamma^{-1}$. Скалярное произведение в этом РГП для функций $f(x)=\sum_{i=1}^{N}a_i\varphi_i(x)$ и $g(x)=\sum_{i=1}^{N}b_i\varphi_i(x)$ определяется как $\langle f,g\rangle_k=\sum_{i,j=1}^{N}\gamma_{ij}a_i b_j$.

2. ЗАДАЧИ СРЕДНЕКВАДРАТИЧНОЙ, МЕДИАННОЙ И КВАНТИЛЬНОЙ РЕГРЕССИИ

Под задачей регрессии (восстановления зависимости между y и x) понимается нахождение определенных (условных, при фиксированном x) характеристик совместного распределения переменных (x,y), например условного среднего, медианы или квантилей. Метод опорных векторов может применяться для решения задач восстановления зависимостей на основе данных наблюдений.

Рассмотрим множество точек $S_m=\{(x_i\in X\subset R^n,\ y_i\in R^1)\}_{i=1}^{m}$, полученных путем наблюдений, которые, предположительно, удовлетворяют соотношению $y=f(x)$, где f — некоторая неизвестная функция. Теория статистического обучения занимается задачей нахождения этой функции на основе собранных наблюдений. Однако, чтобы это было возможно, необходима дополнительная информация о классе функций F, которому принадлежит искомая зависимость f. Например, в теории линейной регрессии предполагается, что f принадлежит линейной оболочке некоторых базисных функций $\{\phi_h(x),h=1,2,\dots,r\}$. Другой подход состоит

в требовании, чтобы f представлялась в наиболее простой форме, например, имела минимальное число ненулевых параметров в линейном представлении. Именно этот путь выбран в той части статистической теории обучения, которая известна как ядерное обучение (kernel learning) [1, 2] и стремится объединить гибкость с простотой представления зависимости f. При этом качество аппроксимации данных измеряется эмпирическим риском, т.е. суммой абсолютных величин отклонений прогнозных значений модели f от наблюденных, а сложность модели оценивается некоторой мерой сложности, например нормой f в рассматриваемом пространстве функций. Оптимальный выбор f осуществляется по этим двум критериям путем минимизации взвешенной суммы эмпирического риска и меры сложности. Критерий, отвечающий за сложность, называется регуляризующим, поскольку он устраняет некорректность задачи, обеспечивает единственность и минимизирует число ненулевых компонентов представления зависимости. Кроме того, статистическая теория обучения линеаризует и, таким образом, упрощает задачу путем отображения исходных данных $\{x_i\}$ из пространства R^n первичных признаков в другое гильбертово пространство вторичных признаков, возможно очень большой или даже бесконечной размерности, такое, что зависимость между зависимой переменной y и вторичными признаками становится линейной и может быть выражена с помощью скалярного произведения. Подобный прием используется в теории обобщенных функций Соболева–Шварца [32], в которой нелинейные функции рассматриваются как линейные функционалы на некотором пространстве основных функций. Таким образом, искомая зависимость ищется в классе функций, которые в расширенном пространстве признаков являются линейными и могут быть заданы в нем вектором. Вычислительные сложности устраняются путем построения такого расширенного пространства признаков, что вычисление скалярного произведения двух векторов очень большой размерности из этого пространства сводится к вычислению значений ядра, т.е. некоторой симметричной функции двух переменных $k(x, y)$. Это и есть репродуктивное гильбертово пространство. Ключевой результат [33] о представлении решения регуляризованной задачи в РГП состоит в том, что решение задачи минимизации регуляризованного эмпирического риска представимо в виде конечной линейной комбинации ядерных функций $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$, привязанных к точкам наблюдения x_i. Следовательно, задача оценивания (или обучения) становится задачей конечномерной оптимизации относительно неизвестных коэффициентов разложения $\{\alpha_i\}_{i=1}^{m}$. В случае кусочно-линейной функции потерь эта задача сводится к решению задачи квадратичной оптимизации, двойственная к которой особенно удобна для решения (см. разд. 4).

Наиболее общей вероятностной моделью зависимости между $y \in Y \subseteq R^1$ и $x \in X \subseteq R^n$ является совместное распределение вероятностей на $X \times Y$, которое можно восстановить из наблюдений $S_m = \{(x_i \in X \subset R^n, y_i \in R^1)\}_{i=1}^{m}$. Однако даже в этой общей постановке интерес представляют только некоторые характеристики распределения, например условное при данном x среднее значение, медиана или квантиль распределения, которые могут быть вычислены, если распределение известно. Условное среднее можно приближенно вычислить по прямым непараметрическим формулам Надараи–Ватсона [34]. В то же время условное среднее может быть получено путем минимизации функционала квадратического риска по всем измеримым зависимостям. Кроме того, рассматриваются неквадратичные функционалы, такие как математическое ожидание абсолютного отклонения, с линейно-квадратичными [35] и ε-нечувствительными функциями потерь [1], которые приводят к робастным оценкам, менее чувствительным к разбросу наблюдений.

Пусть $z = (x, y)$ — векторная случайная величина с некоторым распределением $P(dz)$, сосредоточенным на множестве Z, $c(z, y)$ — выпуклая по второму аргументу функция потерь, $f: X \to R^1$ — функция из некоторого класса F. Определим

функционал риска как математическое ожидание потерь, $R(f)=E_z c(z, f(x))=$ $=\int_Z c(z, f(x))P(dz)$. Задача минимизации риска имеет вид

$$R(f)=E_z c(z, f(x))=\int_Z c(z, f(x))P(dz)\to \inf_{f\in F}. \qquad (1)$$

Множество решений этой задачи обозначим F^*. Для квадратичной функции потерь, $c(z, f)=(y-f)^2$, хорошо известно, что существование условного среднего $\mu(x)=\int_{Y_x} yp(z)dy$ при каждом фиксированном x и $\mu(x)\in F$ влечет $f^*=\mu$ [13].

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. Действительно, для функции потерь вида $c(z, f)=|y-f|=\max\{(f-y),(y-f)\}$ безусловный минимум в (1) функционала $R(f)$ по всем измеримым функциям $f(x)$ достигается на условной медиане распределения $P(dz)$, если она существует. В более общем случае в задаче минимизации риска используется функция потерь

$$c(z, f)=\max\{\mu(f-y), \nu(y-f)\},\ \mu, \nu\geq 0. \qquad (2)$$

Тогда точный минимум (1) является условным квантилем распределения $P(dz)$ уровня $\alpha=\nu/(\mu+\nu)$ (см. [22, 23]; в контексте стохастических минимаксных задач этот факт был установлен в работах [36, с. 95; 37]; он детально обсуждается в [38]). Для медианы имеем $\mu=\nu=1$. Далее формулируются условия, при которых условный α-квантиль распределения является единственным минимумом соответствующего функционала риска.

Предположение 1 (условия идентифицируемости модели):

а) распределение $P(dz)$ имеет плотность $p(z)=p(x, y)$, $z\in Z=$ $=\{z=(x, y): x\in X,\ y\in Y_x\}$ отображение $x\to Y_x$ локально ограничено;

б) x-плотность $p(x)=\int_{Y_x} p(x, y)dy$ невырождена, т.е. $p(x)>0$ для всех $x\in X$;

в) условная функция распределения $F_x(a)=\int_{-\infty}^{a} p_x(y)dy$, где $p_x(y)=$ $p(x, y)/p(x)$, непрерывна по (x, a) и строго монотонна по a для a таких, что $0<F_x(a)<1$.

Условие 1а гарантирует существование моментов распределения; 1б требует, чтобы x-значения заполняли все множество X; 1в гарантирует, что квантили функции распределения $F_x(\cdot)$ и условное среднее $\mu(x)=\int_{Y_x} yp(x, y)dy$ — непрерывные функции x. Теперь можно доказать следующее утверждение.

Теорема 1. Пусть выполнены условия предположения 1. Рассмотрим задачу минимизации риска

$$R(f)=\int_Z \max\{\mu(f(x)-y), \nu(y-f(x))\}p(z)dz\to \inf_f \qquad (3)$$

на классе функций, которые непрерывны и интегрируемы по мере с плотностью $p(x)$ на множестве X. Тогда для данного x минимум в (3) достигается на квантиле $q_\alpha(x)$ уровня $\alpha=\nu/(\mu+\nu)$, $\mu, \nu>0$, и этот минимум единственный.

Доказательство. Рассмотрим условную функцию риска

$$\varphi(x, a)=E_{y|x}\max\{\mu(a-y), \nu(y-a)\}=\int_{Y_x}\max\{\mu(a-y), \nu(y-a)\}dF_x(y). \qquad (4)$$

Известно, что безусловный минимум функции $\varphi(x, a)$ по a достигается на условном квантиле распределения $P(dz)$ [15, 22, 23, 36, 37, 38]. Приводимое доказательство направлено на то, чтобы показать, что в сделанных предположениях этот квантиль является непрерывной функцией от x и единственным непрерывным минимумом функционала (3).

В силу предположения 1а функция $\varphi(x, a)$ определена для всех $a \in R$ и $x \in X$. Отметим, что $\varphi(x, a)$ равномерно непрерывна (даже равномерно липшицева) по a и в силу предположений 1а, 1в непрерывна по x (по теореме Хелли о предельном переходе под знаком интеграла Лебега–Стильтьеса [32]), следовательно, она непрерывна по $(x, a) \in X \times R$. Для любой измеримой функции f имеем

$$R(f) = E_{x,\,y} \max\{\mu(f(x) - y), \nu(y - f(x))\} =$$

$$= E_x E_{y|x} \max\{\mu(f(x) - y), \nu(y - f(x))\} = E_x \varphi(x, f(x)) \geq E_x \inf_a \varphi(x, a).$$

Согласно [36, 37] представим условную функцию риска следующим образом:

$$\varphi(x, a) = (\mu + \nu) \int_{-\infty}^{a} dF_x(y) - \nu a + \nu \int_{-\infty}^{+\infty} y dF_x(y).$$

В силу предположения 1в отсюда вытекает, что функция $\varphi(x, a)$ дифференцируема по a с производной $\varphi_a(x, a) = (\mu + \nu) F_x(a) - \nu$. Так как $\mu, \nu > 0$ по предположению, квантиль $q(x)$ уровня $\alpha = \nu / (\mu + \nu)$ функции распределения $F_x(a)$ является единственным корнем по переменной a уравнения $\varphi_a(x, a) = 0$. Отметим также, что в силу предположения 1в функция $\varphi_a(x, a) = (\mu + \nu) F_x(a) - \nu$ непрерывна по (x, a) и строго монотонна по a; следовательно, в силу классической теоремы о неявной функции корень $q(x)$ уравнения $\varphi_a(x, a) = 0$ является непрерывной функцией x. Очевидно, поскольку производная $\varphi_a(x, a)$ неотрицательна при $a < q(x)$ и положительна при $a > q(x)$, а $F_x(\cdot)$ строго монотонна, функция $\varphi(x, \cdot)$ при фиксированном x достигает своего минимума в точке $q(x)$. Таким образом, для любой измеримой функции f имеем

$$\inf_f R(f) \geq E_x \inf_a \varphi(x, a) = E_x \varphi(x, q(x)) =$$

$$= E_{x,y} \max\{\mu(q(x) - y), \nu(y - q(x))\} = V(\mu, \nu)$$

и инфинум достигается при $f(x) = q(x)$. Однако остается еще вопрос, является ли минимальная функция единственной.

Так как функция $\varphi(x, a)$ непрерывна по (x, a), суперпозиции $\varphi(x, q(x))$ и $\Delta(x) = \varphi(x, f(x))$ непрерывны по x для непрерывных $f(x)$. В силу оптимальности $q(x)$ имеем $\varphi(x, f(x)) \geq \varphi(x, q(x))$ для любого $x \in X$. Теперь предположим, что $f(x') \neq q(x')$ в некоторой точке $x' \in X$. Тогда ввиду единственности минимума $q(x)$ функции $\varphi(x, a)$ по a имеем $\varphi(x', f(x')) > \varphi(x', q(x'))$. Благодаря непрерывности $\Delta(x)$ и $\varphi(x, q(x))$ это неравенство имеет место не только в точке x', но и в некоторой окрестности точки x', и в силу невырожденности плотности $p(x) = \int_{Y_x} p(x, y) dy$ получаем

$$R(f) = E_x \varphi(x, f(x)) > E_x \varphi(x, q(x)) = R(q).$$

Это доказывает, что $q(x)$ — единственный непрерывный минимум функционала $R(\cdot)$.

Следствие 1. Пусть выполнены условия предположения 1. Рассмотрим задачу минимизации риска

$$R(f) = E_{x,\,y} |f(x) - y| \to \inf_f \tag{5}$$

на множестве функций, непрерывных на X. Тогда для данного x минимум достигается на медиане $m(x) = q_{1/2}(x)$ и это решение единственно.

Более общая, чем, например, оценивание медианы или среднего, постановка задачи оценивания квантилей может быть полезной потому, что кроме оценивания некоторой средней тенденции обычно представляет интерес установление области значимости, в которую прогнозные значения попадают с заданной вероятностью. Другими словами, исследователь также стремится оценить ошибки $\Delta(x) = E_{y|x} |y - m(x)|$, которые являются компонентами ожидаемого риска

$$R(m) = E_{x,\,y} |y - m(x)| = E_x E_{y|x} |y - m(x)| = E_x \Delta(x) = \int_X \Delta(x) p(x) dx.$$

Для этого строится регрессия $\Delta(x)$ на основе наблюдений (x_i, η_i), где $\eta_i = |y_i - f(x_i)|$. Альтернативный способ описания ошибок состоит в последовательном или одновременном построении ряда условных квантильных функций $q_\alpha(x)$ для уровней α, меняющихся от 0 до 1, медиане соответствует уровень 0,5; этот подход обсуждается в [15, 22]. Таким образом, оценка квантилей для различных доверительных уровней дает возможность построить пакет статистик, например, медиану и доверительный интервал, 5 %-й квантиль слева и 95 %-й квантиль справа, совместно дающие 90 %-ю доверительную область для прогноза, когда сам прогноз проводится, скажем, по медиане. Для уровней, близких к нулю или единице, оценка квантилей сводится к оценке верхней и нижней границ распределения.

Заметим, что функционалы риска (3), (5) штрафуют все ошибки, большие и маленькие. В статистических методах обучения, методе опорных векторов для регрессии и классификации, в частности, маленькие отклонения зачастую не штрафуются, а рассматривается возмущенная задача минимизации риска, т.е. используются ε-нечувствительные функции потерь, такие как

$$c_\varepsilon(z, f(x)) = \max\{0, \mu(f(x) - y) - \varepsilon, \nu(y - f(x)) - \varepsilon\}, \qquad (6)$$

и соответствующий ε-нечувствительный функционал риска $R_\varepsilon(f) = E_{x,y} c_\varepsilon(z, f(x))$. Очевидно, что $R_\varepsilon(f)$ равномерно сходится к $R(f)$, когда $\varepsilon \to 0$, но этого еще недостаточно для доказательства равномерной сходимости минимумов $f_\varepsilon \in \in \arg\min_f R_\varepsilon(f)$ к $q = \arg\min_f R(f)$, потому что риск, являющийся интегралом, нечувствителен к изменению подынтегральной функции на произвольно малой области. Действительно, при таких возмущениях можно только установить сходимость по мере P_x к этому квантилю при $\varepsilon \to 0$.

Теорема 2 (об идентификации квантиля посредством ε-нечувствительных решений при $\varepsilon \to 0$). Путь f_ε — любой минимум $R_\varepsilon(\cdot)$ на множестве измеримых функций, определенных на замкнутом множестве $X \subseteq R^n$. Тогда при условиях предположения 1 f_ε сходятся к квантилю $q(x)$ по мерее P_x, генерируемой плотностью $p(x) = \int_{Y_x} p(x, y)\, dy$, т.е. для любого $\delta > 0$ имеем

$$P_x\{|q(x) - f_\varepsilon(x)| \ge \delta\} \to 0 \text{ при } \varepsilon \to 0.$$

Доказательство. Легко видеть, что $c(z, f(x)) - \varepsilon \le c_\varepsilon(z, f(x)) \le c(z, f(x))$ и, значит, $R(f) - \varepsilon \le R_\varepsilon(f) \le R(f)$ равномерно по всем измеримым функциям f. Следовательно,

$$R(q) - \varepsilon \le R(f_\varepsilon) - \varepsilon \le R_\varepsilon(f_\varepsilon) \le R_\varepsilon(q) \le R(q),$$

т.е. $R(q) \le R(f_\varepsilon) \le R(q) + \varepsilon$ и $\lim_{\varepsilon \to 0} R(f_\varepsilon) = R(q)$. Проведем доказательство от противного. Предположим, что существуют константы $\delta > 0$ и $\gamma > 0$, для которых последовательность $\varepsilon_m \to 0$ такова, что $P_x\{|q(x) - f_{\varepsilon_m}(x)| \ge \delta\} \ge \gamma > 0$. Вероятностная мера P_x на замкнутом множестве $X \subseteq R^n$ является плотной, поэтому существует компактное подмножество $X' \subseteq X$ такое, что $P_x\{X'\} \ge 1 - \gamma/2$. Тогда $P_x\{x \in X' : |q(x) - f_{\varepsilon_m}(x)| \ge \delta\} \ge \gamma/2 > 0$. Рассмотрим функцию $\varphi(x, a)$, определенную формулой (4). Как было показано при доказательстве теоремы 1, при предположении 1 функция $\varphi(x, a)$ непрерывна по (x, a), имеет единственный минимум $q(x)$ по a, который является непрерывной функцией переменной x. Поэтому суперпозиция $\Delta(x) = \varphi(x, q(x))$ также непрерывна. Обозначим $\bar{y}(x) = \int_{Y_x} y\, dF_x(y)$ среднеквадратичную функцию регрессии. При предположениях 1а, 1в функция $\bar{y}(x)$ непрерывна по $x \in X$ в силу теоремы Хелли о предельном переходе под знаком интеграла Лебега–Стильтьеса [32]. В силу неравенства Иенсена $\varphi(x, a) \ge \max\{\mu(a - \bar{y}(x)), \nu(\bar{y}(x) - a)\}$. Обозначим $\bar{y}_{\min} = \min_{x \in X'} \bar{y}(x)$ и

$\bar{y}_{\max} = \max_{x \in X'} \bar{y}(x)$. Очевидно, $\varphi(x, a) \geq \max\{\mu(a - \bar{y}_{\max}), \nu(\bar{y}_{\min} - a)\}$ для всех $x \in X'$ и $a \in R^1$. Тогда непрерывная функция $\varphi(x, a)$ достигает своего минимума на множестве $\{(x, a): x \in X', a \in R^1, |q(x) - a| \geq \delta\}$ и в силу единственности минимума $q(x) = \arg\min_{a \in R^1} \varphi(x, a)$ имеет место

$$\inf_{\{(x, a): x \in X', a \in R^1\}} \{\varphi(x, a) - \varphi(x, q(x)): |q(x) - a| \geq \delta\} \geq \nu > 0.$$

Следовательно, $\varphi(x, f_{\varepsilon_m}(x)) \geq \varphi(x, q(x))$ для всех $x \in X'$ и $\varphi(x, f_{\varepsilon_m}(x)) \geq$ $\geq \varphi(x, q(x)) + \nu$ на некотором подмножестве положительной меры $\gamma / 2$, поэтому $R(f_{\varepsilon_m}) = E_x \varphi(x, f_{\varepsilon_m}(x)) \geq E_x \varphi(x, q(x)) + \nu\gamma / 2$. Однако это противоречит свойству $\lim_{m \to \infty} R(f_{\varepsilon_m}) = R(q)$.

Теорема доказана.

Данная теорема допускает, что при исчезновении возмущений при $\varepsilon \to 0$ функция f_ε может отклоняться от точного минимума f^* на бесконечном множестве точек и не на малую величину, хотя и на множестве малой меры. Таким образом, f_ε может быть нестабильной при изменении ε и не сходиться к f^* равномерно.

Задачи классификации тесно связаны с оптимизацией функционалов риска. Под задачей классификации понимается построение классификатора, на котором байесовский риск (вероятность ошибочной классификации) принимает минимальное значение. Для любой измеримой функции $f(x)$ бинарное классифицирующее правило определяется по формуле

$$I_{1/2}(f(x)) = \begin{cases} 1, & f(x) > 1/2, \\ 0 & \text{противном случае.} \end{cases} \tag{7}$$

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 $y \in \{0, 1\}$. Заметим, что байесовский риск достигает минимального значения P^* на решающем правиле g_η, задаваемом функцией условной вероятности $\eta(x) = P\{y = 1 | x\}$ [39, с. 10], но она неизвестна. Таким образом, один возможный путь построения оптимальных классификаторов состоит в аппроксимации условной вероятности $\eta(x) = P\{y = 1 | x\}$. Например, в [20, 21] условная вероятность $\eta(x)$ аппроксимируется байесовским методом при (сильном) предположении независимости компонент случайного входного вектора наблюдений.

Известно ([39, с. 16] и ссылки в этой работе), что

$$P\{I_{1/2}(f(x)) \neq y\} - \min_{f \in F} P\{I_{1/2}(f(x)) \neq y\} \leq 2E|f(x) - \eta(x)|$$

и $\eta(x) = \arg\min_f E(y - f(x))^2$. Если $\tilde{f}(x)$ — некоторое приближенное решение задачи минимизации квадратичного риска $R(f) = E(f(x) - y)^2$, то соответствующее классифицирующее правило определяется по формуле (7).

Справедлива оценка [39, с. 20]

$$P\{I_{1/2}(f(x)) \neq y\} - \min_{f \in F} P\{I_{1/2}(f(x)) \neq y\} \leq 2(R(f) - \min_{f \in F} R(f)), \tag{8}$$

где минимумы берутся по множеству измеримых функций. Таким образом, минимизация функционала $R(f) = E|f(x) - y|$ по множеству измеримых функций F в силу (8) автоматически ведет к минимизации функционала байесовского риска. В силу следствия 1 минимум функционала $R(f) = E|f(x) - y|$ достигается на условных медианах распределения $P(\cdot)$. Если есть основания полагать, что условные медианы распределения $P(\cdot)$ принадлежат некоторому классу функций, например некоторому репродуктивному гильбертову пространству H_k, то в (8) можно положить $F = H_k$.

3. РЕГУЛЯРИЗАЦИЯ ЗАДАЧ ОПТИМИЗАЦИИ РИСКА В РЕПРОДУКТИВНОМ ГИЛЬБЕРТОВОМ ПРОСТРАНСТВЕ

Оптимизация интегрального функционала риска (1) осложняется следующими обстоятельствами:

1) задача может быть некорректной, иметь не единственное решение, ее приближенные решения могут быть численно неустойчивыми;

2) задача является бесконечномерной, поэтому необходима та или иная ее аппроксимация;

3) оценка значений функционала связана с многократным вычислением интегралов.

Рассмотрим первую проблему, которая обычно решается методом регуляризации Тихонова. Метод регуляризации для решения операторных уравнений детально представлен в [40], для бесконечномерных задач оптимизации — в [41], а для задач статистического оценивания — в [1, 5]. Необходимо адаптировать теорию метода регуляризации к задачам оптимизации выпуклых интегральных функционалов риска в РГП. Вторая и третья проблемы будут детально рассмотрены в следующем разделе, где для аппроксимации задачи и интегралов применяется метод эмпирических аппроксимаций.

Наряду с (1) рассмотрим регуляризованную задачу

$$R^{\lambda}(f) = R(f) + \lambda\,\Omega(\|f\|) = E_{x,y}c(z, f(x)) + \lambda\Omega(\|f\|) \to \inf_{f\in F}, \qquad (9)$$

где $R(f) = E_{x,y}c(z, f(x))$, $c(z, f): Z \times R^1 \to R^1$, $\lambda > 0$ — параметр регуляризации, $\Omega(\cdot)$ — неотрицательная функция одной переменной, F — множество в некотором гильбертовом пространстве H с нормой $\|\cdot\|$. Кроме (9) рассмотрим также возмущенную задачу

$$R^{\lambda}(f) = \widetilde{R}(f) + \lambda\,\Omega(\|f\|) \to \inf_{f\in F}, \qquad (10)$$

где $\widetilde{R}(f)$ — некоторая аппроксимация или возмущение функционала $R(f)$. Выведем условия сходимости решений f^{λ} задачи (9) и решений $\widetilde{f}^{\lambda}$ задачи (10) к решению f^* исходной задачи (1), когда параметр регуляризации стремится к нулю, $\lambda \to 0$.

Предположение 2 (об исходной, возмущенной и регуляризованной задачах):

а) функционал $R(f)$ ограничен снизу, является выпуклым и непрерывным на $F \subseteq H$;

б) F — замкнутое выпуклое множество в гильбертовом пространстве H;

в) множество $F^* = \{f \in H: R(f) = \inf_{f\in F} R(f)\}$ не пусто;

г) $\Omega(f) = \|f\|^2$;

д) $\sup_{f\in D} |\widetilde{R}(f) - R(f)| \le \delta_{\lambda}$, $\lim_{\lambda\to+0} \delta_{\lambda}/\lambda = 0$.

Предположения 2 являются общими и не связаны со спецификой задач обучения.

Напомним один общий результат о сходимости для семейства возмущенных задач (10).

Теорема 3 [40, гл. VIII, § 1; 41, гл. 2, § 5, теорема 2]. В условиях предположения 2 решения (если они существуют) $\widetilde{f}^{\lambda}$ возмущенной регуляризованной задачи сходятся к нормальному (минимальному по норме) решению f^* исходной задачи при $\lambda \to +0$ в сильной топологии пространства H (по норме $\|\cdot\|$).

Сформулируем результат о сходимости метода регуляризации для задач минимизации функционалов риска вида (1).

Предположение 3 (о задаче минимизации риска). Функция потерь $c(z, f(x))$ в (1) удовлетворяет условиям:

а) $c(z,a) \geq c_0 > -\infty$ для всех $z \in Z$, $a \in R^1$;

б) $c(z,a)$ выпукла по a;

в) $|c(z,a)| \leq c_1(z) + c_2(z)\rho(|a|)$, $\int_Z c_1(z)p(z)dz < +\infty$, $\int_Z c_2(z)p(z)dz < +\infty$,

где $\rho(\cdot)$ — монотонно возрастающая функция;

г) $H(X) = H_k(X)$ — РГП с ограниченным на диагонали ядром, $\sup_{x \in X} |k(x,x)| = K^2 < +\infty$.

Теорема 4 (о свойствах функционала риска). В предположениях 3 функционал риска $R(f) = E_z c(z, f(x))$, определенный на репродуктивном гильбертовом пространстве $H_k(X)$ с ограниченным на диагонали ядром, является выпуклым и непрерывным по норме $\|\cdot\|_k$ этого пространства.

Доказательство. Ограниченность снизу и выпуклость функционала риска следуют из предположений 3а, 3б. В бесконечномерном пространстве из выпуклости функционала вытекает только его полунерерывность снизу [42]. При сделанных предположениях докажем его непрерывность с помощью теоремы Лебега о предельном переходе под знаком интеграла. Пусть $\lim_{s \to \infty} \| f^s - f \|_k = 0$ и, значит, $\lim_{s \to \infty} \| f^s \|_k = \| f \|_k$. Тогда $\{f^s\}$ сходится к f равномерно. Действительно, для любого $x \in X$ в силу воспроизводящего свойства ядра (см. определение 1б) и неравенства Коши–Шварца имеем

$$|f^s(x) - f(x)| \leq \sqrt{k(x,x)} \| f^s - f \|_k \leq K \| f^s - f \|_k \to 0 \ , \ s \to \infty .$$

В силу выпуклости функция $c(z,\cdot)$ непрерывна, следовательно, $\lim_{s \to \infty} c(z, f^s(x)) = c(z, f(x))$ для любого $z \in Z$. Из предположений 3в, 3г получаем

$$|c(z, f^s(x))| \leq c_1(z) + c_2(z)\rho(|f^s(x)|) \leq c_1(z) + c_2(z)\rho(|k(x,x)| \, \| f^s \|_k \,) \leq$$

$$\leq c_1(z) + c_2(z)\rho(K \sup_s \| f^s \|_k \,) = U(z).$$

В силу предположения 3в мажоранта $U(z)$ интегрируема. Теперь непрерывность функционала $R(f) = E_z c(z, f(x))$ следует из теоремы Лебега о предельном переходе под знаком интеграла.

Теорема 5 (о равномерной сходимости регуляризованных решений при $\lambda \to 0$). При предположениях 2б–2д и 3 решения (если они существуют) f^λ, $\tilde{f}^\lambda$ регуляризованных задач (9), (10) с функционалом риска $R(f) = E_z c(z, f(x))$ и $F \subseteq H_k(X)$ равномерно сходятся при $\lambda \to 0$ к нормальному решению f^* (минимальному по норме $\|\cdot\|_k$ из множества решений F^* исходной задачи (1)).

Доказательство. В предположениях 3 функционал риска $R(f) = E_z c(z, f(x))$ удовлетворяет условию 2а в силу теоремы 4. Тогда по теореме 3 регуляризованные решения сильно (по норме) сходятся к нормальному решению исходной задачи, $\lim_{\lambda \to +0} \| \tilde{f}^\lambda - f^* \| = 0$, но такая сходимость может быть неравномерной. Однако в репродуктивном гильбертовом пространстве $H_k(X)$ с ограниченным на диагонали ядром k (предположение 3г) сильная сходимость влечет равномерную, $\lim_{\lambda \to +0} \sup_{x \in X} |\tilde{f}^\lambda(x) - f^*(x)| = 0$. Результат о сходимости для f^λ следует из доказанного при $\delta_\lambda = 0$.

Теорема доказана.

Регуляризация делает решения более стабильными. Заметим, что вместо $\| f \|^2$ могут использоваться другие регуляризующие функционалы, например $\Omega(\| f \|) = \| f - f^0 \|^2$, где f^0 — некоторая функция отсчета. Путем трансляционной замены $f = \bar{f} + f^0$ задача сводится к однородной форме (9), (10), и, таким обра-

зом, к ней применимы доказанные результаты о сходимости. Более того, точки отсчета могут меняться от вычислений к вычислениям (при переходе от одних λ к другим), как это делается в методах последовательной регуляризации (см., например, [43]).

4. ОПТИМИЗАЦИЯ РЕГУЛЯРИЗОВАННЫХ ФУНКЦИОНАЛОВ ЭМПИРИЧЕСКОГО РИСКА: МЕТОД ОПОРНЫХ ВЕКТОРОВ

Рассмотрим возможности аппроксимации бесконечномерной задачи оптимизации функционалов риска и входящих в них интегралов по вероятностной мере $P(\cdot)$. На практике распределение $P(\cdot)$ обычно не известно полностью, а имеется набор независимых наблюдений $\{z_i, i = 1,\dots, m\}$ векторной случайной величины z 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 $R(f) = Ec(z, f(x))$ — эмпирическими средними (эмпирическим риском) $\widetilde{R}_m(f) = (1/m)\sum_{i=1}^{m} c(z_i, f(x_i))$. Таким способом решается проблема оценки интегралов. Однако задача минимизации эмпирического риска $\widetilde{R}_m(f)$ по-прежнему бесконечномерна, поскольку множество F допустимых зависимостей бесконечномерно, например, является подмножеством некоторого гильбертова или иного линейного пространства функций H. В классическом регрессионном анализе предполагается, что H — линейная оболочка конечного множества функций $\{\varphi_i(x), i = 1,\dots, n\}$, и чем больше этих функций, тем богаче класс H и больше возможностей точнее аппроксимировать искомую неизвестную зависимость. В непараметрической статистике используются ядерные порождающие функции, например $\{k((x - x_i)/\theta)\}$, где k — некоторая фиксированная функция, $\{x_i\}$ — произвольные точки множества X, θ — числовой параметр. Наибольшая гибкость достигается, когда порождающее множество функций бесконечно. Однако необходимо соблюдать компромисс между степенью аппроксимации данных наблюдений, которая измеряется величиной эмпирического риска $\widetilde{R}_m(f)$, и сложностью аппроксимирующей модели $f(x)$, поскольку из-за излишне точной подгонки модели под имеющуюся случайную обучающую выборку $\{z_i, i = 1,\dots, m\}$ она может потерять обобщающую способность, т.е. будет плохо работать на другом наборе данных. Таким образом, модель должна хорошо аппроксимировать данные и быть, по возможности, наиболее простой. В качестве меры сложности зависимости f может использоваться, например, число ненулевых элементов разложения f по базису пространства, сумма абсолютных величин коэффициентов разложения или сумма их квадратов. Для ортонормированного базиса сумма квадратов коэффициентов разложения — это норма элемента f из гильбертова пространства H, поэтому обобщенной мерой сложности зависимости f можно считать норму $\|f\|$ или некоторую функцию $\Omega(\|f\|)$ от нормы. Таким образом, задача выбора модели двухкритериальна с критериями $\widetilde{R}_m(f)$ и $\|f\|$. Эти критерии свертываются в один $\widetilde{R}_m(f) + \lambda\Omega(\|f\|)$ с коэффициентом λ, выбор значения которого является важнейшей проблемой.

Возможен и другой формально математический взгляд на проблему восстановления зависимости f по экспериментальным данным $\{z_i, i = 1,\dots, m\}$ с позиций теории стохастических некорректных задач. Исходная задача минимизации функционала риска , вообще говоря, может быть некорректной, т.е. иметь неоднозначные решения, быть неустойчивой по отношению к возмущениям функционала. 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 $\delta_m(f) = \widetilde{R}_m(f) - R(f)$. Для нахождения приближенных решений применяется ме-

тод регуляризации А.Н. Тихонова в функциональном (гильбертовом) пространстве H. Метод регуляризации с детерминированными возмущениями хорошо изучен [40, 41]. Метод регуляризации со случайными возмущениями исследован значительно меньше, некоторые результаты для линейных операторных уравнений (квадратичных оптимизационных задач) в рефлексивном банаховом пространстве имеются в [1, 5], а для оснащенных гильбертовых пространств с общими случайными возмущениями — в [44] (в этих работах можно найти дальнейшие ссылки).

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

$$\widetilde{R}_m(f)+\lambda\Omega_k(f)=\frac{1}{m}\sum_{i=1}^{m}c(z_i,f(x_i))+\lambda\,\|f\|_k^2\to\inf_{f\in H_k},\qquad(11)$$

где H_k — некоторое РГП, порожденное ядром k. Оказывается, что решение регуляризованной задачи (11) 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 [2, Theorem 4.2, p. 90; 33] в РГП решение задачи существует и может быть представлено в виде

$$f_m^{\lambda}(x)=\sum_{i=1}^{m}\alpha_i k(x_i,x),\qquad(12)$$

где $\alpha^m=\{\alpha_i\}$ — набор действительных чисел. Подставляя выражение (12) в (11) и используя репродуктивное свойство ядра (см. определение 1б), приходим к следующей конечномерной задаче оптимизации:

$$R_m(\alpha^m)=\frac{1}{m}\sum_{i=1}^{m}c\left(z_i,\sum_{j=1}^{m}\alpha_j k(x_i,x_j)\right)+\lambda\sum_{i,j=1}^{m}\alpha_i\alpha_j k(x_i,x_j)\to\min_{\alpha^m}.\qquad(13)$$

Если функция потерь $c(z,f)$ непрерывна и неотрицательна, а матрица $\{k(x_i,x_j)\}$ положительно определена, то эта задача имеет единственное решение f_m^{λ}. Более того, для квантильных функций потерь (2), (6), задача (13) является выпуклой и негладкой, однако с помощью дополнительных переменных $\xi^m=\{\xi_i\}_{i=1}^m$ (ошибок регрессии на наблюдениях) она сводится к задаче квадратичного программирования при линейных ограничениях

$$\mathrm{R_m}(\alpha^{\mathrm m},\xi^{\mathrm m})=\frac{1}{\mathrm m}\sum_{\mathrm i=1}^{\mathrm m}\xi_{\mathrm i}+\lambda\sum_{\mathrm{i,j=1}}^{\mathrm m}\alpha_{\mathrm i}\alpha_{\mathrm j}\mathrm{k(x_i,x_j)}\to\min_{\xi^{\mathrm m}\ge 0,\alpha^{\mathrm m}},$$

$$\xi_i\ge\nu\,y_i-\nu\sum_{j=1}^{m}\alpha_j k(x_j,x_i)-\varepsilon,\qquad(14)$$

$$\xi_i\ge\mu\sum_{j=1}^{m}\alpha_j k(x_j,x_i)-\mu\,y_i-\varepsilon,\ \ i=1,\dots,m.$$

В [2, 28] рассматриваются модификации задачи (14), в которых параметр нечувствительности $\varepsilon\ge 0$ — переменная величина. В этом случае в целевую функцию задачи добавляется некоторый штраф $\theta(\varepsilon)\ge 0$, $\theta(0)=0$, монотонно возрастающий по $\varepsilon\ge 0$.

В некоторых приложениях желательно, чтобы искомая функция регрессии обладала некоторыми заранее известными теоретическими свойствами (монотонности, выпуклости или вогнутости, инвариантности по отношению к некоторым пре-

образованиям и т.п.). Пусть множество G таких функций можно представить в виде $g(x)=\sum_{h=1}^{r}\gamma_h\phi_h(x)$ линейных комбинаций базисных функций $\{\phi_h(x)\}$ с коэффициентами $\gamma^r=(\gamma_1,\ldots,\gamma_r)\in\Gamma$. Тогда искомая зависимость $f(x)$, с одной стороны, должна минимизировать эмпирический риск, а с другой — быть наиболее близка к множеству G. Соответствующая задача оптимизации принимает вид (см., например, [2, 9, 15])

$$\frac{1}{m}\sum_{i=1}^{m}c(z_i,f(x_i))+\lambda\left\|f-\sum_{h=1}^{r}\gamma_h\phi_h(x)\right\|_k^2\to\inf_{f\in H_k,\gamma^r\in\Gamma},\qquad(15)$$

где λ — весовой коэффициент. В этом случае решение $f_m^\lambda(x)$ задачи (15) представимо в виде [2, p. 91]

$$f_m^\lambda(x)=\sum_{i=1}^{m}\alpha_i k(x_i,x)+\sum_{h=1}^{r}\gamma_h\phi_h(x).\qquad(16)$$

Подстановка выражения (16) в (15) снова приводит к задаче конечномерной оптимизации относительно неизвестных коэффициентов разложения $\{\alpha_i\}$ и $\{\gamma_h\}$, которая, в свою очередь, для кусочно-линейных функций потерь сводится к задаче квадратичной оптимизации.

5. СХОДИМОСТЬ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

Рассмотрим асимптотические свойства при $m\to\infty$ и $\lambda\to 0$ эмпирических ядерных оценок $f_m^\lambda(x)$, которые являются решениями задачи минимизации регуляризованного эмпирического риска (11). В основополагающих работах [1, 4, 5] исследован вопрос сходимости $R(f_m^\lambda)\to\inf_{f\in F}R(f)$ в предположении ограниченной емкости класса функций F. Этод 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 $R(f)=Ec(z,f(x))$. Однако не всегда подходящий класс функций имеет конечную емкость (конечную размерность в смысле Вапника–Червоненкиса). К тому же условие равномерной сходимости аппроксимаций $R_m^\lambda(f)$ к $R(f)$ не является необходимым для сходимости минимумов [45]. Поэтому следуем другому подходу [14], основанному на использовании свойств устойчивости оценок $f_m^\lambda(x)$ по отношению к отдельным наблюдениям. Подобный подход использовался в [2, Section 12.1; 15, 16, 19], где исследовалась сходимость оценок по вероятности. В отличие от этих работ установим 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 $\|f^*\|_k$.

Сделаем дополнительные специфические предположения относительно функции потерь $c(\cdot,\cdot)$.

Предположение 4 (дополнительные свойства функции потерь):

а) функция потерь $c(z,\cdot)$ выпукла и липшицева по второму аргументу с константой L равномерно по $z\in Z$;

б) $\sup_{z\in Z}c(z,0)=C<+\infty$.

Следующая теорема дает оценку неоптимальности (в среднем) SVM-оценок f_m^λ искомой функциональной зависимости как функцию m и λ. Эти оценки являются случайными величинами со значениями в пространстве непрерывных функ-

ций и определены на счетном произведении исходного вероятностного пространства (X, B_X, P).

Теорема 6 [46]. Пусть решение задачи (1) существует, функции f_m^λ являются решениями задачи (11). Тогда в предположениях 3, 4 для любого $\lambda > 0$ и m имеет место оценка

$$E_m R(f_m^\lambda) \le R(f^*) + 2\frac{2C + L\| f^* \|_\infty}{\sqrt{m}} + \frac{LK(5LK + 2\sqrt{\lambda C})}{\lambda\sqrt{m}} + \lambda \| f^* \|_k^2, \quad (17)$$

где математическое ожидание E_m берется по всем выборкам $\{\omega^1, \ldots, \omega^m\}$ с независимыми одинаково распределенными наблюдениями, f^* — любое решение задачи (1).

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 $m \to \infty$.

Установим 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 3 [40]. Решение $f^* \in F^*$ задачи (1) называется нормальным, если оно имеет минимальную норму, $\| f^* \|_k = \min_{f \in F^*} \| f \|_k$.

Следующие две теоремы дают достаточные условия равномерной сходимости с вероятностью единица SVM-оценок $f_m^{\lambda(m)}$ к нормальному решению $f^* \in F^*$ задачи (1), т.е. $\lim_{m\to\infty} \sup_{\omega\in\Omega} \left| f_m^{\lambda(m)}(\omega) - f^*(\omega) \right| = 0$.

Теорема 7 (достаточные условия сильной состоятельности SVM-оценок) [46]. Пусть решение задачи (1) существует и выполнены предположения 3, 4. Рассмотрим семейство решений $f_m^{\lambda(m)}$ задачи (11), причем $\lim_{m\to\infty} \lambda(m) = 0$. Тогда 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 (11) равномерно по $x \in X$ сходятся к нормальному решению f^* задачи (1) с вероятностью единица при $m \to +\infty$.

Теорема 8 (оценка скорости сходимости SVM-оценок) [46]. 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 7 и имеет место оценка

$$E_m R(f_m^{\lambda(m)}) - R(f^*) \le 2\frac{2C + L\| f^* \|_\infty}{\sqrt{m}} + \\ + \frac{LK(5LK + 2\sqrt{2\Lambda C})}{\Lambda(\ln m)^\varepsilon \sqrt[4]{m}} + \frac{\| f^* \|_k^2 \Lambda(\ln m)^\varepsilon}{\sqrt[4]{m}}. \quad (18)$$

ЗАКЛЮЧЕНИЕ

В настоящей работе установлена скорость сходимости метода опорных векторов (SVM) В.Н. Вапника, а именно показано, что значения функционала риска на SVM-оценках сходятся к минимуму в среднем пропорционально $1/\sqrt[4]{m}$, где m — число элементов в обучающей выборке. Доказано, что метод опорных векторов является асимптотически устойчивым, т.е. он сходится к некоторому (нормальному) решению задачи с вероятностью единица при уменьшении параметра регуляризации пропорционально $(\ln m)/\sqrt[4]{m}$.

СПИСОК ЛИТЕРАТУРЫ

1. Vapnik V.N. Statistical learning theory. — New York: Wiley, 1998. — 736 p.
2. Schoelkopf B., Smola A.J. Learning with kernels. Support vector machines, regularization, optimization, and beyond. — Cambridge, MA: MIT Press, 2002. — 626 p.
3. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 384 с.
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. Cristianini N., Shawe-Taylor J. An introduction to support vector machines. — Cambridge, UK: Cambridge Univ. Press, 2000. — 189 p.
7. Evgeniou T., Pontil M., Poggio T. Regularization networks and support vector machines // Adv. Comput. Math. — 2000. — **13**, N 1. — P. 1–50.
8. Cucker F., Smale S. On the mathematial foundations of learning // Bull. Amer. Math. Soc. (N.S.) — 2002. — **39**, N 1. — P. 1–49.
9. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data mining, inference, and prediction. — New York; Berlin; Heidelberg: Springer, 2001. — 533 p.
10. Poggio T., Smale S. The mathematics of learning: Dealing with data // Notices Amer. Math. Soc. — 2003. — **50,** N 5. — P. 537–544.
11. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наук. думка, 2004. — 536 с.
12. Draper N.R., Smith H. Applied regression analysis. Third edition. — Chichester: Wiley, 1998. — 736 p.
13. Györfi L., Kohler M., Krzyżak A., Walk H. A distribution free theory of nonparametric regression. — New York; Berlin; Heidelberg: Springer, 2002. — 647 p.
14. Bousquet O., Elisseeff A. Stability and generalization // J. Mach. Learn. Res. — 2002. — **2.** — P. 499–526.
15. Takeuchi I., Le Q.V., Sears T., Smola A.J. Nonparametric quantile estimation // Ibid. — 2006. — **7.** — P. 1231–1264.
16. 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.
17. Cucker F., Smale S. Best choices for regularization parameters in learning theory: on the bias-variance problem // Ibid. — 2002. — **2,** N 4. — P. 413–428.
18. Smale S., Zhou D.X. Shanon sampling and function reconstruction from point values // Bull. Amer. Math. Soc. (N.S.) — 2004. — **41,** N 3. — P. 279–305.
19. Smale S., Zhou D.X. Shannon sampling II: Connections to learning theory // Appl. Comput. Harmon. Anal. — 2005. — **19,** N 3. — P. 285–302.
20. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры классификации объектов // Кибернетика и системный анализ. — 1995. — № 4. — С. 76–89.
21. Сергиенко И.В., Гупал А.М. Оптимальные процедуры распознавания и их применение // Там же. — 2007. — № 6. — С. 41–54.
22. Koenker R., Bassett G.W. Regression quantiles // Econometrica. — 1978. — **46.** — P. 33–50.
23. Koenker R. Quantile regression. — Cambridge; New York: Cambridge Univ. Press, 2005. — 366 p.
24. McDiarmid C. On the method of bounded differences // Survey of Combinatorics. — Cambridge: Cambridge Univ. Press, 1989. — P. 148–188.
25. Smale S., Yao Y. Online learning algorithms // Found. Comput. Math. — 2006. — **6,** N 2. — P. 145–170.
26. Knopov P.S., Kasitskaya E.J. Empirical estimates in stochastic optimization and identification. — Dordrecht; Boston; London: Kluwer Acad. Publ., 2002. — 250 p.
27. Ермольев Ю.М., Кнопов П.С. Метод эмпирических средних в задачах стохастического программирования // Кибернетика и системный анализ. — 2006. — № 6. — С. 3–18.
28. Keyzer M.A. Rule-based and support vector (SV-) regression/classification algorithms for joint processing of census, map, survey and district data // Working Paper WP-05-01. — Amsterdam: Centre for World Food Studies, 2005. — 88 p. (http://www.sow.vu.nl/pdf/wp05.01.pdf)

29. 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.
30. Ароншайн Н. Теория воспроизводящих ядер // Математика (Период. сб. перев. иностр. статей). — М.: Изд-во иностр. лит., 1963. — **7,** № 2. — С. 67–130.
31. Гельфанд И.М., Виленкин Н.Я. Обобщенные функции. Вып. IV. Некоторые применения гармонического анализа. Оснащенные гильбертовы пространства. — М.: Физматгиз, 1961. — 472 с.
32. Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. Изд. пятое. — М.: Наука, 1981. — 544 с.
33. Wahba G. Spline models for observational data // CBMS-NSF Regional Conference Series in Applied Mathematics. — Philadelphia, PA: SIAM, 1990. — 59 p.
34. Надарая Э.А. Непараметрическое оценивание плотности и кривой регрессии. — Тбилиси: Изд-во Тбилис. гос. ун-та, 1983. — 194 с.
35. Huber P.J. Robust statistics. — New York: Wiley, 1981. — 308 p.
36. Ермольев Ю.М., Ястремский А.И. Стохастические модели и методы в экономическом планировании. — М.: Наука, 1979. — 254 с.
37. Ermoliev Y.M., Leonardi G. Some proposals for stochastic facility location models // Math. Modelling. — 1982. — **3.** — P. 407–420.
38. Ruszczynski A., Shapiro A. (Eds.) Stochastic programming // Handbooks in OR & MS. — Amsterdam: Elsevier, 2003. — **10.** — 682 p.
39. Devroye L., Györfi L., Lugosi G. A probabilistic theory of pattern recognition. — New York: Springer, 1996. — 634 p.
40. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. Изд. 3-е, испр. — М.: Наука, 1986. — 288 с.
41. Васильев Ф.П. Методы решения экстремальных задач. Задачи минимизации в функциональных пространствах, регуляризация, аппроксимация. — М.: Наука, 1981. — 400 с.
42. Экланд И., Темам Р. Выпуклый анализ и вариационные проблемы. — М.: Мир, 1979. — 397 с.
43. Kaplan A., Tichatschke R. Stable methods for ill-posed variational problems: Prox-regularization of elliptic variational inequalities and semi-infinite problems. — Berlin: Akad. Verlag, 1999. — 437 p.
44. Федотов А.М. Линейные некорректные задачи со случайными ошибками в данных / Отв. ред. акад. М.М. Лаврентьев. — Новосибирск: Наука. Сиб. отд-ние, 1982. — 190 с.
45. Rockafellar R.T., Wets R.J.-B. Variational analysis. — Berlin: Springer, 1998. — 733 p.
46. Norkin V., Keyzer M. On stochastic optimization and statistical learning in reproducing kernel Hilbert spaces by Support Vector Machines (SVM) // Informatika (Vilnius). — 2009. — **20,** N 2. — P. 1–19.

Поступила 16.09.2008