Исследование эффективности некоторых линейных методов классификации на модельных распределениях*

В. М. Неделько
nedelko@math.nsc.ru
Институт математики им. С. Л. Соболева СО РАН
Россия, г. Новосибирск, пр. акад. Коптюга, д. 4

Рассматривается проблема построения вероятностных моделей, позволяющих выявлять свойства методов построения решающих функций и проводить исследование этих методов. В частности, ставилась задача построения моделей, на которых заданный метод наиболее эффективен среди сравниваемых методов. Для метода логистической регрессии были построены модели, на которых этот метод эквивалентен методу максимального правдоподобия (ММП). Для метода SVM (support vector machine) построена модель, на которой этот метод приближенно эквивалентен ММП. Для дискриминанта Фишера подобной модели построить не удалось. Проведенное исследование демонстрирует перспективность подхода, основанного на построении набора «эталонных» вероятностных моделей, для исследования и сравнения методов построения решающих функций. Под эталонной моделью понимается вероятностная модель, на которой наиболее выраженно проявляется некоторое свойство исследуемого метода, например модель, на которой метод демонстрирует наибольшее превосходство, или модель, на которой проявляется некоторый недостаток метода (например, неустойчивость к «выбросам»). Также выявлены некоторые неочевидные свойства метода SVM и особенности его поведения, учет которых позволяет более эффективно применять данный метод.

Ключевые слова: *машинное обучение; логистическая регрессия; дискриминант Фишера; метод максимального правдоподобия*

DOI: 10.21469/22233792.2.3.04

1 Введение

Полезность методов построения решающих функций определяется, главным образом, точностью, которой они достигают при решении практических задач анализа данных. При этом известно, что на разных задачах лучшие результаты могут давать разные методы.

Общепринятым способом сравнения эффективности методов анализа данных является их исследование на задачах репозитория UCI. Однако такой подход обладает рядом недостатков. Первый недостаток состоит в том, что задачи попадают в репозиторий «случайно» в том смысле, что включение задачи в репозиторий зависит не от ее свойств, а от стечения обстоятельств.

Вместе с тем, предпринимались попытки создания подобного репозитория путем целенаправленного подбора задач, в частности была реализована идея включения в репозиторий для каждого из наиболее известных методов хотя бы по одной задаче, на которой этот метод был бы наиболее эффективен [1].

Другим недостатком репозиториев реальных задач является то, что каждая задача в них представлена единственной выборкой, зачастую небольшого объема, в силу чего получаемые выводы не являются строго достоверными, а носят вероятностный характер.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 14-01-00590 и № 14-07-00249.

Возникает естественная идея составить репозиторий, в котором в качестве задач были бы распределения. При этом очевидно, что придумывать распределение «наугад» неконструктивно.

В данной работе будут исследоваться возможности целенаправленного конструирования таких распределений, с тем чтобы на полученном наборе задач можно было наиболее полно выявить особенности различных методов построения решающих функций.

В качестве исследуемых методов классификации выбраны линейные методы.

В настоящее время существует несколько методов классификации, использующих линейные решающие функции [2]. Наиболее известные из них: дискриминант Фишера, машина опорных векторов (SVM) и логистическая регрессия.

Несмотря на широкое использование перечисленных методов, остается открытым вопрос, какой из методов эффективнее для тех или иных задач [3].

То, что эти методы существенно различны, дает основания предполагать, что классы задач, на которых они эффективны, также существенно различаются.

Заметим, что дискриминант Фишера и SVM в определенном смысле ближе друг к другу, чем к логистической регрессии, поскольку оба являются непараметрическими и не предполагают не только вид распределений, но и вероятностную природу данных. В отличие от этого, логистическая регрессия основана на задании определенного вида функции условной вероятности. Исходя из этого можно ожидать, что дискриминант Фишера и SVM будут эффективны на более широком классе задач [4]. Вместе с тем следует заметить, что дискриминант Фишера по факту оказывается практически идентичным решению, получаемому в предположении нормальности распределений и равенства ковариационных матриц для классов. Тот факт, что метод, полученный из сильных вероятностных предположений, указывается практически совпадающим с методом, полученным из чисто геометрических эвристик, свидетельствует о том, что класс задач, на которых метод является эффективным [5], может быть радикально шире класса задач, на которые он изначально ориентирован.

В данной работе сделана попытка подобрать семейства вероятностных моделей, которые бы позволили проявить различия в эффективности методов и проанализировать причины [6] различия эффективности.

Под вероятностной моделью в данной работе будем понимать распределение (или семейство распределений) в пространстве переменных (включая целевую), сконструированное (выбранное) в соответствии с целями проводимого исследования.

2 Основные понятия

Для введения основных понятий рассмотрим сначала общую постановку задачи построения решающих функций [7].

2.1 Постановка задачи классификации

Пусть X — пространство значений переменных, используемых для прогноза, а Y — пространство значений прогнозируемых переменных, и пусть C — множество всех вероятностных мер на мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство $\langle D, \mathfrak{B}, \mathsf{P}_c \rangle$, где $\mathfrak{B}$ — σ-алгебра, P_c — вероятностная мера. Параметр c будем называть стратегией природы.

В качестве значений целевой переменной возьмем множество $Y = \{-1, 1\}$. Как уже говорилось, мы рассматриваем случай двух классов. Для обозначения классов можно выбрать любые числовые (и нечисловые) значения. Значения -1 и 1 выбраны из соображе-

ний удобства использования в дальнейших формулах и являются одним из общепринятых вариантов.

Решающей функцией называется соответствие $\lambda\colon X \to Y$. Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$. Под риском [8] будем понимать средние потери:

$$R(c, \lambda) = \mathsf{E}_c \mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x))\ \mathsf{P}_c(dx, dy), \quad x \in X,\ y \in Y.$$

В данной работе будем рассматривать самую простую и наиболее распространенную функцию потерь $\mathcal{L}(y, y') = I(y \neq y')$. В этом случае риск есть вероятность ошибочной классификации. Здесь $I(\cdot)$ — индикаторная функция (равна 1, когда условие истинно, и 0, когда ложно).

Пусть $V_N = \{(x^i, y^i) \in D \mid i = 1, \ldots, N\}$, $V_N \in D^N$ — случайная независимая выборка из распределения P_c. В дальнейшем объем выборки N будет, как правило, фиксированным, поэтому этот параметр в обозначении выборки обычно будем опускать.

Метод (алгоритм) построения решающих функций есть отображение $Q\colon \mathcal{V} \to \Lambda$, где Λ — заданный класс решающих функций, $\mathcal{V} = \bigcup\limits_{N=1}^{\infty} D^N$ — множество всевозможных выборок, а $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q. Задача распознавания образов в стандартной постановке заключается в выборе и обосновании [9] подходящего алгоритма [10] построения решающих функций.

Критерием качества метода классификации может служить средний риск

$$\mathcal{F}_N(c, Q) = \mathsf{E}_V R(c, \lambda_{Q,V}).$$

Усреднение производится по выборкам объема N.

В случае индикаторной функции потерь качество метода характеризуется математическим ожиданием вероятности ошибочной классификации.

2.2 Оценивание функции условной вероятности

Задача распознавания образов в стандартной постановке заключается в построении и обосновании [11] подходящего алгоритма построения решающих функций. Будем также рассматривать более общую постановку задачи, когда под решающей функцией понимается оценка $\tilde{g}(x)$ функции условной вероятности

$$g(x) = \mathsf{P}_c(y = 1 \mid x) = \frac{\mathsf{P}_c(dx, y = 1)}{\mathsf{P}_c(dx)}.$$

Если в качестве значений целевой переменной выбрать 0 и 1, то функция $g(x)$ была бы функцией регрессии [12], т. е. условным математическим ожиданием (что объясняет название метода логистической регрессии). При нашем выборе значений Y функция $g(x)$ формально регрессией не является, но мы, тем не менее, следуя традиции, будем говорить о восстановлении регрессии, поскольку содержание задачи при изменении обозначений классов не меняется.

Качество решения $\tilde{g}(x)$ можно определять как вероятность ошибочной классификации для некоторого порогового классификатора, построенного по $\tilde{g}(x)$. Однако если нас интересует не просто классификация, а точность оценивания условной вероятности, то меру

качества решения $\tilde{g}(x)$ следует определять как ее точность в качестве оценки $g(x)$. Наиболее естественным представляется определение качества через использование следующей логарифмической функции потерь:

$$\mathcal{L}_g(y, \tilde{g}(x)) = -I(y = 1) \ln \tilde{g}(x) - I(y = -1) \ln(1 - \tilde{g}(x)),$$

где $I(\cdot)$ — индикаторная функция (принимает значение 1, если условие истинно, и 0 — если ложно). Выборочное среднее данной функции потерь есть взятая со знаком минус функция правдоподобия выборки по отношению к оценке условной вероятности. Математическое ожидание $R(c, \tilde{g}) = \mathsf{E}_c \mathcal{L}_g(y, \tilde{g}(x))$ этой функции потерь характеризует степень отличия $\tilde{g}(x)$ от $g(x)$, т. е. содержательно может интерпретироваться как погрешность оценки $\tilde{g}(x)$.

2.3 Эмпирические функционалы качества

Заметим, что большинство методов классификации сводятся к минимизации некоторого эмпирического критерия в заданном классе решающих функций, а многие эмпирические критерии могут быть представлены как выборочное среднее эмпирической функции потерь. Такое представление полезно для сравнения и систематизации методов построения решающих функций. При этом обоснование выбора эмпирического функционала является нетривиальной задачей [13, 14].

Простейший пример эмпирического функционала — эмпирический риск, который определяется как средние потери на выборке

$$\widetilde{R}(V, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda(x^i)).$$

Усреднение логарифмической функции потерь по выборке дает критерий правдоподобия:

$$\widetilde{R}_{MMP}(V, \tilde{g}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_g(y^i, \tilde{g}(x^i)).$$

Последнее выражение есть общепринятая функция правдоподобия, взятая с противоположным знаком. Такой знак выбран для единообразия, чтобы критерий был на минимум.

Заметим, что функция потерь $\mathcal{L}(\cdot)$ является частью постановки задачи классификации, т. е. задана, в то время как эмпирическая функция потерь, которую будем обозначать $\widetilde{\mathcal{L}}(\cdot)$, вовсе не обязана с ней совпадать и может выбираться произвольной. Поэтому эмпирический критерий качества может выглядеть как

$$\widetilde{R}(V, \tilde{g}) = \frac{1}{N} \sum_{i=1}^{N} \widetilde{\mathcal{L}}(y^i, \tilde{g}(x^i)).$$

Заметим, что в таком виде эмпирический критерий представляется не всегда (контрпримером является дискриминант Фишера). Кроме того, критерий может содержать регуляризатор.

Функцию $\widetilde{\mathcal{L}}(\cdot)$ часто называют функцией потерь. Однако мы функцией потерь называем величину $\mathcal{L}(y, \lambda(x))$ — это функция, которая является неотъемлемой частью постановки задачи классификации и отражает объективные потери от неверного решения. Функция потерь задается «заказчиком», т. е. является внешними требованиями.

В отличие от нее функция $\widetilde{\mathcal{L}}(\cdot)$ является эвристикой и частью метода решения задачи. Эта функция может выбираться произвольно, на основе интуиции исследователя. Будем называть эту функцию эмпирической функцией потерь.

Связь между этими функциями заключается в том, что разработчик метода ожидает, что минимизация эмпирической функции потерь на выборке в определенной степени соответствует минимизации функции потерь на (неизвестном) распределении (контрольных выборках). При этом ожидание, как правило, основывается на интуиции либо на полуформальной аргументации.

3 Линейные методы классификации

Пусть X представлено декартовым произведением количественных переменных $X = \prod_{j=1}^{n} X_j$, тогда $x \in X$ представляет собой n-мерный вектор.

3.1 Логистическая регрессия

Метод логистической регрессии [15] подразумевает параметрическое оценивание (построение оценки $\tilde{g}(x)$) функции условной вероятности $g(x)$.

Для вывода модели логистической регрессии рассмотрим случай нормальных распределений классов с равными ковариационными матрицами S, т. е. условные меры $\mathsf{P}_c(dx \mid y)$ задаются плотностями:

$$\varphi_y(x) = \frac{1}{(2\pi)^{n/2}|S|^{1/2}} e^{-0{,}5(x-\mu_y)^T S^{-1}(x-\mu_y)}.$$

Имеем

$$g(x) = \mathsf{P}_c(y = 1 \mid x) = \frac{p\varphi_1(x)}{p\varphi_1(x) + (1-p)\varphi_{-1}(x)},$$

где $p = \mathsf{P}_c(y = 1)$ — безусловная вероятность первого класса.

Подставив нормальную плотность, после элементарных преобразований получаем:

$$g(x) = \frac{1}{1 + e^{-(w'x + w'_0)}} = \sigma(w'x + w'_0),$$

где $w' = 0{,}5S^{-1}(\mu_1 - \mu_{-1})$, $w'_0 = 0{,}5\mu_{-1}^T S^{-1}\mu_{-1} - 0{,}5\mu_1^T S^{-1}\mu_1 + \ln p - \ln(1-p)$.

Здесь $\sigma(z) = 1/(1 + e^{-z})$ — так называемая логистическая функция (иногда также называемая сигмоидом или логит-функцией). Логистическая функция широко используется в методах анализа данных, в частности как функция активации в нейронных сетях, а также как функция для замены переменных при необходимости перейти от ограниченного множества значений (например, от вероятностей) к неограниченному множеству значений (например, для построения линейной регрессии) и наоборот.

Метод логистической регрессии основан на оценивании функции условной вероятности моделью $\tilde{g}(x) = \sigma(wx + w_0)$, в которой w и w_0 — настраиваемые параметры. Знак транспонирования для вектора в записи скалярного произведения будем опускать.

Как следует из выкладок, модель логистической регрессии является точной для случая нормальных распределений с равными ковариационными матрицами. Очевидно, что она является точной и для гораздо более широкого класса распределений, поскольку она определяет только условное распределение, но никак не зависит от безусловного распределения в X.

На практике параметры модели обычно оцениваются путем максимизации критерия правдоподобия

$$\mathfrak{R}_\sigma(V,\tilde{g}) = \frac{1}{N}\sum_{i=1}^{N} -I(y^i=1)\ln\tilde{g}(x^i) - I(y^i=-1)\ln(1-\tilde{g}(x^i)).$$

Учитывая, что $1-\sigma(z)=\sigma(-z)$, предыдущее выражение можно привести к виду:

$$\mathfrak{R}_\sigma(V,w,w_0) = \frac{1}{N}\sum_{i=1}^{N} -\ln\sigma(-y^i(wx^i+w_0)) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathcal{L}}(y^i(wx^i+w_0)).$$

Здесь $\widetilde{\mathcal{L}}(z) = -\ln\sigma(-z)$ — эмпирическая функция потерь.

3.2 Метод опорных векторов

Метод опорных векторов ранее был известен как метод обобщенного портрета.

Так же как и в дискримнанте Фишера, идея метода опорных векторов заключается в поиске такого направления в пространстве переменных, по которому классы были бы наиболее разделимы. Отличие заключается в критерии качества разделимости.

В методе опорных векторов мера разделимости основывается на понятии зазора, который понимается как ширина разделяющей полосы между классами.

Запишем линейный пороговый классификатор в виде:

$$\lambda(x) = \operatorname{sign}(wx - w_0).$$

Требуется найти вектор w и скаляр w_0, минимизирующие эмпирический риск и одновременно максимизирующие ширину разделяющей полосы.

Если классы линейно разделимы, то задача может быть записана как задача максимизации зазора при ограничениях, обеспечивающих безошибочную классификацию обучающей выборки.

Поскольку норма вектора w не влияет на направление, выберем удобную нормировку из условия

$$\min_{(x^i,y^i)\in V} y^i(x^iw - w_0) = 1.$$

Для граничных точек имеем:

$$x_+w - w_0 = 1; \quad -(x_-w - w_0) = 1.$$

Ширина разделяющей полосы:

$$(x_+ - x_-)\frac{w}{|w|} = \frac{(w_0+1)-(w_0-1)}{|w|} = \frac{2}{|w|}.$$

Получаем задачу квадратичной оптимизации:

$$\begin{cases} w^2 \to \min\limits_{w,w_0}; \\ y^i(x^iw - w_0) \geqslant 1\,, \quad i=1,\ldots,N\,. \end{cases}$$

Условие нормировки выполняется автоматически.

В случае линейно неразделимой выборки метод также сводится к задаче квадратичного программирования, которая выглядит как

$$\begin{cases} \dfrac{w^2}{2} + C\sum\limits_{i=1}^{N}\xi_i \to \min\limits_{w,w_0,\xi}; \\ y^i(x^i w - w_0) \geqslant 1 - \xi_i; \\ \xi_i \geqslant 0, \quad i = 1,\ldots,N, \end{cases}$$

где $C > 0$ — параметр.

Задача эквивалентна задаче безусловной минимизации (по w и w_0) следующего эмпирического критерия:

$$\mathfrak{R}(V, w, w_0) = \frac{w^2}{2} + C\sum_{i=1}^{N}\big(1 - y^i(x^i w - w_0)\big)_+ = \frac{w^2}{2} + C\sum_{i=1}^{N}\widetilde{\mathcal{L}}(y^i(wx^i + w_0)).$$

Здесь $(z)_+ = zI(z > 0)$ обозначает функцию, которая «зануляет» отрицательные значения аргумента.

Критерий приведен к виду с эмпирической функцией потерь $\widetilde{\mathcal{L}}(z) = (1 - z)_+$.

Слагаемое $w^2/2$ имеет смысл регуляризатора (т. е. слагаемого, вводимого для повышения устойчивости решений).

Такая форма выявляет большое сходство SVM и логистической регресии [16], которое становится особенно наглядным, если сравнить графики эмпирических функций потерь для этих методов (рис. 1).

Это сходство представляется довольно неожиданным ввиду того, что SVM — непараметрический метод, а логистическая регрессия — параметрический (с неполной вероятностной моделью).

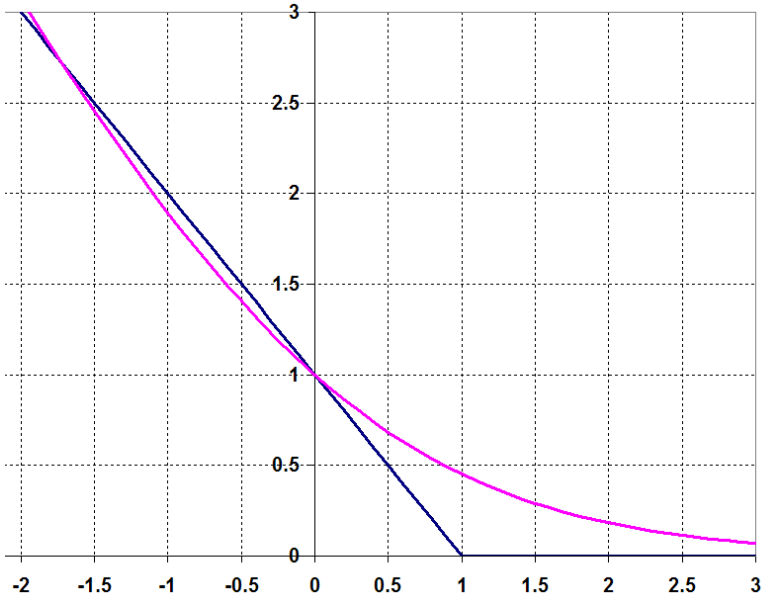


Рис. 1 Эмпирические функции потерь для SVM и логистической регрессии

Нам представляется более удобным записать критерий в несколько ином виде:

$$\mathfrak{R}(V, w, w_0) = \varkappa \frac{w^2}{2} + \frac{1}{N} \sum_{i=1}^{N} \widetilde{\mathcal{L}}(y^i(wx^i + w_0)).$$

Здесь вместо параметра C введен параметр $\varkappa = N/C$. Преимущество этой формы записи критерия состоит в том, что нулевое значение параметра C не имеет смысла, в то время как $\varkappa = 0$ – вполне допустимое и разумное значение параметра, которое соответствует отсутствию регуляризатора. При этом следует оговориться, что для случая линейно разделимой выборки $\varkappa = 0$ все же не годится, поскольку решение будет определяться неоднозначно (и не будет максимизироваться зазор), однако остаются допустимыми сколь угодно малые значени $\varkappa$.

Утверждение 1. *Оптимальное значение критерия SVM не превосходит* 1, *т. е.*

$$\min_{w, w_0, \xi} \mathfrak{R}(V, w, w_0) \leqslant 1.$$

Доказательство. Рассмотрим решение с параметрами

$$w = 0, \quad w_0 = \arg\max_{y} \sum_{i=1}^{N} I(y^i = y),$$

т. е. w_0 принимает значение того класса, представителей которого в выборке больше (при равенстве частот — любого). Эта решающая функция для всех x прогнозирует класс с наибольшей частотой в обучающей выборке.

Очевидно, что эмпирический риск для такой решающей функции не превосходит 0,5. При этом непосредственной подстановкой можно убедиться, что значение критерия SVM для такого решения есть удвоенное значение эмпирического риска. ■

Данная простая оценка объясняет, почему в случае, когда классы плохо разделимы, метод SVM предпочитает их не разделять вовсе, а относить все объекты к одному классу. Особенно часто это происходит, когда частоты классов в обучающей выборке сильно различаются.

Заметим, что для подобных решающих функций, которые все объекты относят к одному классу, ширина раздеряющей полосы (зазор) формально бесконечна. Увеличение параметра $\varkappa$ повышает вероятность того, что метод SVM построит такое решение.

3.3 Дискриминант Фишера

Дискриминант Фишера, как и SVM, использует исключительно метрические свойства конфигурации выборочных точек и не требует не только никаких предположений о распределениях, но и вообще статистической постановки задачи классификации.

Идея дискриминанта Фишера заключается в выборе такого направления в пространстве переменных, при проецировании выборки на которое образы классов оказываются в некотором смысле наиболее удаленными друг от друга. Формально это выражается в максимизации следующего критерия:

$$\Phi(w) = \frac{(\tilde{\mu}_{w,1} - \tilde{\mu}_{w,-1})^2}{\tilde{S}_w},$$

где $\tilde{\mu}_{w,y} = (1/N_y) \sum_{i=1}^{N} wx^i I(y^i = y)$ — среднее проекций точек выборки y-го класса на направление w, N_y — число объектов y-го класса в выборке, а $\tilde{S}_w = \sum_{i=1}^{N} (wx^i - \tilde{\mu}_{w,y^i})^2$ —

суммарный квадрат отклонений проекций точек выборки y-го класса на направление w от среднего этого класса,

Критерий приводится к виду:

$$\Phi(w) = \frac{(w\tilde{\mu}_1 - w\tilde{\mu}_{-1})^2}{w^T\tilde{S}w} = \frac{w^T(\tilde{\mu}_1 - \tilde{\mu}_{-1})(\tilde{\mu}_1 - \tilde{\mu}_{-1})^T w}{w^T\tilde{S}w},$$

где $\tilde{\mu}_{w,y} = (1/N_y)\sum_{i=1}^{N} x^i I(y^i = y)$ — среднее точек выборки y-го класса, $\tilde{S}_y$ — выборочная ковариационная матрица y-го класса, $\tilde{S} = N_1\tilde{S}_1 + N_{-1}\tilde{S}_{-1}$.

Последняя форма критерия имеет вид отношения Релея.

Известно, что максимум $\Phi(w)$ достигается при $w = w_\Phi = \tilde{S}^{-1}(\tilde{\mu}_1 - \tilde{\mu}_{-1})$.

Заметим, что выражение для w_Φ совпадает (с точностью до нормировки) с выражением для нормали к разделяющей гиперплоскости для случая нормальных распределений с равными матрицами ковариаций. Такое сходство приводит к тому, что в литературе эти методы иногда смешиваются, несмотря на их принципиальное различие по подходу и предположениям.

После выбора направления w_Φ задача классификации становится одномерной и может быть достаточно легко решена. Более того, в некоторых случаях полученное (проецированием) упорядочивание объектов само по себе может считаться решением.

3.4 Модификации методов

Рассмотренные методы обладают схожими чертами, в частности они допускают одинаковые усовершенствования.

Первое усовершенствование касается регуляризации.

Заметим, что SVM регуляризующий член содержит изначально — это $\varkappa w^2/2$, слагаемое, отвечающее за максимизацию отступа (зазора).

Ровно это же слагаемое можно добавить к критерию логистической регрессии. И это слагаемое также будет отвечать за максимизацию отступа. Хотя отступ для логистической регрессии имеет не такой очевидный смысл, как для SVM, его можно определить как ширину полосы между классами, в которой эмпирическая функция потерь для обоих классов не превышает 1. Это определение годится и для SVM.

Для дискриминанта Фишера аналогичный регуляризатор выглядит как добавка к матрице S в виде некоторой диагональной матрицы (аналогично гребневой регрессии).

Второе усовершенствование касается использования функции ядра (так называемый kernel trick). Заметим, что исторически kernel trick был изначально разработан для метода SVM и лишь впоследствии распространен на остальные методы. Тем не менее, этот прием [17] полностью аналогичен для всех трех методов.

Возможность перехода к ядрам обусловлена двумя факторами. Во-первых, линейностью решения, т. е. тем, что решение дается через скалярное произведение вектора, описывающего объект, с направляющим вектором. Во-вторых, критерии всех трех методов таковы, что оптимальное направление может быть представлено как линейная комбинация векторов выборки. Это справедливо не только в том очевидном случае, когда ранг системы векторов выборки равен размерности пространства, но и в общем случае, в том числе для бесконечномерных пространств.

Из этих фактов элементарно следует, что решение может быть записано с использованием только скалярных произведений между объектами, но не требует даже признаковых описаний объектов [18].

Если бы второй факт не имел места, то оптимальное направление в спрямляющем пространстве могло бы не иметь прообраза в исходном пространстве, и тогда это спрямляющее пространство пришлось бы вводить явно.

4 Решающие функции на распределениях

Естественным первым шагом в оценивании эффективности метода построения решающих функций является исследование метода на распределениях. Фактически это означает изучение асимптотических свойств [19, 20] метода (при объеме выборки, стремящемся к бесконечности).

Применение методов к распределениям позволяет, например, оценить, насколько в принципе метод способен приблизиться к Байесовскому решению. В данной работе это также позволит аналитически исследовать поведение отступа.

В отличие от работ, в которых вероятностная модель конструируется на основе выборки [21], будем конструировать модели таким образом, чтобы исследуемые особенности методов проявлялись на них наиболее выраженно.

В качестве первого примера рассмотрим модель нормальных распределений с равными ковариационными матрицами. Положим для простоты безусловные вероятности классов равными, т. е. $\mathsf{P}(y) = 0{,}5$.

Для удобства выберем начало координат посредине между центрами классов, т. е. так, чтобы для векторов математических ожиданий выполнялось $\mu_{-1} = -\mu_1$. Введем параметр μ, через который выразим вектора математических ожиданий как $\mu_y = y\mu$.

Вычислим для данной модели значение критерия SVM:

$$\mathfrak{R}(c, w, w_0) = \varkappa\frac{w^2}{2} + \int\limits_D \widetilde{\mathcal{L}}(y(wx + w_0))\ \mathsf{P}_c(dx, dy).$$

При сделанном выборе начала координат оптимальное значение параметра w_0 равно 0, поэтому w_0 можно исключить из выражения.

Рассмотрим для начала одномерный случай, когда x, w и μ — скаляры.

Для модели с одномерными нормальными распределениями можем записать:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \sum_{y\in\{-1,1\}} \frac{\mathsf{P}(y)}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-(x-\mu_y)^2/(2\sigma)}(1 - y(wx))_+\, dx.$$

Учитывая симметрию распределений классов, после преобразований получаем:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{1}{w}} e^{-(x-\mu)^2/(2\sigma)}(1 - wx)\, dx,$$

Делая стандартную замену $t = (x - \mu)/\sigma$, получаем:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{(1-\mu w)/(w\sigma)} e^{-t^2/2}(1 - w(t\sigma + \mu))\, dt.$$

После элементарных преобразований выражение приводится к виду:

$$\mathfrak{R}(c, w) = \varkappa\frac{w^2}{2} + \frac{zF_{\mathcal{N}}(z) + F'_{\mathcal{N}}(z)}{z + m},$$

где $m = \mu/\sigma$; $z = 1/(w\sigma) - m$; $F_{\mathcal{N}}(\cdot)$ — функция (интегральная) стандартного нормального распределения; $F'_{\mathcal{N}}(\cdot)$ — плотность нормального распределения.

Полученное выражение позволяет установить некоторые свойства SVM.

Введем обозначение:

$$w^* = \arg\min_w \mathfrak{R}(c, w).$$

Величина $1/w^*$ называется зазором или отступом (margin).

Введем обозначение $s = 1/(w^*\sigma)$.

Утверждение 2. *Величина зазора s для метода SVM без регуляризации (при $\varkappa = 0$) монотонно убывает с ростом m.*

Данное свойство выглядит парадоксальным: при отдалении распределений классов друг от друга величина зазора (т. е. ширина разделяющей полосы) уменьшается.

Доказательство. Пусть $\varkappa = 0$. Для нахождения w^* вычислим производную:

$$\frac{\partial \mathfrak{R}(c, w)}{\partial z} = \frac{mF_{\mathcal{N}}(z) - F'_{\mathcal{N}}(z)}{(z+m)^2}.$$

Производная равна нулю при $m = F'_{\mathcal{N}}(z^*)/F_{\mathcal{N}}(z^*)$, где $z^* = s - m$.

Требуется установить, что зависимость s от m, неявно задаваемая выражением $m = F'_{\mathcal{N}}(s-m)/F_{\mathcal{N}}(s-m)$, является монотонно убывающей функцией. Для этого вычислим производную $m' = dm/ds$. Получаем:

$$m' = \frac{F''_{\mathcal{N}}(s-m)F_{\mathcal{N}}(s-m) - (F'_{\mathcal{N}}(s-m))^2}{(F_{\mathcal{N}}(s-m))^2}(1-m') = -sm(1-m'),$$

откуда выражаем

$$m' = \frac{sm}{sm-1}.$$

Легко убедиться, что $sm < 1$. Действительно,

$$sm = (z+m)m = \left(z + \frac{F'_{\mathcal{N}}(z)}{F_{\mathcal{N}}(z)}\right)\frac{F'_{\mathcal{N}}(z)}{F_{\mathcal{N}}(z)}.$$

Остается убедиться, что

$$zF'_{\mathcal{N}}(z)F_{\mathcal{N}}(z) + (F'_{\mathcal{N}}(z))^2 - (F_{\mathcal{N}}(z))^2 < 0$$

при любых z. Это можно сделать стандартным методом анализа функций, т. е. вычислив и проанализировав несколько производных.

Таким образом, установлено, что $m' < 0$. ■

Оптимальное значение критерия SVM при $\varkappa = 0$ есть

$$\mathfrak{R}(c, w^*) = \frac{z^*F_{\mathcal{N}}(z^*) + F'_{\mathcal{N}}(z^*)}{z^*+m} = F_{\mathcal{N}}(z^*).$$

Только что было доказано, что величина s убывает с ростом m, но тогда $z^* = s - m$ — тем более монотонно убывающая функция m. Учитывая монотонность $F_{\mathcal{N}}(\cdot)$, заключаем, что $\mathfrak{R}(c, w^*)$ монотонно убывает с ростом m.

Данный факт вполне естественен и очевиден. Его можно доказать намного проще, но в приведенных выкладках были также получены полезные выражения для отступа.

Вернемся теперь к случаю многомерного пространства переменных.

Проекция нормального распределения на произвольное направление w есть одномерное нормальное распределение, параметры которого обозначим μ_w и σ_w.

Значение критерия SVM по направлению w полностью определяется (при $\varkappa = 0$) величиной $m_w = \mu_w/\sigma_w$, причем монотонно от нее зависит.

Заметим, что критерий дискриминанта Фишера для рассматриваемой модели нормальных распределений есть просто $4m_w^2$.

Из сказанного следует, что оптимальные направления для разделяющих функций SVM и дискриминанта Фишера совпадают.

Утверждение 3. *При $\varkappa = 0$ на модели нормальных распределений с равными матрицами ковариаций решения методами SVM, логистической регрессии и дискриминанта Фишера совпадают с Байесовским решением.*

Данный факт является известным.

Применительно к дискриминанту Фишера это классический результат. Для логистической регрессии утверждение доказывается элементарно.

Что касается SVM, то автору не удалось найти работы, где этот метод применялся бы не к выборке, а к распределениям. Однако известны результаты о состоятельности метода SVM [22], из которых следуют схожие выводы.

Справедливость утверждения для метода SVM следует из установленного выше факта совпадения получаемого решения с решением дискриминанта Фишера.

Заметим, что исследуемые методы дают оптимальные решения далеко не на всех вероятностных моделях.

Утверждение 4. *Существуют вероятностные модели, для которых Байесовская разделяющая функция линейна, но решения, полученные методами SVM, логистической регрессии и дискриминанта Фишера, не являются оптимальными.*

Доказательство. Для доказательства явно построим модель, обладающую требуемым свойством.

Эта модель является смесью исходной модели нормальных распределений с равными матрицами ковариаций и модели с распределениями, сосредоточенными в некоторой точке $\check{x}$, причем для этой второй компоненты вероятности классов в точке $\check{x}$ одинаковы.

Очевидно, что добавление описанной компоненты к исходной модели не меняет Байесовского решения, поскольку для этой компоненты классы неразделимы (любое решение дает одинаковую вероятность ошибочной классификации).

При этом вес второй компоненты и положение точки $\check{x}$ существенно влияют на решения, получаемые перечисленными методами. В частности, метод SVM при достаточно большом весе второй компоненты изменит решение так, чтобы точка $\check{x}$ попала внутрь разделяющей полосы. ■

Выясним теперь, как влияет на решение параметр регуляризации $\varkappa$ в методе SVM.

На рис. 2 приведены разделяющие функции, построенные методом SVM для модели с нормальными распределениями. Ковариационные матрицы для обоих классов одинаковы и соответствуют стандартным отклонениям, равным по главным осям 1 и 2. Математические ожидания для классов равны соответственно -1 и 1. Распределения на рисунке изображены соответственно синим и зеленым эллипсами (кривые равной плотности вероятности). Черная прямая соответствует решению, построенному методом SVM при $\varkappa = 0$,

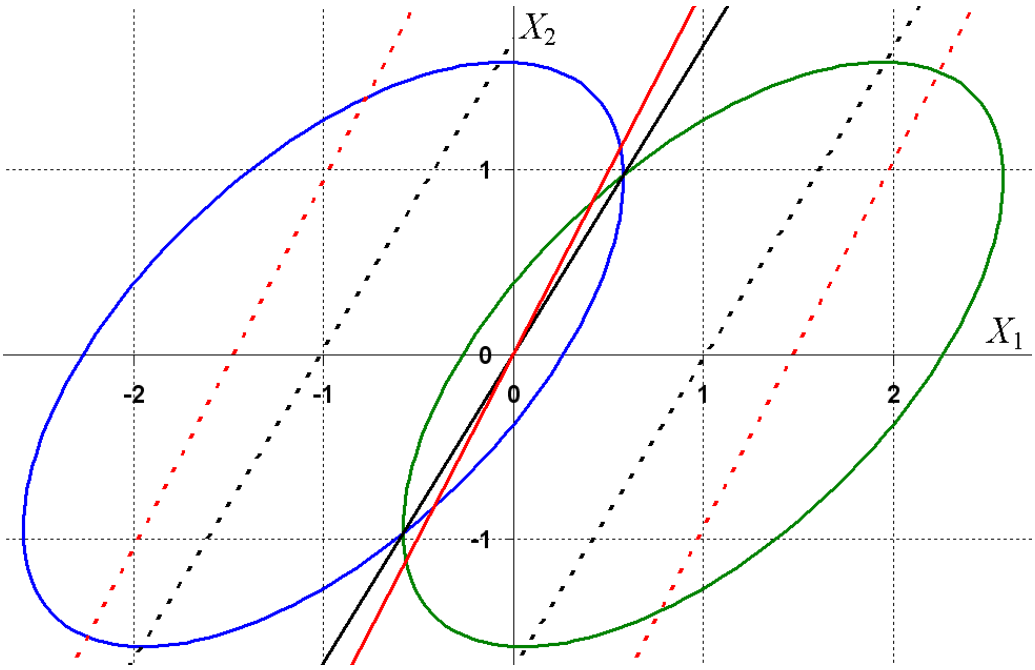


Рис. 2 Решения, построенные методом SVM для модели с нормальными распределениями при нулевом (черная линия) и ненулевом (красная линия) значениях параметра регуляризации $\varkappa$.

которое совпадает с Байесовской решающей функцией. Красная прямая соответствует $\varkappa = 0{,}2$. Пунктирные линии отмечают разделяющую полосу для решений соответствующего цвета.

Как видно из примера, увеличение $\varkappa$ приводит к повороту разделяющей прямой, как если бы эллипсы распределений стали менее вытянутыми. Такое же изменение решения можно получить и на основе дискриминанта Фишера, добавляя к оценке ковариационной матрицы единичную матрицу с некоторым коэффициентом.

5 Модели максимального правдоподобия

Для исследования эффективности методов выберем подходящие задачи, т. е. вероятностные модели, в соответствии с которыми будут генерироваться выборки для тестирования методов.

Вообще говоря, модель для тестирования можно придумать совершенно произвольную. Выберем для каждого метода модель, которая ему в некотором смысле наиболее подходит, а именно: модель, при которой метод классификации соответствует ММП (или близок к нему).

Следует уточнить, в каком смысле мы говорим о соответствии.

5.1 Метод максимального правдоподобия

В классическом виде ММП каждой выборке сопоставляет распределение (из заданного параметрического семейства), для которого функция правдоподобия данной выборки максимальна.

В более универсальном виде ММП можно сформулировать как метод, сопоставляющий выборке распределение, которое критерием отношения правдоподобия будет выбрано против любой альтернативы.

Пусть Θ — заданное множество вероятностных мер (в параметрическом случае будем отождествлять Θ со множеством значений параметров).

Будем говорить, что вероятностная мера θ_1 не менее правдоподобна, чем мера θ_2, по отношению к выборке V, если во всех точках выборки существует производная Радона–Никодима меры θ_2 по мере θ_1 и произведение этих производных по всем точкам выборки не превосходит 1.

Метод максимального правдоподобия сопоставляет выборке меру θ^*, которая не менее правдоподобна, чем любая мера из Θ.

Заметим, что ММП определен не для любого семейства распределений Θ, поскольку не в любом Θ найдётся θ^* с требуемыми свойствами. Однако классический вариант ММП, основанный на функции правдоподобия, является частным случаем приведенного.

Такое обобщение нужно, чтобы иметь возможность определить ММП для случаев, когда семейство распределений очень широкое. В частности, так можно определить ММП даже для случая всех вероятностных мер (заданных на одной и той же σ-алгебре). При этом наиболее правдоподобная мера будет эмпирическим распределением для выборки (когда в каждой точке выборки сосредоточена вероятность $1/N$).

Будем считать, что ММП задает отображение множества выборок в некоторое множество Θ, элементами которого являются в зависимости от постановки задачи либо вероятностные меры, либо значения параметров распределений. Будем обозначать это отображение как $\theta^*(V)$, где $V \in D^N$.

Определение 1. *Будем говорить, что метод классификации Q, отображающий множество выборок D^N во множество решающих функций Λ, соответствует ММП для семейства распределений Θ, если существует отображение $\zeta : \Theta \to \Lambda$, такое что $\lambda_{Q,V} = \zeta(\theta^*(V))$ для всех $V \in D^N$.*

Такое отображение, очевидно, существует тогда и только тогда, когда не существует двух выборок, образы которых для ММП совпадают, а для метода Q различаются.

Заметим, что если в качестве Θ взять множество всех вероятностных мер, то любой метод классификации будет соответствовать ММП. Поэтому факт такого соответствия практически значим только в случае, если класс Θ достаточно узок.

Определение 2. *Будем говорить, что метод классификации Q эквивалентен ММП для семейства распределений Θ, если существует взаимно однозначное отображение $\zeta : \Theta \to \Lambda$, такое что $\lambda_{Q,V} = \zeta(\theta^*(V))$ для всех $V \in D^N$.*

Попытаемся построить для рассматриваемых в работе линейных методов классификации модели, для которых эти методы будут эквивалентны методам максимального правдоподобия.

5.2 Общий вид модели

Логистическая регрессия была построена как модель максимального правдоподобия. Возникает вопрос, можно ли остальные методы интерпретировать как методы максимального правдоподобия для некоторой вероятностной модели.

Если критерий выражается через эмпирическую функцию потерь, то соответствующей вероятностной моделью может быть параметрическое семейство распределений с плотностью вида

$$\varphi(x, y) = A(w, w_0)e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}\varphi_0(x). \tag{1}$$

Идея построения этой модели очевидна. Если мы хотим, чтобы функция потерь совпадала с функцией правдоподобия (с обратным знаком), то напрашивается задать плотность вероятности как экспоненту от функции потерь. Однако сама по себе функция $e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}$

не может быть плотностью, поскольку интеграл от нее бесконечен. По этой причине приходится включить множитель $\varphi_0(x)$ — это некоторая функция, не зависящая от параметров w и w_0, которая обеспечивает конечность интеграла. Кроме того, требуется добавить нормировочный множитель $A(w, w_0)$.

В общем случае нормировочный множитель зависит от параметров w и w_0, поэтому добиться точного совпадения критерия на основе эмпирической функции потерь и критерия максимального правдоподобия не всегда удается. Однако отличие (связанное с наличием этого нормировочного множителя) на практике несущественно.

Утверждение 5. *Для того чтобы эмпирическая функция потерь была функцией правдоподобия для условной вероятности, т. е. чтобы было возможно представление* $\mathsf{P}(y \mid x) = e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}$, *необходимо и достаточно, чтобы выполнялось соотношение*

$$\widetilde{\mathcal{L}}(z) = -\ln\left(1 - e^{-\widetilde{\mathcal{L}}(-z)}\right).$$

Доказательство. Для условной вероятности должно выполняться соотношение

$$\mathsf{P}(y = 1 \mid x) = 1 - \mathsf{P}(y = -1 \mid x).$$

Подставив требуемое представление, имеем

$$e^{-\widetilde{\mathcal{L}}(+1(wx+w_0))} = 1 - e^{-\widetilde{\mathcal{L}}(-1(wx+w_0))}.$$

Элементарными преобразованиями получаем искомое. ■

Следствие 1. *При выполнении условия из утверждения* 1 *модель* (1) *может быть представлена в виде:*

$$\varphi(x, y) = e^{-\widetilde{\mathcal{L}}(y(wx+w_0))}\varphi_0(x),$$

где $\varphi_0(x)$ *— безусловная плотность для* x.

Нормировочный множитель $A(w, w_0)$ в этом случае не требуется, поскольку, умножив плотность на условную вероятность, получим автоматически нормированную совместную плотность.

5.3 Логистическая регрессия

Логистическая регрессия изначально построена как модель максимального правдоподобия для условной вероятности. Это означает, что условие утверждения 1 должно выполняться (в чем легко убедиться непосредственно).

В качестве модели, на которой логистическая регрессия будет построена ММП, возьмем распределение

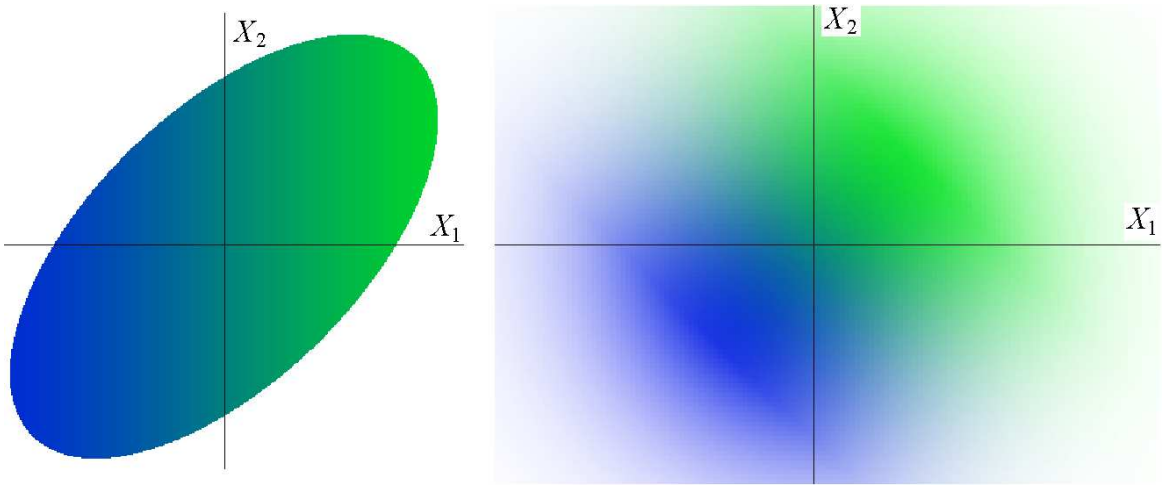
$$\varphi(x, y) = \sigma(y(wx))\varphi_{\mathcal{E}}(x),$$

где $\varphi_{\mathcal{E}}(x)$ — равномерное распределение внутри эллипса с осями 2 и 1, главная ось повернута на угол $\pi/4$.

Пример распределения из этого семейства приведен на рис. 3, *а*. Интенсивность синего и зеленого цветов отражает плотности вероятности классов 1 и -1 при $w = (1, 0)$.

5.4 Метод опорных векторов

Для метода SVM невозможно подобрать модель максимального правдоподобия [23], где бы от параметров зависела только условная вероятность.



$(а)$ $(б)$

Рис. 3 Плотности вероятности двух классов для моделей логистической регрессии $(а)$ и SVM $(б)$

Выберем следующее семейство распределений (является модификацией модели из [24]):

$$\varphi(x,y) = A\left(|w|\right) e^{-(1-y(wx))_+} \varphi_{\mathcal{N}}(x)\,, \tag{2}$$

где $\varphi_{\mathcal{N}}(x)$ — нормальное распределение с нулевым средним и единичной ковариационной матрицей.

Пример распределения из этого семейства приведен на рис. 3, *б*. Интенсивность синего и зеленого цветов отражает плотности вероятностей классов 1 и -1 при $w\sqrt{2} = (1,1)$.

Данную плотность можно представить как произведение условной вероятности

$$\mathsf{P}(y \mid x) = \sigma\left(y\left((1+wx)_+ - (1-wx)_+\right)\right)$$

и безусловной плотности

$$\varphi(x) = A\left(|w|\right)\left(e^{-(1+wx)_+} + e^{-(1-wx)_+}\right)\varphi_{\mathcal{N}}(x).$$

Для наглядности преобразуем функцию условной вероятности:

$$g(x) = \sigma\left((1+wx)_+ - (1-wx)_+\right) = \begin{cases} \sigma(wx-1), & wx < -1\,; \\ \sigma(2wx), & -1 \leqslant wx \leqslant 1\,; \\ \sigma(wx+1), & wx > 1. \end{cases}$$

Данная функция изображена синей кривой на рис. 4. Красная кривая изображает функцию $e^{-(1+wx)_+} + e^{-(1-wx)_+}$.

Вычислим $A(|w|)$. Из условия нормировки имеем:

$$\frac{1}{A\left(|w|\right)} = \sum_{y\in\{-1,1\}} \int\limits_X \varphi(x,y)\,dx = \sum_{y\in\{-1,1\}} (2\pi)^{-n/2} \int\limits_X e^{-x^2/2} e^{-(1-y(wx))_+}\,dx\,.$$

Повернем систему координат так, чтобы переменная X_1 была в направлении вектора w. Учитывая симметричность распределений при разных y, имеем:

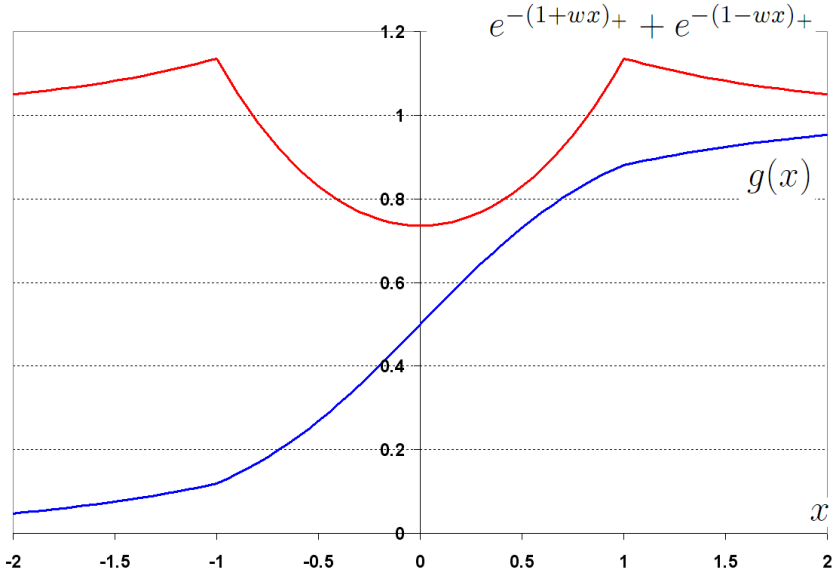


Рис. 4 Функция условной вероятности $g(x) = \sigma\left((1+wx)_+ - (1-wx)_+\right)$ и параметрическая компонента распределения $e^{-(1+wx)_+} + e^{-(1-wx)_+}$ для модели SVM

$$\frac{1}{2A(|w|)} = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} e^{-x^2/2} e^{-(1-|w|x_1))_+}\, dx_1 = e^{w^2/2-1} F_{\mathcal{N}}\left(\frac{1}{|w|} - |w|\right),$$

где $F_{\mathcal{N}}(\cdot)$ — функция (интегральная) нормального распределения.

Заметим, что при малых $|w|$ имеет место $-\ln A(|w|) \approx w^2/2 - 1 + \ln 2$. В этом случае функция правдоподобия (с обратным знаком) с точностью до аддитивной константы может быть представлена как

$$-\ln \varphi(x, y) \approx \frac{w^2}{2} + (1 - y(wx))_+.$$

Получаем, что при малых $|w|$ для рассмотренной модели критерий максимального правдоподобия совпадает с критерием SVM при $\varkappa = N$.

Заметим, что во многих источниках регуляризующее слагаемое $\varkappa w^2/2$ интерпретируется в рамках Байесовского подхода [25], когда параметр w полагается случайным [26,27]. Здесь, однако, получена модель максимального правдоподобия для критерия SVM без вероятностной интерпретации параметра w.

5.5 Дискриминант Фишера

Труднее всего подобрать модель максимального правдоподобия для дискриминанта Фишера.

Это утверждение может показаться неожиданным, поскольку дискриминант Фишера принято связывать с вероятностной моделью нормальных распределений с равными ковариационными матрицами.

Следует, однако, уточнить, в чем заключается эта связь.

Известно, что дискриминант Фишера дает оптимальное (Байесовское) решение, будучи примененным непосредственно к нормальным распределениям с равными ковариационными матрицами. Однако, как мы выяснили в разд. 1, ровно этим же свойством обладают и логистическая регрессия, и SVM при $\varkappa = 0$.

Также связь заключается в том, что дискриминант Фишера выражается через оценки максимального правдоподобия нормальных распределений. Это означает, что дискриминант Фишера соответствует ММП на классе нормальных распределений (с равными ковариационными матрицами), но не эквивалентен ему, поскольку по решающей функции параметры вероятностной модели не восстанавливаются.

При этом критерий дискриминанта Фишера не может являться функцией правдоподобия, хотя бы потому, что он не аддитивен по объектам выборки.

Гипотеза 1. Не существует невырожденной вероятностной модели, на которой дискриминант Фишера был бы эквивалентен ММП.

Под невырожденностью здесь понимается, что размерность пространства переменных больше 1 и что вероятности не сосредоточены на многообразиях нулевой меры Лебега.

6 Численный эксперимент

Сравнение методов проводилось неоднократно (см., например, [28]), однако доступные в литературе выводы об области применимости каждого метода носят частный характер.

В ближайшее время вряд ли следует ожидать появления исчерпывающего описания семейств вероятностных моделей, для которых был бы наиболее предпочтителен заданный метод, например SVM. Задача построения такого описания представляется чрезвычайно сложной.

В данной работе численный эксперимент проводится также на конкретных частных примерах, однако вероятностные модели подбираются как характерные представители определенных классов моделей с заданными свойствами (например, модели с редкими большими уклонениями). Это позволяет надеяться, что обнаруженные закономерности в поведении исследуемых методов будут иметь более общий характер.

В качестве задач, на которых будут тестироваться методы, выбраны следующие модели.

В первую очередь, для каждого метода была сконструирована вероятностная модель, на которой этот метод предположительно должен давать наилучший результат. Для дискриминанта Фишера это модель с нормальными распределениями, для SVM и логистической регрессии это модели максимального правдоподобия.

Предварительный численный эксперимент, однако, показал, что на всех трех моделях лучшим оказывается дискриминант Фишера. В связи с этим были целенаправлено подобраны модели, позволяющие каждому методу продемонстрировать преимущество.

Список моделей:

1) нормальные распределения с равными матрицами ковариаций. Модель описана в разд. 1, параметры те же: стандартные отклонения по главным осям 1 и 2, математические ожидания для классов -1 и 1, главная ось повернута на угол $\pi/4$ (см. рис. 2);
2) нормальные распределения с «шумом». К предыдущей модели добавлена «шумовая» компонента, для которой классы имеют одинаковые нормальные распределения с нулевыми средними и стандартным отклонением 5 по любому направлению. Вес (вероятность) компоненты равен 0,1;
3) модель с логистической функцией условной вероятности. Безусловное распределение $\mathsf{P}(dx)$ выбрано равномерным внутри эллипса с осями 2 и 1, главная ось повернута на угол $\pi/4$. Условная вероятность имеет вид $g(x) = \sigma(x)$;
4) к предыдущей модели добавлена шумовая компонента, как в модели 2;

Усредненные вероятности ошибочной классификации на различных моделях

Вероятностная модель	Методы классификации				
	Байесовское решающее правило	ЛДФ	Логистическая регрессия	SVM, $\varkappa = 0$	SVM, $\varkappa = 0{,}2$
1. Нормальное распределение	0,216	0,235	0,237	0,241	0,259
2. Нормальное распределение, шум	0,244	0,313	0,309	0,305	0,326
3. Логистическая ($g(x) = \sigma(x)$)	0,345	0,380	0,382	0,389	0,428
4. Логистическая, шум	0,359	0,433	0,430	0,431	0,464
5. Логистическая, смесь распределений	0,320	0,365	0,352	0,364	0,414
6. Логистическая ($g(x) = \sigma(2{,}5x)$)	0,202	0,220	0,223	0,226	0,256
7. Правдоподобия для критерия SVM	0,207	0,228	0,232	0,233	0,258

5) безусловное распределение $\mathsf{P}(dx)$ есть смесь равномерного распределения внутри эллипса из модели 3 и нормального распределения со стандартным отклонением 5 (по любому направлению). Вес второй компоненты 0,1. Условная вероятность есть $g(x) = \sigma(x)$;
6) модель, как в варианте 3, за исключением того, что $g(x) = \sigma(2{,}5x)$;
7) модель задается формулой (2) при $w\sqrt{2} = (1, 1)$ и изображена на рис. 3, *б*.

В таблице приведены полученные методом статистического моделирования оценки математических ожиданий вероятностей ошибочной классификации $\mathcal{F}(c, Q)$ для следующих методов: Байесовское (оптимальное) решающее правило; линейный дискриминант Фишера (ЛДФ); логистическая регрессия; метод SVM с параметром $\varkappa = 0$; метод SVM с $\varkappa = 0{,}2$.

Заметим, что на самом деле метод SVM запускался не при нулевом $\varkappa$, а при при $\varkappa = 0{,}0001$. Однако говорить о нулевом значении параметра $\varkappa$ все же допустимо, поскольку существует предел решения при $\varkappa \to 0$, и достаточно малые значения $\varkappa$ можно практически отождествлять с нулем.

Погрешность (в смысле стандартного отклонения) приведенных в таблице значений около 0,002.

Результаты приведены для объема выборки $N = 30$. Данное значение выбрано эмпирически как объем выборки, при котором различие методов проявляется в наибольшей степени. При других значениях N результаты качественно согласуются с приведенными.

Анализируя таблицу, можно заметить, что дискриминант Фишера оказался лучшим методом на моделях 1, 3, 6 и 7. Превосходство этого метода на модели 1 соответствует общепринятым ожиданиям (хотя убедительные обоснования для таких ожиданий отсутствуют), согласно которым ЛДФ принято ассоциировать с нормальными распределениями. Однако модель 6 существенно отличается от нормальной, но ЛДФ и на ней существенно лучше остальных методов.

Вместе с тем, на всех моделях, состоящих из смеси распределений, метод ЛДФ существенно проигрывает. Причина этого, очевидно, в том, что критерий дискриминанта Фишера содержит квадраты отклонений выборочных точек и поэтому неустойчив к редким большим отклонениям («выбросам»).

В целом, можно сделать вывод, что эффективность ЛДФ связана не с нормальностью распределений, а с наличием «выбросов». Действительно, в рамках исследования ЛДФ оказался лучшим на всех моделях без «выбросов».

Логистическая регрессия оказалась лучшей на большинстве оставшихся моделей, особенно на модели 5. Это вполне объяснимо, поскольку в этой модели функция условной вероятности имеет вид логистической кривой и при этом имеют место «выбросы».

В моделях 2 и 4 вид «шумовой» компоненты нарушает логистическую форму кривой условной вероятности, а наличие «выбросов» не позволяет получать хорошие решения методом ЛДФ. В результате, на модели 2 лучшим оказывается метод SVM.

Что касается параметра регуляризации, то во всех случаях решение при $\varkappa = 0$ оказывалось лучше, чем при $\varkappa = 0{,}2$. Вероятно, это связано как раз с тем, что ненулевые значения $\varkappa$ увеличивают вероятность того, что SVM не разобъет выборку, а отнесет ее к одному классу. Но все вероятностные модели таковы, что вероятность ошибочной классификации для такого решения равна 0,5.

Как и следовало ожидать, результаты логистической регрессии и SVM отличаются в большинстве случаев незначительно. Это объясняется тем, что методы различаются лишь эмпирическими функциями потерь, которые при этом достаточно близки.

7 Заключение

В работе поднята проблема построения вероятностных моделей, позволяющих выявлять свойства методов построения решающих функций и проводить исследование этих методов. В частности, ставилась задача построения моделей, на которых заданный метод наиболее эффективен среди сравиваемых методов.

Для метода логистической регрессии были построены модели, на которых этот метод эквивалентен ММП. Для метода SVM построена модель, на которой этот метод приближенно эквивалентен ММП. Для дискриминанта Фишера подобной модели построить не удалось.

Проблема построения набора «эталонных» вероятностных моделей для исследования и сравнения методов построения решающих функций остается практически полностью открытой. Вместе с тем, проведенное исследование демонстрирует принципиальную возможность продвижения в ее решении.

Также в работе выявлены некоторые неочевидные свойства метода SVM и особенности его поведения, учет которых позволяет более эффективно применять данный метод.

В работе были, в частности, установлены следующие любопытные факты:

- при применении метода SVM к модели нормальных распределений с равными матрицами величина зазора (ширина разделяющей полосы) уменьшается при удалении распределений друг от друга;
- существуют вероятностные модели, для которых Байесовская разделяющая функция линейна, но решения, полученные (на распределениях) методами SVM, логистической регрессии и дискриминанта Фишера не являются оптимальными;
- модель нормальных распределений с равными ковариационными матрицами не является моделью максимального правдоподобия для дискриминанта Фишера;
- дискриминант Фишера превосходит методы SVM и логистической регрессии на многих моделях с распределениями, далекими от нормального.

Полученные результаты позволяют лучше понять особенности исследуемых методов, что дает возможности для их дальнейшего усовершенствования.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Сравнение алгоритмов распознавания с помощью программной системы «Полигон» // Анализ данных и знаний в экспертных системах. — Новосибирск: Вычислительные системы, 1990. Вып. 134. С. 56–66.

[2] *Неделько В.М.* Регрессионные модели в задаче классификации // Сиб. ж. индустриальной математики, 2014. Т. XVII. № 1. С. 86–98.

[3] *Mease D., Wyner A.* Evidence contrary to the statistical view of boosting // J. Mach. Learn. Res., 2008. Vol. 9. P. 131–156.

[4] *Krasotkina O. V., Mottl V. V., Turkov P. A.* Bayesian aproach to the pattern recognition problem in nonstationary environment // Pattern recognition and machine intelligence / Eds. S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, S. K. Pal. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2011. Vol. 6744. P. 24–29.

[5] *Nedel'ko V. M.* Misclassification probability estimations for linear decision functions // Structural, syntactic, and statistical pattern recognition / Eds. A. Fred, T. M. Caelli, R. P. W. Duin, *et al.* — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2004. Vol. 3138. P. 780–787.

[6] *Неделько В. М.* К вопросу об эффективности бустинга в задаче классификации // Вестник Новосибирского гос. ун-та. Серия: математика, механика, информатика, 2015. Т. 15. Вып. 2. С. 72–89.

[7] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Институт математики СО РАН, 1999. 211 с.

[8] *Nedel'ko V.* Decision trees capacity and probability of misclassification // Autonomous intelligent systems: Agents and data mining / Eds. V. Gorodetsky, J. Liu, V. A. Skormin. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag, 2005. Vol. 3505. P. 193–199.

[9] *Кельманов А. В., Пяткин А. В.* NP-трудность некоторых квадратичных евклидовых задач 2-кластеризации // Докл. РАН, 2015. Т. 464. № 5. С. 535–538.

[10] *Смердов С. О., Витяев Е. Е.* Синтез логики, вероятности и обучения: формализация предсказания // Сиб. электронные математические известия, 2009. Т. 6. С. 340–365.

[11] *Torshin I. Yu., Rudakov K. V.* On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification // Pattern Recogn. Image Anal., 2015. Vol. 25. No. 4. P. 577–587.

[12] *Лисицын Д. В.* Комбинированные регрессионные модели для описания данных, представленных в разных шкалах // Сб. научн. тр. Новосибирского гос. техн. ун-та, 2013. № 3(73). С. 41–48.

[13] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recogn. Image Anal., 2010. Vol. 20. No. 3. P. 269–285.

[14] *Motrenko A., Strijov V., Weber G.-W.* Sample size determination for logistic regression // J. Comput. Appl. Math., 2014. Vol. 255. P. 743–752.

[15] *Friedman J., Hastie T., Tibshirani R.* Additive logistic regression: A statistical view of boosting // Ann. Stat., 2000. Vol. 28. No. 2. P. 337–407.

[16] *Красоткина О. В., Турков П. А., Моттль В. В.* Байесовская логистическая регрессия в задаче обучения распознаванию образов при смещении решающего правила // Изв. Тульского гос. ун-та. Технические науки, 2013. № 2. С. 177–187.

[17] *Zhu J., Hastie T.* Support vector machines, kernel logistic regression and boosting // Multiple classifier systems / Eds. F. Roli, J. Kittler. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer, 2002. Vol. 2364. P. 16–26.

[18] *Середин О. С., Моттль В. В.* Метод опорных объектов для обучения распознаванию образов в произвольных метрических пространствах // Изв. Тульского гос. ун-та. Естественные науки, 2015. № 4. С. 49–66.

[19] *Lugosi G., Vayatis N.* On the Bayes-risk consistency of regularized boosting methods // Ann. Stat., 2004. Vol. 32. No. 1. P. 30–55.

[20] *Liu Y.* Fisher consistency of multicategory support vector machines // 11th Conference (International) on Artificial Intelligence and Statistics Proceedings. — San Juan, Puerto Rico, 2007. Vol. 2. P. 291–298.

[21] *Muandet K., Fukumizu K., Dinuzzo F., Schölkopf B.* Learning from distributions via support measure machines // Advances in neural information processing systems 25 / Eds. F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger. — MIT Press, 2012. P. 10–18.

[22] *Steinwart I.* Consistency of spport vector machines and other regularized kernel classifiers // IEEE Trans. Inform. Theory, 2005. Vol. 51. No. 1. P. 128–142.

[23] *Sollich P.* Bayesian methods for support vector machines: Evidence and predictive class probabilities // Mach. Learn., 2002. Vol. 46. P. 21–52.

[24] *Vojtěch F., Zien A., Schölkopf B.* Support vector machines as probabilistic models // Conference (International) on Machine Learning Proceedings. — New York, NY, USA: ACM, 2011. 665–672.

[25] *Боровков А. А.* О задаче распознавания образов // Теория вероятностей и её применение, 1971. Т. 16. № 1. С. 132–136.

[26] *Seredin O., Mottl V., Tatarchuk A., Razin N., Windridge D.* Convex support and relevance vector machines for selective multimodal pattern recognition // 21st Conference (International) on Pattern Recognition Proceedings. — Tsukuba, Japan, 2012. P. 1647–1650.

[27] *Татарчук А. И.* Байесовские методы опорных векторов для обучения распознаванию образов с управляемой селективностью отбора признаков. Дисс. ... канд. физ.-мат. наук, 2014. 125 с.

[28] *Salazar D. A., Vélez J. I., Salazar J. C.* Comparison between SVM and logistic regression: Which one is better to discriminate? // Rev. Colomb. Estad., 2012. Vol. 35. No. 2. P. 223–237.

Поступила в редакцию 31.08.2016

Investigation of effectiveness of several linear classifiers by using synthetic distributions*

V. M. Nedel'ko

nedelko@math.nsc.ru

S. L. Sobolev Institute of Mathematics SB RAS, 4 Acad. Koptyug Ave., Novosibirsk, Russia

The most common way to compare the effectiveness of data analysis methods is testing on tasks from UCI repository. However, this approach has several disadvantages, in particular, the incompleteness of the set of tasks and limited sample sizes. The present authors consider the possibility of building a repository of probabilistic distributions. The distributions are constructed purposefully in such a way as to reveal properties of the studied methods. Such distributions are called the probabilistic models. Some linear classification methods have been chosen for research: logistic regression, Fisher discriminant, and support vector machine. Several probabilistic models have been constructed to investigate the properties of these methods,

*The research was supported by the Russian Foundation for Basic Research (grants 14-01-00590 and 14-07-00249).

in particular, for each method, there was built a model on which this method outperformed the other methods. In addition, these models allow one to explain why a particular method was the best.

Keywords: *pattern recognition; machine learning; support vector machine; deciding function; logistic regression; misclassification probability*

DOI: 10.21469/22233792.2.3.04

References

[1] Lbov, G. S., and N. G. Starceva. 1990. Sravnenie algoritmov raspoznavaniya s pomoshch'yu programmnoy sistemy "Poligon" [Comparison of recognition algorithms with the software system "Poligon"]. *Analiz dannykh i znaniy v ekspertnykh sistemakh. Vychislitel'nye sistemy* [Analysis of data and knowledge in expert systems. Computer systems]. Novosibirsk. 34:56–66.

[2] Nedel'ko, V. M. 2014. Regressionnye modeli v zadache klassifikatsii [Regression models in the classification problem]. *Sib. zh. industrial'noy matematiki* [Siberian J. Industrial Mathematics] XVII(1):86–98.

[3] Mease, D., and A. Wyner. 2008. Evidence contrary to the statistical view of boosting. *J. Mach. Learn. Res.* 9:131–156.

[4] Krasotkina, O. V., V. V. Mottl, and P. A. Turkov. 2011. Bayesian approach to the pattern recognition problem in nonstationary environment. *Pattern recognition and machine intelligence*. Eds. S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, and S. K. Pal. Lecture notes in computer science ser. Berlin–Heidelberg: Springer-Verlag. 6744:24–29.

[5] Nedel'ko, V. M. 2004. Misclassification probability estimations for linear decision functions. *Structural, syntactic, and statistical pattern recognition*. Eds. A. Fred, T. M. Caelli, R. P. W. Duin, *et al.* Lecture notes in computer science ser. Berlin–Heidelberg: Springer-Verlag. 3138:780–787.

[6] Nedel'ko, V. M. 2015. K voprosu ob effektivnosti bustinga v zadache klassifikatsii [On the boosting efficiency in the classification problem]. *Vestnik Novosibirskogo gos. un-ta. Seriya: Matematika, mekhanika, informatika* [Bull. Novosibirsk State University. Ser. mathematics, mechanics, computer science] 15(2):72–89.

[7] Lbov, G. S., and N. G. Starceva. 1999. *Logicheskie reshayushchie funktsii i voprosy statisticheskoy ustoychivosti resheniy* [Logical decision functions and problem of statistical robustness of the solutions]. Novosibirsk: Institute of Mathematics SB RAS. 211 p.

[8] Nedel'ko, V. 2005. Decision trees capacity and probability of misclassification. *Autonomous intelligent systems: Agents and data mining*. Eds. V. Gorodetsky, J. Liu, and V. A. Skormin. Lecture notes in computer science ser. — Berlin–Heidelberg: Springer-Verlag. 3505:193–199.

[9] Kel'manov, A. V., and A. V. Pyatkin. 2015. NP-hardness of some Quadratic Euclidean 2-clastering problems. *Dokl. Math.* 92(2):634–637.

[10] Smerdov, S. O., and E. E. Vityaev. 2009. Sintez logiki, veroyatnosti i obucheniya: Formalizatsiya predskazaniya [Probability, logic and learning synthesis: Formalizing prediction concept]. *Sibirskie Elektronnye Matematicheskie Izvestiya* [Siberian Electronic Math. Rep.] 6:340–365.

[11] Torshin, I. Yu., and K. V. Rudakov. 2015. On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification. *Pattern Recogn. Image Anal.* 25(4):577–587.

[12] Lisitsin, D. V. 2013. Kombinirovannye regressionnye modeli dlya opisaniya dannykh, predstavlennykh v raznykh shkalakh [Combined regression models for the data represented in different scales]. *Sb. nauchn. tr. Novosibirskogo gos. tekhn. un-ta* [Contributions of the Novosibirsk State Technical University] 3(73):41–48.

[13] Vorontsov, K. V. 2010. Exact combinatorial bounds on the probability of overfitting for empirical risk minimization. *Pattern Recogn. Image Anal.* 20(3):269–285.

[14] Motrenko, A., V. Strijov, and G.-W. Weber. 2014. Sample size determination for logistic regression. *J. Comput. Appl. Math.* 255:743–752.

[15] Friedman, J., T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *Ann. Stat.* 28(2):337–407.

[16] Krasotkina, O. V., P. A. Turkov, and V. V. Mottl'. 2013. Bayesovskaya logisticheskaya regressiya v zadache obucheniya raspoznavaniyu obrazov pri smeshchenii reshayushchego pravila [Bayesian logistic regression in the problem of pattern recognition learning on shifting decision rule] // *Izv. Tul'skogo gos. un-ta. Tehnicheskie nauki* [Proceedings of the Tula State University. Engineering] 2:177–187.

[17] Zhu, J., and T. Hastie. 2002. Support vector machines, kernel logistic regression and boosting. *Multiple classifier systems*. Eds. F. Roli and J. Kittler. Lecture notes in computer science ser. Berlin–Heidelberg: Springer. 2364:16–26.

[18] Seredin, O. S., and V. V. Mottl'. 2015. Metod opornykh ob"ektov dlya obucheniya raspoznavaniyu obrazov v proizvol'nykh metricheskikh prostranstvakh [The method of support objects for pattern recognition in arbitrary metric spaces]. *Izv. Tul'skogo gos. un-ta. Estestvennye nauki* [Proceedings of the Tula State University. Natural Sciences] 4:49–66.

[19] Lugosi, G., and N. Vayatis. 2004. On the Bayes-risk consistency of regularized boosting methods. *Ann. Stat.* 32(1):30–55.

[20] Liu, Y. 2007. Fisher consistency of multicategory support vector machines. *11th Conference (International) on Artificial Intelligence and Statistics Proceedings*. San Juan, Puerto Rico. 2:291–298.

[21] Muandet, K., K. Fukumizu, F. Dinuzzo, and B. Schölkopf. 2012. Learning from distributions via support measure machines. *Advances in neural information processing systems 25*. Eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. MIT Press. 10–18.

[22] Steinwart, I. 2005. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory* 51(1):128–142.

[23] Sollich, P. 2002. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Mach. Learn.* 46:21–52.

[24] Vojtěch, F., A. Zien, and B. Schölkopf. 2011. Support vector machines as probabilistic models. *Conference (International) on Machine Learning Proceedings*. New York, NY: ACM. 665–672.

[25] Borovkov, A. A. 1971. On the problem of pattern recognition. *Theor. Probab. Appl.* 16(1):141–144.

[26] Seredin, O., V. Mottl, A. Tatarchuk, N. Razin, and D. Windridge. 2012. Convex support and relevance vector machines for selective multimodal pattern recognition. *21st Conference (International) on Pattern Recognition Proceedings*. Tsukuba, Japan. 1647–1650.

[27] Tatarchuk, A. I. 2014. Bayesovskie metody opornykh vektorov dlya obucheniya raspoznavaniyu obrazov s upravlyaemoy selektivnost'yu otbora priznakov [Bayesian methods of support vector machines for pattern recognition training with controlled selectivity feature selection]. PhD Thesis. 125 p.

[28] Salazar, D. A., J. I. Vélez, and J. C. Salazar. 2012. Comparison between SVM and logistic regression: Which one is better to discriminate? // *Rev. Colomb. Estad.* 35(2):223–237.

Received August 31, 2016