Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 974 ГРУППЫ

«Совместный выбор объектов и признаков при построении моделей в задачах банковского скоринга»

Выполнил:
студент 4 курса 974 группы
Адуенко Александр Александрович

Научный руководитель:
к.ф-м.н., н.с. ВЦ РАН
Стрижов Вадим Викторович

Москва, 2013

Содержание

Аннотация

Для построения скоринговой модели нужно решить несколько проблем. Первая: как выбрать самые важные вопросы в анкете; иначе говоря, самые важные признаки в модели. Вторая: как составить представительную группу клиентов или группу информативных объектов. Более того искомое множество вопросов может отличаться для разных социальных групп, то есть данные могут описываться несколькими моделями, а потому требуется найти число этих моделей. Для решения этих задач использовался мультимодельный подход и предложенный алгоритм отбора объектов и признаков, основанный на анализе ковариационной матрицы оценок параметров модели. Работа алгоритма проиллюстрирована на данных по потребительским кредитам, данных по сердечным заболеваниям. Мультимодельный подход проиллюстрирован также на данных конкурса Яндекс Интернет-математика 2009.

Ключевые слова: *потребительский кредит, кредитный скоринг, вероятность дефолта, число моделей, совместный выбор объектов и признаков.*

1 Введение

Актуальность темы. Задача кредитного скоринга становится все более актуальной вместе с распространением и широким использованием разного рода кредитов, особенно потребительских. Если кредитные риски крупных компаний оцениваются международными рейтинговыми агенствами на основании публичной отчетности, а потому решение о выдаче кредита и процентной ставке принять относительно легко, то никаких общепризнанных данных о «кредитном рейтинге» отдельного заемщика нет. А поэтому решение о выдаче кредита и ставке принимается либо экспертным путем, либо с помощью некоторой скоринговой системы. Под скоринговой системой подразумевается автоматизированная система, которая по предоставленным заемщиком данным оценивает вероятность дефолта по кредиту.

При проектировании таких систем возникает несколько основных проблем. Во-первых, нужно выделить информацию, которую нужно получить от заемщика. Ясно, что список потенциальных вопросов очень широк. Однако большая часть собранных данных не коррелирует с текущей платежеспособностью, да и правдивость ответов на многие сложно проверить. В связи с этим возникает проблема отбора важных данных — признаков, которые с одной стороны коррелируют с платежеспособностью, а с другой стороны — легко проверить на подлинность.

Второй проблемой является выбор информативного множества клиентов. Клиентские базы содержат миллионы записей, часть из которых имеют ошибки. Невозврат кредита может быть обусловлен не какими-то причинами, имевшимися на момент получения кредита, а чем-то произошедшим после, что было трудно или невозможно предсказать. Данные таких клиентов не нужно использовать при построении скоринговой модели, поскольку они ухудшат прогноз [1].

Третьей проблемой является неоднородность данных. Для регионов с равномерно высокими доходами у населения можно предположить, что уровень дохода не столь важен для возврата кредита, как в регионах с низким средним доходом и высоким имущественным расслоением. В работе решается задача разделения объектов на однородные совокупности, определения их числа и построения своей модели для каждой совокупности.

Альтернативой скоринговой системе является использование экспертов для принятия решений по кредитам. За год крупный российский банк выдает несколько миллионов потребительских кредитов на общую сумму в сотни миллиардов рублей, из которых 5-16% не возвращают [1]. Улучшение показателя возврата даже на несколько процентов позволит банку сэкономить несколько миллиардов рублей. Поэтому внедрение скоринговых систем, которые автоматически проводят отбор объектов и информативных признаков, а также строят модели данных, в банковском кредитовании оправдано.

Цель работы. Построить алгоритм классификации объектов, определяющий требуемое число моделей для описания данных и их параметры, а также проводящий фильтрацию выбросов и отбор информативных признаков.

Методы исследований. При построении алгоритма использовались методы оценки характеристик распределения по выборке, проверки гипотез о виде распределения, методы обработки категориальных признаков. Для программной реализации

разработанного алгоритма использовалась среда MATLAB.

Научная новизна.

- Предложена модификация правила выбора модели на обучении в многоуровневых моделях.
- Предложена модификация алгоритма многоклассовой логистической регрессии для ранжирования документов внутри классов.
- Разработан алгоритм отбора признаков, основанный на предложенной модификации метода Белсли.
- Разработан алгоритм отбора выбросов с использованием ковариационной матрицы оценок параметров.

Практическая ценность. Разработан программный модуль, который

- фильтрует выбросы;
- отбирает признаки;
- строит многоуровневые модели или смесь логистических моделей;
- определяет требуемое количество моделей;
- визуализирует результаты.

Положения, выносимые на защиту:

- Алгоритм отбора признаков, основанный на предложенной модификации метода Белсли.
- Алгоритм отбора выбросов с использованием ковариационной матрицы оценок параметров.
- Модификация алгоритма многоклассовой логистической регрессии для ранжирования документов внутри классов.

Апробация. Результаты квалификационной работы бакалавра были использованы для решения задачи ранжирования поисковой выдачи Яндекса [23], а также применены для классификации на два класса для данных по немецким [21] и венгерским [22] потребительским кредитам и заболеваниям сердца в Южной Африке [24].

Обзор литературы. В качестве модели, использующейся для определения вероятности невозврата рассмотрим модель логистической регрессии [2,3]. Тогда получим, что каждая модель, построенная по данным, будет зависеть от двух множеств: некоторого множества объектов и некоторого множества признаков, которые использовались для ее построения. Рассмотрим задачу отбора признаков. В качестве базового алгоритма, который будет применяться для отбора признаков будет использоваться генетический алгоритм [6,7]. Путем несложной модификации генетические алгоритмы использованы для совместного отбора объектов и признаков. Другим подходом к отбору признаков является подход, основанный на оценке информативности признаков [8], при котором в модели оставляют некоторое количество наиболее информативных признаков. Также для отбора признаков используют шаговые алгоритмы [9].

В качестве еще одного базового алгоритма используются разные модификации SVM с отбором признаков. Отбор объектов в SVM производится автоматически: происходит выделение некоторого множества опорных объектов. К отбору же признаков в SVM существует два подхода: фильтрация и внедрение отбора внутрь алгоритма классификации [15,16]. В качестве возможного способа внедрения отбора внутрь алгоритма классификации предлагается изменение оптимизационной задачи, возникающей при нахождении весов признаков в SVM [11,12]. Существует также алгоритм, являющийся синтезом предложенных двух базовых, в котором генетический алгоритм применяется для отбора признаков при классификации с помощью SVM [14].

Для решения задачи отбора объектов и признаков применяют регуляризацию [13]. При этом вводится штраф за сложность модели. В работе [13] рассматриваются разные виды штрафов и указывается, что целесообразно применять смеси непрерывных штрафов за рост модулей параметров модели и дискретных, связанных с числом объектов и признаков в ней. Разные способы регуляризации для задачи многоклассовой логистической регрессии рассматриваются в [17]. В этой же работе указано на быстрое разреживание модели (на скорую стабилизацию числа объектов и признаков) при применении регуляризации. Результаты и время работы сравниваются с работой SVM с отбором признаков и пакета SVM^{lite}.

Подходом, связанным с регуляризацией, является подход, в котором векторы параметров моделей предполагаются случайными векторами из некоторого априорного распределения [18,19]. При этом вместо оптимизации обычной функции правдоподобия происходит оптимизация совместной функции правдоподобия данных и модели. Оптимизация совместной функции правдоподобия обычно улучшает свойства оптимизируемой функции, уменьшая число обусловленности гессиана, поскольку, например, в случае предположения о нормальной априорной плотности распределения параметров, фактически происходит регуляризация оптимизируемой функции.

В данной работе для отбора объектов и признаков предлагается итерационный алгоритм. При этом категориальные признаки проходят предварительную обработку [26]. На каждой итерации алгоритма будет оцениваться матрица ковариаций параметров модели, с помощью анализа свойств которой по аналогии с методом Белсли для линейной регрессии [19] и производился отбор признаков. Затем производился отбор объектов. При этом объекты предполагаются независимыми. Случай скореллированных объектов для одной модели рассмотрен в [9]. Для проверки того, что повышение качества в терминах функционала AUC [27] статистически значимо, сэмплировалось распределение AUC по выборке. Для проверки нормальности полученного эмпирического распределения AUC использовался критерий Шапиро-Уилка [25].

Для решения проблемы неоднородности данных предполагается построение своей модели для каждой однородной части данных. Предложено правило для определения необходимого числа моделей в алгоритме построения смеси моделей [2,4] и для многоуровневых моделей [28, 29]. Существуют также алгоритмы построения иерархических моделей [5], но в данной работе они не рассматриваются.

2 Постановка задачи

Имеются исходные данные – выборка $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$. По исходным данным составим матрицу признаков $\mathbf{X} = [\mathbf{x}_1^\intercal, \ \ldots, \ \mathbf{x}_m^\intercal] \in \mathbb{R}^{m \times n}$, ($m$–число записей данных, n–количество признаков) и вектор ответов $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$, $y_i \in \{0, 1\}$. Здесь 1 означает, что заемщик кредит не вернул, а 0— вернул. Для индексации признаков введем обозначение $\mathcal{J} = \{1, \ \ldots, \ n\}$.

Введем разбиение $\mathcal{I} = \mathcal{S} \sqcup \mathcal{T}$ исходной выборки на обучающую выборку $S\{(x_i, y_i)\}$, $i \in \mathcal{S}$ и тестовую $T\{(x_i, y_i)\}$, $i \in \mathcal{T}$. Будем осуществлять это разбиение случайно.

Предполагается, что каждый элемент y_i есть реализация бернуллевской случайной величины $Y_i \ \sim \ Be(p_i)$, $\{Y_i\}_{i=1}^m$ независимы в совокупности. При этом вероятность невозврата кредита p_i заемщиком определяется с помощью модели логистической регрессии

$$p_i = f(\mathbf{x}_i, \mathbf{w}) = f(\mathbf{x}_i^\intercal \mathbf{w}) = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\intercal \mathbf{w}\right)}. \tag{1}$$

Здесь $\mathbf{w} \in \mathcal{W} = \mathbb{R}^n$ вектор параметров модели. В работе $\mathbf{w}$ считается фиксированным вектором параметров, значение которого нужно оценить по выборке. Для нахождения оценки $\hat{\mathbf{w}}$ вектора параметров $\mathbf{w}$ будет использоваться метод наибольшего правдоподобия (2).

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \in \mathbb{R}^n} L(\mathbf{w}|\mathbf{X}, \ \mathbf{y}), \tag{2}$$

где $L(\mathbf{w}|\mathbf{X}, \mathbf{y})$ функция правдоподобия данных.

Введем обозначения $\mathbf{X}(\mathcal{B}, \ \mathcal{A})$ и $\mathbf{y}(\mathcal{B})$, соответствующие усеченной матрице объект-признак $\mathbf{X}(\mathcal{B}, \ \mathcal{A}) = [x_{ij}]$, $i \in \mathcal{B} \subseteq \mathcal{I}$, $j \in \mathcal{A} \subseteq \mathcal{J}$ и усеченному вектору ответов $\mathbf{y}(\mathcal{B}) = [y_i]$, $i \in \mathcal{B} \subseteq \mathcal{I}$. Тогда задачу отбора объектов и признаков можно записать в виде

$$[\mathcal{B}, \ \mathcal{A}] = \arg\max_{\mathcal{B} \in \mathcal{S}, \ \mathcal{A} \in \mathcal{J}} L(\hat{\mathbf{w}}|\mathbf{X}(\mathcal{T}, \ \mathcal{A}), \ \mathbf{y}(\mathcal{T})), \tag{3}$$

где

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w}|\mathbf{X}(\mathcal{B}, \ \mathcal{A}), \ \mathbf{y}(\mathcal{B})). \tag{4}$$

Считается также, что данные могут описываться не одной моделью, а несколькими, а потому ставится задача определения числа логистических моделей, описывающих данные. Эта задача решается с помощью EM – алгоритма с последовательным увеличением числа моделей.

2.1 Функция правдоподобия и ее свойства

Запишем функцию правдоподобия данных для одной модели и покажем, что она является вогнутой, а потому имеет единственный максимум. В силу предположения о независимости в совокупности случайных величини $\{Y_i\}_{i=1}^m$ для функции правдоподобия имеем выражение

$$L(\mathbf{w}|\mathbf{X}(\mathcal{B},\,\mathcal{A}),\,\mathbf{y}(\mathcal{B})) = \prod_{t\in\mathcal{B}} f(\mathbf{x}_t^\intercal(\mathcal{A})\mathbf{w})^{y_t}\,(1 - f(\mathbf{x}_t^\intercal(\mathcal{A})\mathbf{w}))^{1-y_t}. \tag{5}$$

Рассмотрим функцию

$$l(\mathbf{w}|\mathbf{X}(\mathcal{B},\,\mathcal{A}),\,\mathbf{y}(\mathcal{B})) = -\ln L(\mathbf{w}|\mathbf{X}(\mathcal{B},\,\mathcal{A}),\,\mathbf{y}(\mathcal{B})). \tag{6}$$

Далее для удобства записи опустим $\mathcal{B}$ и $\mathcal{A}$ в выражении (6), введем обозначение $r = |\mathcal{B}|$ и без ограничения общности предположим $\mathcal{B} = \{1,\,\ldots,\,r\}$, тогда

$$l(\mathbf{w}|\mathbf{X},\,\mathbf{y}) = -\sum_{t=1}^{r} y_t \ln f(\mathbf{x}_t^\intercal\mathbf{w}) + (1 - y_t)\ln(1 - f(-\mathbf{x}_t^\intercal\mathbf{w})). \tag{7}$$

Продифференцируем (7) по $\mathbf{w}$ с учетом $f^{'}(x) = f(x)\cdot f(-x)$, получим

$$\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} = -\sum_{t=1}^{r} \mathbf{x}_t(y_t - f(\mathbf{x}_t^\intercal\mathbf{w})) = \mathbf{X}^\intercal(\mathbf{f} - \mathbf{y}). \tag{8}$$

Найдем гессиан $l(\mathbf{w}|\mathbf{X},\,\mathbf{y})$ и покажем, что он положительно определен, откуда и получим выпуклость $l(\mathbf{w}|\mathbf{X},\,\mathbf{y})$, а в силу монотонности преобразования $g(x) = -\ln(x)$, получим вогнутость $L(\mathbf{w}|\mathbf{X},\,\mathbf{y})$ и единственность максимума функции правдоподобия.

$$\mathbf{H} = \frac{\partial^2 l(\mathbf{w})}{\partial \mathbf{w}^2} = \sum_{t=1}^{r} \mathbf{x}_t f(\mathbf{x}_t^\intercal\mathbf{w}) f(-\mathbf{x}_t^\intercal\mathbf{w})\mathbf{x}_t^\intercal = \mathbf{X}^\intercal\mathbf{R}\mathbf{X}, \tag{9}$$

где $\mathbf{R}$ — диагональная матрица с элементами $f(\mathbf{x}_t^\intercal\mathbf{w})f(-\mathbf{x}_t^\intercal\mathbf{w}) > 0$ на диагонали. Рассмотрим произвольный вектор $\mathbf{u} \neq \mathbf{0}$.

$$\mathbf{u}^\intercal\mathbf{H}\mathbf{u} = \mathbf{u}^\intercal\mathbf{X}^\intercal\mathbf{R}\mathbf{X}\mathbf{u} = (\mathbf{X}\mathbf{u})^\intercal\,\mathbf{R}\,(\mathbf{X}\mathbf{u}) > 0,$$

откуда $\mathbf{H}$ положительно определенная, то есть $l(\mathbf{w})$ выпуклая, а потому имеет единственный минимум. Перейдем к алгоритму нахождения этого минимума.

2.2 Оценка параметров скоринговой модели

В случае линейной регрессии существует явное выражение для оценки наибольшего правдоподобия, поскольку в этом случае функция $l(\mathbf{w})$ квадратичная. В силу нелинейности сигмоидной функции получить явное выражение для $\hat{\mathbf{w}}$ не удается, но можно предложить итерационную процедуру, основанную на методе Ньютона-Рафсона для нахождения оценки наибольшего правдоподобия $\hat{\mathbf{w}}$ для вектора параметров логистической модели.

На начальном шаге задается вектор $\hat{\mathbf{w}}_0$ оценок параметров. На каждом следующем шаге вычисляется новое приближение $\hat{\mathbf{w}}_i$ к оценке наибольшего правдоподобия $\hat{\mathbf{w}}$ по формуле

$$\hat{\mathbf{w}}_i = \hat{\mathbf{w}}_{i-1} - \mathbf{H}^{-1}\nabla l(\hat{\mathbf{w}}_{i-1}) = \hat{\mathbf{w}}_{i-1} - (\mathbf{X}^\intercal \mathbf{R}\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{f} - \mathbf{y}). \quad (10)$$

Вычисления по формуле (10) продолжаем до тех пор, пока изменение нормы вектора $\mathbf{w}$ не перестанет быть значительным.

3 Выбор оптимального множества объектов и признаков

3.1 Алгоритм устойчивого отбора признаков

Оценка ковариационной матрицы $\hat{\mathbf{w}}$. Оценим ковариационную матрицу оценок параметров, чтобы в дальнейшем использовать ее для отбора признаков. Заметим, что эта полученная матрица в общем случае не является несмещенной оценкой ковариационной матрицы параметров, поскольку та зависит определяется еще априорным распределением параметров (18). С учетом того, что в точке $\hat{\mathbf{w}}$ $\nabla l(\hat{\mathbf{w}}) = 0$, пользуясь формулой Тейлора, получим

$$\ln \frac{L(\mathbf{w})}{L(\hat{\mathbf{w}})} \approx -\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\intercal \mathbf{H}(\mathbf{w} - \hat{\mathbf{w}}). \quad (11)$$

Пользуясь локально нормальной аппроксимацией $\hat{\mathbf{w}}$, получим, что $\hat{\mathbf{w}} \sim N(\mathbf{w}_0, \mathbf{H}^{-1})$, где $\mathbf{w}_0 = \mathbf{w}$. Таким образом, оценкой ковариационной матрицы оценок параметров является матрица $\mathbf{H}^{-1}$.

Модификация метода Белсли для задачи логистической регрессии. Воспользуемся найденной оценкой ковариационной матрицы $\hat{\mathbf{w}}$ для отбора признаков. Для матрицы $\mathbf{H}$ имеем

$$\mathbf{H} = \mathbf{X}^\intercal \mathbf{R}\mathbf{X} = \tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}},$$

где введено обозначение

$$\tilde{\mathbf{X}} = \tilde{\mathbf{R}}\mathbf{X},$$

а $\tilde{\mathbf{R}}$ диагональная матрица, на диагонали которой стоят корни из соответствующих диагональных элементов диагональной матрицы $\mathbf{R}$. Фактически матрица $\tilde{\mathbf{R}}$ задает параметры на множестве объектов, при этом больший вес получают объекты с неустойчивой классификацией, поскольку $r_{ii} = p_i(1 - p_i)$. Выполним сингулярное разложение матрицы $\mathbf{X}$

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal.$$

Тогда

$$\mathbf{var}(\mathbf{w}) = \mathbf{H}^{-1} = (\tilde{\mathbf{X}}^\intercal \tilde{\mathbf{X}})^{-1} = (\mathbf{V}\mathbf{\Lambda}\mathbf{U}^\intercal \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal)^{-1} = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^\intercal.$$

Таким образом, дисперсия j-го регрессионного коэффициента — это j-й диагональный элемент матрицы $\mathbf{H}^{-1}$.

Для обнаружения мультиколлинеарности признаков построим таблицу, в которой каждому индексу обусловленности η_j соответствуют значения q_{ij} — долевые коэффициенты. Сумма долевых коэффициентов по индексу j равна единице.

$$\mathbf{var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2} = (q_{i1} + q_{i2} + \ldots + q_{in}) \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2},$$

где q_{ij} — отношение соответствующего слагаемого в разложении вектора $\mathbf{var}(w_i)$ ко всей сумме, а $\mathbf{V} = (v_{ij})$.

Таблица 1: Разложение $\mathbf{var}(\mathbf{w})$

Индекс обусловленности	$\mathbf{var}(w_1)$	$\mathbf{var}(w_2)$	...	$\mathbf{var}(w_n)$
η_1	q_{11}	q_{21}	$\cdots$	q_{n1}
η_2	q_{12}	q_{22}	$\cdots$	q_{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
η_n	q_{1n}	q_{2n}	$\cdots$	q_{nn}

Чем больше значение долевого коэффициента q_{ij}, тем больший вклад вносит j-ый признак в дисперсию i-го регрессионного коэффициента.

Из табл. 1 определяется мультиколлинеарность: большие величины η_j означают, что, возможно, есть зависимость между признаками. Если один из признаков является линейной комбинацией остальных, то матрица $\tilde{\mathbf{X}}$ будет матрицей неполного ранга, а потому у матрицы $\mathbf{\Lambda}$ одним из собственных значений будет 0. В случае мультиколлинеарности признаков матрица $\mathbf{\Lambda}$ будет иметь близкие к нулю собственные значения, которым соответствуют большие коэффициенты обусловленности. На этом и будет основан метод отбора признаков, изложенный ниже.

Алгоритм отбора признаков, основанный на методе Белсли. Заметим, что чем больше параметров в модели, тем точнее она описывает имеющиеся данные, тем меньше будет ее применимость для произвольных данных, то есть наблюдается переобучение. Поэтому если при последовательном добавлении признаков учитывать только внутренний критерий качества модели на обучающей выборке, то чтобы уменьшить переобучение, нужно требовать значительного роста качества модели после добавления признака (фактически происходит введение штрафа за сложность модели). В данной работе использовался другой подход.

Опишем два этапа алгоритма: Add и Del. На первом этапе (Add) последовательно добавляются признаки. На втором этапе (Del) происходит последовательное удаление признаков, согласно методу Белсли. Пусть перед началом работы алгоритма модель задается множеством индексов объектов $\mathcal{B} \subseteq \mathcal{S}$ и множеством индексов признаков $\mathcal{A} \subseteq \mathcal{J}$. Опишем как построить новую модель, то есть определить новое множество индексов признаков $\tilde{\mathcal{A}}$. Множество индексов объектов $\mathcal{B} \subseteq \mathcal{S}$, соответствующих модели, остается неизменным. Выделим из контрольной выборки объектов, задаваемой множеством индексов $\mathcal{T}$, некоторую подвыборку, задаваемую множеством индексов $\mathcal{T}_1 \subseteq \mathcal{T}$, то есть $\mathcal{T} = \mathcal{T}_1 \sqcup \mathcal{T}_2$. Выделенная выборка будет применяться для отбора признаков, а выборка, задаваемая индексным множеством $\mathcal{T}_2$ — для контроля качества модели.

Этап Add. Для каждого $j \in \mathcal{J} \setminus \mathcal{A}$ найдем $\hat{\mathbf{w}}_j$ согласно (4)

$$\hat{\mathbf{w}}_j = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|+1}} L(\mathbf{w}|\mathbf{X}(\mathcal{B},\, \mathcal{A} \cup \{j\}),\, \mathbf{y}(\mathcal{B}))$$

и сосчитаем согласно (6)

$$l(j) = l(\hat{\mathbf{w}}_j|\mathbf{X}(\mathcal{T}_1,\, \mathcal{A} \cup \{j\},\, \mathbf{y}(\mathcal{T}_1)).$$

Найдем признак $j^* \in \mathcal{J} \setminus \mathcal{A}$, для которого l_j принимает минимальное значение. Обозначим значение той же функции для модели с множеством признаков $\mathcal{A}$ как l_0, то есть

$$l_0 = l(\hat{\mathbf{w}}_0 | \mathbf{X}(\mathcal{T}_1,\, \mathcal{A},\, \mathbf{y}(\mathcal{T}_1)),$$

где

$$\hat{\mathbf{w}}_0 = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w} | \mathbf{X}(\mathcal{B},\, \mathcal{A}),\, \mathbf{y}(\mathcal{B})).$$

Если $l_{j^*} - l_0 \leq Z_1 < 0$, где $Z_1 \in \mathbb{R}$, то есть произошло значительное улучшение качества модели, признак j^* добавляется в модель, то есть $\mathcal{A} \to \mathcal{A} \cup \{j^*\}$. Этап повторяется до тех пор, пока происходит добавление признаков в модель.

Этап Del. После того, как добавить признаки в модель на этапе Add же не получается, переходим к этапу Del. Находим индексы обусловленности и долевые коэффициенты для текущего набора признаков $\mathcal{A}$ согласно методу Белсли, описание которого приведено выше. Далее находим количество достаточно больших индексов обусловленности. Достаточно большими будем считать индексы, квадрат которых превосходит максимальный индекс обусловленности η_t, где $t = |\mathcal{A}|$, количество признаков в текущем наборе $\mathcal{A}$. Количество таких индексов обозначим

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right]. \tag{12}$$

Затем ищем в матрице долевых коэффициентов $\mathbf{var}(\mathbf{w})$ столбец j^* с максимальной суммой по последним i^* долевым коэффициентам

$$j^* = \arg\max_{j \in \mathcal{A}} \sum_{g=t-i^*+1}^{t} q_g^j. \tag{13}$$

Как и для этапа Add считаем по обучающей выборке $\hat{\mathbf{w}}_0$ и $\hat{\mathbf{w}}_{j^*}$

$$\begin{aligned} \hat{\mathbf{w}}_0 &= \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w} | \mathbf{X}(\mathcal{B},\, \mathcal{A}),\, \mathbf{y}(\mathcal{B})), \\ \hat{\mathbf{w}}_{j^*} &= \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|-1}} L(\mathbf{w} | \mathbf{X}(\mathcal{B},\, \mathcal{A} \setminus \{j^*\}),\, \mathbf{y}(\mathcal{B})), \end{aligned}$$

считаем l_0 и l_{j^*}

$$\begin{aligned} l_0 &= l(\hat{\mathbf{w}}_0 | \mathbf{X}(\mathcal{T}_1,\, \mathcal{A},\, \mathbf{y}(\mathcal{T}_1)), \\ l_{j^*} &= l(\hat{\mathbf{w}}_{j^*} | \mathbf{X}(\mathcal{T}_1,\, \mathcal{A} \setminus \{j^*\},\, \mathbf{y}(\mathcal{T}_1)). \end{aligned} \tag{14}$$

Если $l_{j^*} - l_0 \leq Z_2$, то есть не происходит значительного ухудшения качества модели, признак j^* исключается из модели, то есть $\mathcal{A} \to \mathcal{A} \setminus \{j^*\}$. Этап повторяется до тех пор, пока происходит удаление признаков из модели.

Поочередное повторение этапов Add и Del осуществляется до тех пор, пока происходит удаление или добавление признаков. Заметим, что условием остановки алгоритма является $Z_1 + Z_2 < 0$.

Генетический алгоритм совместного отбора объектов и признаков. Опишем итеративный алгоритм, который применялся для решения задачи отбора объектов и признаков (3). Будем характеризовать набор индексов использующихся объектов и признаков $\mathcal{B}$, $\mathcal{A}$ вектором $\mathbf{b}$ из 0 и 1 размерности $|\mathcal{B} \sqcup \mathcal{A}| = n + m$. Пусть перед r-ой итерацией алгоритма есть некоторый набор векторов $V = \{\mathbf{b}_1, \ldots, \mathbf{b}_v\}$, где $v = |V|$. Каждому вектору $\mathbf{b}_i$ из V сопоставим число

$$e_i = l(\hat{\mathbf{w}}(\mathcal{B}(\mathbf{b}_i), \mathcal{A}(\mathbf{b}_i))|\mathbf{X}(\mathcal{T}, \mathcal{A}), \mathbf{y}(\mathcal{T})).$$

Исключим из V долю α векторов с наибольшими значениями e_i и заменим их дубликатами векторов с долей α с наименьшими значениями e_i. Затем разобьем векторы на пары $\{(\mathbf{b}_{i_k}, \mathbf{b}_{j_k})\}_{k=1}^{\lfloor \frac{v}{2} \rfloor}$(если v нечетно, то один вектор может остаться без пары). Внутри каждой пары проведем операцию скрещивания, которая заменяет пару векторов на некоторую другую пару векторов. Правила этой замены будут описаны ниже. Затем с каждым из получившихся векторов с вероятностью p происходит мутация, то есть случайный бит $\mathbf{b}$ меняется на противоположный.

Опишем операцию скрещивания. Рассмотрим пару векторов $\mathbf{b}_i = (b_1^i, \ldots, b_{n+m}^i)^\intercal$, $\mathbf{b}_j = (b_1^j, \ldots, b_{n+m}^j)^\intercal$. Сгенерируем случайное натуральное число $z \in 1, \ldots, n + m$. Тогда результатом операции скрещивания, примененной к векторам $\mathbf{b}_i$ и $\mathbf{b}_j$ будет пара векторов $\mathbf{b}_i^{'} = (b_1^i, \ldots, b_{z-1}^i b_z^j, \ldots, b_{n+m}^j)^\intercal$ и $\mathbf{b}_j^{'} = (b_1^j, \ldots, b_{z-1}^j b_z^i, \ldots, b_{n+m}^i)^\intercal$.

Для генерации начальной совокупности подвекторов признаков воспользуемся χ^2 тестом Пирсона. Признак с номером j будет появляться в векторах из совокупности пропорционально величине $1 - \alpha_j + \varepsilon$, $\varepsilon > 0$, то есть признаки, для которых уровень значимости близок к единице (то есть наименее информативные) будут реже появляться в совокупности на начальном шаге. При этом наличие $\varepsilon > 0$ не дает совсем их исключить из совокупности, что необходимо, поскольку незначимость признака в отдельности, как было показано в примере выше, еще не значит, что этот признак не информативен совместно с остальными.

Генетический алгоритм останавливается либо после заданного числа итераций, либо когда значение целевой функции перестанет существенно меняться. Полученный на последнем шаге $\hat{\mathbf{w}}$ и применяется для классификации объектов.

SVM с отбором признаков. SVM – алгоритм отбирает множество опорных объектов. Требуется его несколько модифицировать, чтобы происходил отбор не только объектов, но и признаков. В алгоритме SVM штрафуется l_2–норма вектора $\mathbf{w}$, ставится следующая задача безусловной оптимизации

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^n}{argmin} \left[\lambda \sum_{i=1}^{m} (1 - y_i(\mathbf{w}^\intercal \mathbf{x}_i))_+ \right] + (1 - \lambda)\mathbf{w}^\intercal \mathbf{w}.$$

Эта задача многоэкстремальна, а потому вместо нее рассматривают задачу квадратичного программирования, двойственную к исходной задаче оптимизации с ограничениями. Для того, чтобы происходил и отбор признаков, рассмотрим $l_2 - l_1 - SVM$.

В этом методе рассматривается следующая задача оптимизации с ограничениями

$$\min_{\mathbf{w}\in\mathbb{R}^n,\,\boldsymbol{\xi}\in\mathbb{R}^m,\,\mathbf{v}\in\mathbb{R}^n}\left[\frac{\mu}{m}\mathbf{e}^\intercal\boldsymbol{\xi}+\frac{1}{2}\mathbf{w}^\intercal\mathbf{w}+\nu\mathbf{e}^\intercal\mathbf{v}\right],$$
$$y_i\mathbf{w}^\intercal\mathbf{x}_i\geq 1-\xi_i,\ i=1,\ \ldots,\ m, \tag{15}$$
$$\xi_i\geq 0,\ -\mathbf{v}\leq\mathbf{w}\leq\mathbf{v},$$

где $\mu>0$ — некоторое заданное число, $\mathbf{e}=[1\ldots1]^\intercal\in\mathbb{R}^m$, неравенства $-\mathbf{v}\leq\mathbf{w}\leq\mathbf{v}$ понимаются покомпонентно, а ξ, $\mathbf{v}$ – вспомогательные переменные. Однако, как и в случае с SVM лучше рассматривать двойственную задачу [11, 12], поскольку она содержит меньшее число переменных и имеет более простую структуру по сравнению с исходной задачей.

Найденный вектор весов признаков $\hat{\mathbf{w}}$ далее применяется для классификации объектов

$$a(\mathbf{x})=\operatorname{sign}(\hat{\mathbf{w}}^\intercal\mathbf{x}).$$

3.2 Предлагаемый метод отбора объектов и фильтрация выбросов

Отбор объектов наряду с отбором признаков очень важен для построения качественной и устойчивой модели. Под выбросом будем понимать объект, добавление которого в модель значимо влияет на ее параметры. Для того, чтобы определить влияние объекта $(\mathbf{x}_i,\ y_i)$ на модель определим его *специфичность* следующим образом

$$Sp(\mathbf{x}_i)=(\Delta_i\mathbf{w})^\intercal\mathbf{H}(\Delta_i\mathbf{w}), \tag{16}$$

$$\Delta_i\mathbf{w}=\hat{\mathbf{w}}_i-\hat{\mathbf{w}},$$

$\mathbf{H}^{-1}$ – ковариационная матрица оценок параметров модели $\hat{\mathbf{w}}$, а $\hat{\mathbf{w}}_i$ и $\hat{\mathbf{w}}$ оценки параметров модели, полученные согласно (4) по выборке без и с объектом $(\mathbf{x}_i,\ y_i)$ соответственно, то есть

$$\hat{\mathbf{w}}=\arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}}L(\mathbf{w}|\mathbf{X}(\mathcal{S},\ \mathcal{A}),\ \mathbf{y}(\mathcal{S})), \tag{17}$$
$$\hat{\mathbf{w}}_i=\arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}}L(\mathbf{w}|\mathbf{X}(\mathcal{S}\setminus\{i\},\ \mathcal{A}),\ \mathbf{y}(\mathcal{S}\setminus\{i\})),$$

где $\mathcal{A}$ – некоторое ранее отобранное множество признаков, а $\mathcal{S}$, как и ранее, множество индексов объектов обучения.

Как показано в разделе 3.1, оценки параметров модели $\hat{\mathbf{w}}$ локально нормальны, то есть $\hat{\mathbf{w}}\sim N(\mathbf{w}_0,\ \mathbf{H}^{-1})$, то есть в условиях предположения о том, что объект $(\mathbf{x}_i,\ y_i)$ не является выбросом, $\Delta_i\mathbf{w}\sim N(\mathbf{0},\ \mathbf{H}^{-1})$. Матрицу $\mathbf{H}$ считаем невырожденной (хотя, возможно, плохо обусловленной), а потому в условиях предположения о том, что рассматриваемый объект $(\mathbf{x}_i,\ y_i)$ не является выбросом, получаем

$$Sp(\mathbf{x}_i)=(\Delta_i\mathbf{w})^\intercal\mathbf{H}(\Delta_i\mathbf{w})\sim\chi^2(|\mathcal{A}|),$$

где $|\mathcal{A}|$ – число признаков в модели. Задавая уровень значимости α, можно отбирать некоторую долю объектов, которые являются выбросами.

Модификация алгоритма отбора объектов. В случае, когда матрица $\mathbf{H}$ является вырожденной число степеней свободы будет определяться $rg(\mathbf{H})$, а не $|\mathcal{A}|$. Однако определение точного числа степеней свободы затруднительно в силу вычислительных ошибок. То есть требуется определить некоторый порог $\lambda_0 \geq 0$, что собственное число λ матрицы $\mathbf{H}$ считается нулевым, если $\lambda \leq \lambda_0$. Кроме того, в этом случае затруднено использование метода IRLS, поскольку в нем требуется обращать матрицу $\mathbf{H}$. Для стабильности алгоритма IRLS можно заменить матрицу $\mathbf{H}$ на матрицу $\mathbf{H} + \tau\mathbf{I}$, где $\mathbf{I}$ – единичная матрица соответствующего размера. Это будет соотвествовать априорному предположению о нормальном распределении параметров с ковариационной матрицей $\frac{1}{\tau}\mathbf{I}$ и математическим ожиданием $\mathbf{w}_0$, то есть $\mathbf{w} \sim N(\mathbf{w}_0,\, \tau\mathbf{I})$. В таком случае апостериорное распределение вектора оценок весов модели будет нормальным с ковариационной матрицей

$$\hat{\mathbf{w}} \sim N\left(\mathbf{w}_0,\, \left(\mathbf{H} + \tfrac{1}{\tau}\mathbf{I}\right)^{-1}\right). \tag{18}$$

Выполняя соответствующую замену матрицы $\mathbf{H}$ на матрицу $\mathbf{H} + \dfrac{1}{\tau}\mathbf{I}$ в формуле (16), получаем обновленное определение *специфичности* объекта.

$$Sp(\mathbf{x}_i) = (\Delta_i \mathbf{w})^\intercal \left(\mathbf{H} + \tfrac{1}{\tau}\mathbf{I}\right)(\Delta_i \mathbf{w}). \tag{19}$$

При $\tau \to \infty$ (19) совпадает с (16).

При введении указанной регуляризации появляется неопределенность с выбором τ. Для того, чтобы избежать этой неопределенности, предлагается рассмотреть для каждого признака с номером j оценку дисперсии его параметры

$$D_j = \frac{\sum_{i \in \mathcal{S}} (\Delta_i w_j)^2}{|\mathcal{S}| - 1},$$

а специфичность определить как

$$\hat{Sp}(\mathbf{x}_i) = \sum_j \frac{(\Delta_i w_j)^2}{D_j}. \tag{20}$$

Если при этом специфичности, определенные по формулам (20) и (16), будут иметь подобную зависимость от номера объекта, то имеет смысл применять вторую, как вычислительно более простую и не требующую введения регуляризации на случай плохо обусловленной или вырожденной матрицы $\mathbf{H}$. Именно этот случай и реализуется на практике, как показал вычислительный эксперимент (см.рис. 10).

Отделение малочисленной шумовой компоненты от основной выборки. Если в выборке есть шумовая компонента, влияние которой не является определяющим, то можно ее выделить, с помощью введенной специфичности объектов. Для этого отсортируем значения специфичности объектов по убыванию и нарисуем график зависимости специфичности от номера объекта в отсортированном наборе (см. рис. 9). Найдем i^*

$$i^* = \arg\max_i \frac{Sp(\mathbf{x}_i)}{Sp(\mathbf{x}_{i+1})},$$

тогда объект $\mathbf{x}_i$ считаем шумовым объектом, если $Sp(\mathbf{x}_i) \geq Sp(\mathbf{x}_{i^*})$.

4 Мультимодельный подход к задаче классификации

4.1 Смесь логистических моделей

Зачастую данные бывают неоднородны, могут представлять совокупность нескольких разных, но схожих между собой наборов объектов. Эта неоднородность может быть вызвана разными причинами. Например, можно предположить, что на возврат кредита состоятельными гражданами величина их дохода влияет слабее (так как они вполне платежеспособны и имеют возможность вернуть кредит), чем более бедными. Эти и подобные замечания обуславливают неоднородность данных.

Для того, чтобы учесть неоднородность, можно было бы предложить разбиение исходного множества объектов на несколько подмножеств. В каждом из этих подмножеств можно построить свою модель логистической регрессии. При появлении очередного объекта нужно решить, к какой совокупности он относится и применять соответствующую этой совокупности модель. Таким образом, наблюдается разделение объектов между моделями. При таком подходе требуется определить правила отнесения объекта к той или иной модели, а также могут наблюдаться проблемы с классификацией объектов, «лежащих далеко от центров» всех совокупностей. Особенности таких (многоуровневых моделей) будут рассмотрены в следующем параграфе и в соответствующем разделе вычислительного эксперимента.

Здесь предлагается модель с мягким разделением между моделями или смесь моделей. При таком разделении для каждой модели есть вероятность $\pi_k \in [0,\, 1]$ того, что объект описывается этой моделью. Предположим, что число моделей $K \geq 1$, каждой из которых соответствует свой вектор весов признаков $\mathbf{w}_k$. Тогда

$$p(\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K|\mathbf{x}_i,\, y_i) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}. \tag{21}$$

Введем вектор $\boldsymbol{\pi} = [\pi_1 \ldots \pi_K]^\intercal$. Тогда по аналогии с (5) получим для правдоподобия данных выражение

$$L(\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \boldsymbol{\pi}|\mathbf{X},\, \mathbf{y}) = \prod_{i=1}^{m} \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i} \right). \tag{22}$$

Введем бинарную матрицу скрытых переменных $\mathbf{Z} = \|z_{ik}\|$ размеров $m \times K$, при этом z_{ik} определяет принадлежность $\mathbf{x}_i$ модели с номером k. Тогда полная функция правдоподобия будет записана в виде

$$L(\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, ,\boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{y}) = \prod_{i=1}^{m} \prod_{k=1}^{K} \{\pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}\}^{z_{ik}}. \tag{23}$$

Далее опишем применяемый EM-алгоритм для оценки параметров моделей $\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$ и их вероятностей $\pi_1,\, \ldots, \pi_K$. Выберем некоторые начальные приближения для $\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$. На E-шаге считаем апостериорные вероятности каждой из компонент смеси для каждого объекта $\mathbf{x}_i$ γ_{ik}

$$\gamma_{ik} = \mathbb{E}[z_{ik}] = p(k|\mathbf{x}_i,\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \boldsymbol{\pi}) = \frac{\pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}}{\sum_{j=1}^{K} \pi_j f(\mathbf{x}_i,\, \mathbf{w}_j)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_j))^{1-y_i}}.$$

Запишем ожидаемое значение отрицательного логарифма полной функции правдоподобия

$$\tilde{l}(\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K,\,\boldsymbol{\pi}|\mathbf{X},\,\mathbf{y}) = \mathbb{E}_{\mathbf{Z}}[-\log L(\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K,\,\mathbf{Z}|\mathbf{X},\,\mathbf{y})] =$$
$$= -\sum_{i=1}^{m}\sum_{k=1}^{K}\gamma_{ik}\{\log\pi_k + y_i\log(f(\mathbf{x}_i,\,\mathbf{w}_k)) + (1-y_i)\log(1-f(\mathbf{x}_i,\,\mathbf{w}_k))\}. \quad (24)$$

На M-шаге происходит минимизация функции $\tilde{l}(\mathbf{w}_1,\ldots,\mathbf{w}_K,\boldsymbol{\pi}|\mathbf{X},\mathbf{y})$ по $\mathbf{w}_1,\ldots,\mathbf{w}_K,\boldsymbol{\pi}$ при ограничении $\sum_{k=1}^{K}\pi_k = 1$. Решение задачи минимизации для $\boldsymbol{\pi}$ в явном виде дает

$$\pi_k = \frac{1}{m}\sum_{i=1}^{m}\gamma_{ik}.$$

Заметим, что при фиксированных $\{\gamma_{ik}\}$ функция $\tilde{l}(\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K,\,\boldsymbol{\pi}|\mathbf{X},\,\mathbf{y})$ представима в виде

$$\tilde{l}(\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K,\,\boldsymbol{\pi}|\mathbf{X},\,\mathbf{y}) = -\sum_{k=1}^{K}\{\log\pi_k\sum_{i=1}^{m}\gamma_{ik}\} + \sum_{k=1}^{K}\tilde{l}_k(\mathbf{w}_k|\mathbf{X},\,\mathbf{y}). \quad (25)$$

Поэтому минимизация каждой из функций $\tilde{l}_k(\mathbf{w}_k|\mathbf{X},\,\mathbf{y})$ одного из векторов весов $\mathbf{w}_k$ производится независимо от остальных с помощью описанного выше метода IRLS. Но изменится выражение для градиента и гессиана. Опуская выкладки, соответствующие дифференцированию, запишем результат.

$$\frac{\partial\tilde{l}_k}{\partial\mathbf{w}_k} = \mathbf{X}^\intercal\boldsymbol{\Gamma}_k(\mathbf{f}-\mathbf{y}), \quad (26)$$
$$\mathbf{H}_k = \mathbf{X}^\intercal\mathbf{R}_k\mathbf{X}, \quad (27)$$

где $\boldsymbol{\Gamma}_k$ – диагональная матрица с элементами γ_{ik} на диагонали, а $\mathbf{R}_k$ – диагональная матрица с элементами $\gamma_{ik}f(\mathbf{x}_i^\intercal\mathbf{w}_k)f(-\mathbf{x}_i^\intercal\mathbf{w}_k)$ на диагонали.

Так как исходная задача минимизации разбивается на K независимых подзадач, возможно применение алгоритма совместного отбора объектов и признаков, приведенного в этой работе для каждой задачи в отдельности.

Определение требуемого числа моделей. Опишем, как подобрать требуемое число моделей K. Заметим, что чем больше число моделей, тем точнее будет соответствие построенной смеси имеющимся данным и при большом числе моделей в смеси обобщающая способность будет минимальна, поэтому важно найти некоторое небольшое количество моделей K, описывающих данные. Для этой цели воспользуемся модификацией EM-алгоритма с последовательным добавлением моделей.

На начальном шаге число моделей $K = 1$, веса моделей $\pi_1 = 1$ и имеется единственная модель логистической регрессии. На каждом следующем шаге выделяются плохо описанные смесью объекты, то есть объекты правдоподобие которых в соответствии с (21) более, чем в $\alpha > 1$ раз меньше максимального правдоподобия объектов

$$\tilde{\mathcal{A}} = \left\{i : \frac{p(\mathbf{w}_1,\,\ldots,\,\mathbf{w}_{K-1}|\mathbf{x}_i,\,y_i)}{\max_j p(\mathbf{w}_1,\,\ldots,\,\mathbf{w}_{K-1}|\mathbf{x}_j,\,y_j)} < \frac{1}{\alpha}.\right\}$$

Если $|\tilde{\mathcal{B}}| < m_0$, где $m_0 \leq m$ — заданный параметр, определяющий максимальное возможное число шумовых объектов, то алгоритм завершает свою работу. Число моделей на рассматриваемом шаге и есть искомое число моделей K. Иначе производим построение K - й компоненты смеси:

$$\begin{aligned} \pi_K &= \frac{|\tilde{\mathcal{B}}|}{m}, \\ \pi_j \to (1 - \pi_K)\pi_j,\ j \leq K - 1, \\ \mathbf{w}_K &= \mathbf{0}. \end{aligned} \tag{28}$$

Затем применяем приведенный выше EM-алгоритм для смеси K логистических моделей. Добавление компонент продолжаем, пока число шумовых объектов не станет меньше, чем m_0.

4.2 Многоклассовая логистическая регрессия как смесь моделей

В случае многоклассовой логистической регрессии отклик $y_i \in Y \subseteq \mathbb{N}_0$, а не $y_i \in \{0, 1\}$. Здесь Y — линейно-упорядоченное конечное множество, состоящее более, чем из одного элемента.

Модель многоклассовой логистической регрессии — параметрическая функция

$$f : (\boldsymbol{\Theta}, \mathbf{x}) \to \hat{y} \in Y, \tag{29}$$

отображающая пару «параметры, объект» в метку класса $\hat{y}$ из множества Y. Для оценки адекватности модели задачи используется функция качества $S(\boldsymbol{\Theta}|\, \mathcal{X}, \mathcal{B}, \mathcal{A})$, где $\boldsymbol{\Theta}$ — набор параметров модели, $\mathcal{X}$ — набор индексов некоторого множества объектов, $\mathcal{A}$ — набор индексов используемых признаков, а $\mathcal{B}$ — набор индексов используемых при обучении объектов.

Поиск оптимального набора параметров $\hat{\boldsymbol{\Theta}}$ осуществляется следующим образом:

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta} \in \mathbb{R}^L} S(\boldsymbol{\Theta}|\, \mathcal{B}, \mathcal{A}, \mathcal{B}), \tag{30}$$

где L — размерность пространства параметров модели.

Параметр $\boldsymbol{\Theta}$ находится путем максимизации качества модели $Q(\boldsymbol{\Theta}|\, \mathcal{X}, \mathcal{B}, \mathcal{A})$ на обучающей выборке $\mathcal{S}$.

4.3 Алгоритм многоклассовой логистической регрессии

Сопоставим каждому классу C_k, $k = 1, \ldots, K$ вектор параметров $\mathbf{w}_\kappa \in \mathbb{R}^n$, где n– число признаков. Тогда для объекта $\mathbf{x}_i$ вероятность попасть в класс C_κ в модели логистической регрессии равна

$$P(C_\kappa|\mathbf{x}_i) = \frac{\exp \mathbf{w}_\kappa^\intercal \mathbf{x}_i}{\sum_{j=1}^K \exp \mathbf{w}_j^\intercal \mathbf{x}_i}. \tag{31}$$

Введем для $P(C_\kappa|\mathbf{x}_i)$ обозначение $y_{i\kappa}$. Для каждого объекта $\mathbf{x}_i$, $i \in \mathcal{I}$ введем целевой вектор $\boldsymbol{t}_i$, где $t_{i\kappa} \in [0,\ 1]$ – принадлежность объекта $\mathbf{x}_i$ классу C_κ. На обучающей

выборке считаем $t_{i\kappa} = 1$, если объект $\mathbf{x}_i$ лежит в классе C_κ, иначе $t_{i\kappa} = 0$. Обозначим целевую матрицу, составленную из $t_{i\kappa}$ $\mathbf{T} = [t_{i\kappa}]$. Запишем функцию правдоподобия выборки, используя (31).

$$p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{i=1}^{N}\prod_{\kappa=1}^{K} P(C_\kappa|\mathbf{x}_i)^{t_{i\kappa}} \prod_{i=1}^{N}\prod_{\kappa=1}^{K} y_{i\kappa}^{t_{i\kappa}}. \tag{32}$$

Запишем отрицательный логарифм функции правдоподобия (32) и поставим задачу его минимизации:

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{\kappa=1}^{K}\sum_{i=1}^{N} t_{i\kappa} \log y_{ik\kappa} \to \min_{\mathbf{w}_1, \ldots, \mathbf{w}_K}. \tag{33}$$

Для нахождения минимума функции (33) рассчитаем ее градиент и гессиан. Введем обозначение $a_\kappa^i = \mathbf{w}_\kappa^\intercal \mathbf{x}_i$. Рассчитаем $\dfrac{\partial a_\kappa^i}{\partial w_j}$ и $\dfrac{\partial y_{i\kappa}}{\partial a_j^i}$.

$$\frac{\partial a_\kappa^i}{\partial w_j} = \mathbf{x}_i \cdot \mathbf{I}_{\kappa j}, \tag{34}$$

где $\mathbf{I}_{\kappa j}$ — элемент единичной матрицы. Запишем $y_{i\kappa}$ через $\{a_j^i\}_{j=1}^K$ следующим образом:

$$y_{ik} = \frac{\exp a_\kappa^i}{\sum_{l=1}^K \exp a_l^i} \tag{35}$$

и с учетом (34) получим:

$$\frac{\partial y_{i\kappa}}{\partial a_j^i} = \frac{\exp a_\kappa^i}{\sum_{l=1}^K \exp a_l^i} \cdot \frac{\partial a_\kappa^i}{\partial a_j^i} - \frac{\exp a_\kappa^i}{\left(\sum_{l=1}^K \exp a_l^i\right)^2} \frac{\partial \left(\sum_{l=1}^K \exp a_l^i\right)}{\partial a_j^i} = y_{i\kappa}\mathbf{I}_{\kappa j} - y_{i\kappa}y_{ij}.$$

Таким образом, получаем:

$$\frac{\partial y_{i\kappa}}{\partial a_j^i} = y_{i\kappa}(I_{\kappa j} - y_{ij}). \tag{36}$$

Из (34) и (36) получаем:

$$\frac{\partial y_{i\kappa}}{\partial \mathbf{w}_j} = \frac{\partial y_{i\kappa}}{\partial a_j^i} \cdot \frac{\partial a_j^i}{\partial \mathbf{w}_j},$$

то есть

$$\frac{\partial y_{i\kappa}}{\partial \mathbf{w}_j} = y_{i\kappa}(\mathbf{I}_{\kappa j} - y_{ij})\mathbf{x}_i. \tag{37}$$

Искомый градиент $\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K)$ имеет вид

$$\begin{aligned}\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{i=1}^{N}\sum_{\kappa=1}^{K} t_{i\kappa}\frac{1}{y_{i\kappa}}\frac{\partial y_{i\kappa}}{\partial \mathbf{w}_j} = -\sum_{i=1}^{N}\sum_{\kappa=1}^{K} t_{i\kappa}\frac{1}{y_{i\kappa}} y_{i\kappa}(I_{\kappa j} - y_{ij})\mathbf{x}_i = \\ = -\sum_{i=1}^{N} t_{ij}\mathbf{x}_i + \sum_{i=1}^{N} y_{ij}\mathbf{x}_i \sum_{\kappa=1}^{K} t_{i\kappa} = \sum_{i=1}^{N}(y_{ij} - t_{ij})\mathbf{x}_i.\end{aligned} \tag{38}$$

Из выражения для градиента (38) и (37) для подматрицы $\mathbf{H}_{\kappa j}$ размера $n \times n$ гессиана $\mathbf{H}$ получаем:

$$\nabla_{\mathbf{w}_\kappa} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \nabla_{\mathbf{w}_\kappa} \left(\sum_{i=1}^{N} (y_{ij} - t_{ij}) \mathbf{x}_i \right) = \sum_{i=1}^{N} \mathbf{x}_i \nabla_{\mathbf{w}_\kappa} y_{ij} = \sum_{i=1}^{N} y_{ij} (\mathbf{I}_{j\kappa} - y_{i\kappa}) \mathbf{x}\mathbf{x}^T. \tag{39}$$

Гессиан $\mathbf{H}$ есть матрица размера $nK \times nK$ вида

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \cdots & \mathbf{H}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{K1} & \cdots & \mathbf{H}_{KK} \end{pmatrix}.$$

Если бы $\mathbf{H}$ была положительно определенной матрицей, то для нахождения оптимального вектора весов можно было бы воспользоваться методом Ньютона-Рафсона:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \mathbf{H}^{-1} \nabla \mathbf{E}, \; \alpha > 0. \tag{40}$$

Однако из (39) получаем, что в каждой строке матрицы $\mathbf{H}$ сумма равна нулю, то есть матрица $\mathbf{H}$ вырождена и метод Ньютона-Рафсона не применим. Поэтому в работе используются методы безусловной минимизации первого порядка.

Найдя векторы $\mathbf{w}_1, \ldots, \mathbf{w}_K$, по формуле (31), найдем для каждого объекта $\mathbf{x}_i$ вероятности $P(C_\kappa | \mathbf{x}_i)$. Класс C_{κ^*}, к которому будет отнесен объект $\mathbf{x}_i$ найдем из условия

$$\kappa^* = \arg\max_{\kappa = 1, \ldots, K} P(C_\kappa | \mathbf{x}_i). \tag{41}$$

Предлагаемая модификация алгоритма многоклассовой логистической регрессии. Заметим, что алгоритм многоклассовой логистической регрессии, описанный выше, классифицирует объекты и потому не подразумевает сравнения объектов внутри классов (объекты из разных классов сравнимы, так как метки классов линейно упорядочены). Это ведет к неустойчивой и часто неправильной классификации объектов, у которых несколько классов близки к выполнению (41). Поэтому далее откажемся от требования того, что $\hat{y}_i$ принадлежит конечному линейно-упорядоченному множеству Y, заменив его на требование $\hat{y}_i$ принадлежит отрезку $[C_1, C_K]$, считая как и ранее классы линейно упорядоченными. Под записью $\hat{y}_i \in [C_1, C_K]$ понимается, что каждому классу C_κ поставлено в соответствие число C_κ, причем в силу линейной упорядоченности $C_1 < C_2 < \ldots < C_K$. Тогда рассматриваемая задача перестает быть задачей классификации, а становится задачей регрессии на отрезок. Однако для ее решения можно использовать полученное ранее решение задачи классификации. С учетом линейной упорядоченности меток классов в качестве оценки $\hat{y}_i$ рассмотрим

$$\hat{y}_i = \sum_{\kappa=1}^{K} C_\kappa P(C_\kappa | \mathbf{x}_i). \tag{42}$$

4.4 Многоуровневые модели

Определение. Многоуровневой регрессионной моделью называется набор регрессионных моделей f_k, $k = 1, \ldots, K$ такой, что при разбиении множества индексов объектов $\mathcal{I} = \sqcup_{k=1}^{K} \mathcal{I}_k$ для всех объектов из $\mathcal{I}_k$ используется модель f_k.

Опишем правило выбора модели на обучении. Запишем правдоподобие модели f_k

$$p(f_k|\mathbf{x}_i, y_i) = \frac{p(f_k, \mathbf{x}_i, y_i)}{p(\mathbf{x}_i, y_i)} = \frac{p(y_i|f_k, \mathbf{x}_i)p(f_k, \mathbf{x}_i)}{p(\mathbf{x}_i, y_i)}.$$

Рассмотрим две модели, без ограничения общности модели f_1 и f_2 и определим, какая из них предпочтительнее для объекта $(\mathbf{x}_i,\ y_i)$. Для этого запишем отношение правдоподобия моделей

$$\frac{p(f_1|\mathbf{x}_i, y_i)}{p(f_2|\mathbf{x}_i, y_i)} = \frac{p(y_i|f_1, \mathbf{x}_i)}{p(y_i|f_2, \mathbf{x}_i)}\frac{p(f_1)}{p(f_2)}.$$

Модель f_1 будет предпочтительнее, чем f_2, если

$$\frac{p(y_i|f_1, \mathbf{x}_i)}{p(y_i|f_2, \mathbf{x}_i)} > 1 \tag{43}$$

и наоборот. В случае K моделей имеем тогда следующее решающее правило отнесения к модели

$$k_i^* = \arg\max_{k\in\{1..K\}} p(y_i|f_k, \mathbf{x}_i).$$

Требуемое число моделей K будем определять так же, как описано выше для смеси логистических моделей.

Процедура выбора модели для объектов контроля. Для объектов контроля предлагается использовать осторожный выбор модели

$$k_i^* = \arg\max_{k\in\{1..K\}} \min_{u\in\{0,1\}} p(u|f_k, \mathbf{x}_i),$$

который в случае логистической регрессии принимает вид

$$k_i^* = \arg\max_{k\in\{1..K\}} \{\min(f(\mathbf{x}_i^\intercal \mathbf{w}_k), f(-\mathbf{x}_i^\intercal \mathbf{w}_k))\}.$$

Преобразуем это выражение и получаем правило отнесения объектов контроля к модели

$$k_i^* = \arg\max_{k\in\{1..K\}} f(-|\mathbf{x}_i^\intercal \mathbf{w}_k|) = \arg\min_{k\in\{1..K\}} f(|\mathbf{x}_i^\intercal \mathbf{w}_k|).$$

С учетом монотонности сигмоидной функции связи получаем окончательное выражение для решающего правила отнесения объектов контроля к моделям

$$k_i^* = \arg\min_{k\in\{1..K\}} |\mathbf{x}_i^\intercal \mathbf{w}_k|. \tag{44}$$

Формула (44) фактически означает, что с точностью до $\|\mathbf{w}\|$ объект контроля относится к той модели, расстояние до разделяющей гиперплоскости которой меньше (см.рис. 1).

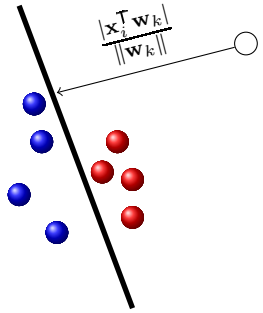


Рис. 1: Иллюстрация отнесения объекта контроля к модели.

Проблемы описанного подхода и их устранение. Проблемой описанного подхода, которая реализуется и на практике (см. рис. 4) является то, что вектор ответов $\mathbf{y}$ на обучении используется не только для подбора вектора параметров размерности $n \ll m$, но и для отнесения объектов к моделям, то есть вводится еще m булевых параметров, которые оптимизируются по обучающей выборке, определяющих отношение объекта каждого объекта обучающей выборки к одной из двух моделей. Из-за наличия этих параметров две модели уже идеально разделяют выборку.

Теорема. Для любой обучающей выборки при описанном методе выбора модели для объектов обучения существует многоуровневая модель, содержащая две модели, идеально разделяющие обучающую выборку.

Доказательство. Для доказательства рассмотрим модели, у которых разделяющие гиперплоскости совпадают, но направления нормали в сторону полупространства, где лежат объекты класса 1, противоположны, то есть $\mathbf{w}_1 = \mathbf{w}$, $\mathbf{w}_2 = -\mathbf{w}$. Тогда при выборе для каждого объекта обучения $(\mathbf{x}_i, y_i)$ в согласии с (43) модели, которой этот объект описывается, будет выбрана модель, которая правильно его классифицирует, поскольку модели f_1 и f_2 имеют противоположные векторы весов $\mathbf{w}_2 = -\mathbf{w}_1$, а потому для каждого объекта ровно одна из двух моделей классифицирует его правильно.

После перераспределения объектов по этим двум моделям дальнейшего перераспределения не происходит, поскольку ситуация, когда правдоподобие многоуровневой модели с ошибками выше, чем правдоподобие модели без ошибок является нереализуемой, как показывает вычислительный эксперимент (см. рис. 4).

Поэтому предлагается для обучающей выборки использовать то же правило (44) отнесения объектов к моделям, что и для контрольной выборки. Это правило не использует известного вектора отклика на обучении $\mathbf{y}$, а потому можно ожидать, что существенного переобучения не будет. Вычислительный эксперимент подтверждает это предположение.

5 Вычислительный эксперимент

5.1 Сравнение алгоритмов на синтетических данных

Сравним описанные алгоритмы смеси логистических моделей и многоуровневую модель с SVM и одной логистической моделью на синтетических данных. В качестве синтетических данных (см.рис. 3) рассмотрим выборку из 2000 объектов

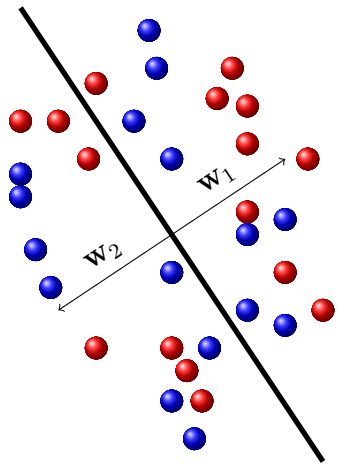


Рис. 2: Две модели, идельно разделяющие любую обучающую выборку на плоскости.

с двумя признаками. Тысяча из этих объектов сгенерирована из распределения $N([-3,\ 0]^\intercal,\ \mathrm{diag}(1,\ 1))$, а еще тысяча из распределения $N([3,\ 0]^\intercal,\ \mathrm{diag}(1,\ 1))$. Из первой тысячи к первому классу были отнесены объекты с положительным вторым признаком, а из второй тысячи — те, у которых первый признак превышает 3. Во всей выборке к классу 1 отнесены 1010 объектов, а классу 0 – 990 объектов.

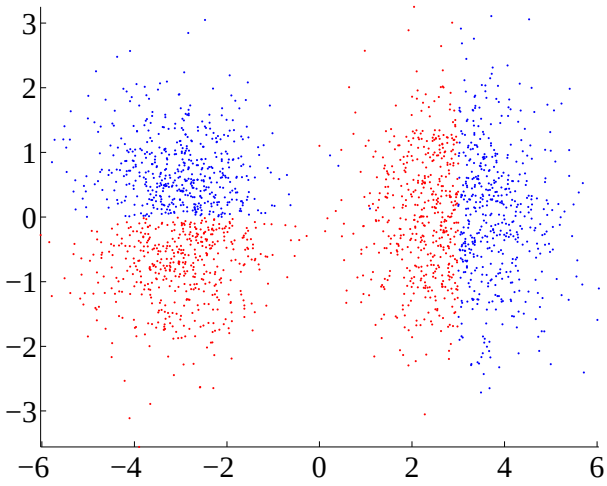


Рис. 3: Синтетическая выборка.

Проиллюстрируем описанный ранее теоретически (см. теорему из раздела 4.4) эффект для многоуровневых моделей, заключающийся в том, что две модели идеально разделяют любую обучающую выборку, если при отнесении обучающих объектов к моделям используется вектор ответов y. Для этого построим с помощью описанного алгоритма две модели, разделяющие синтетические данные. В качестве обучения возьмем 1000 случайных объектов выборки. Синтетическая выборка и две построенные модели приведены на рис. 4 В зазоре между разделяющими прямыми моделей есть объекты только класса 0, а потому полученный результат в точности соответствует описанному теоретически ранее. Все объекты обучения относятся к той из двух моделей, которая их правильно классифицирует. Так как для объектов контроля вектор ответов $\mathbf{y}$ неизвестен, то применение осторожной стратегии выбора

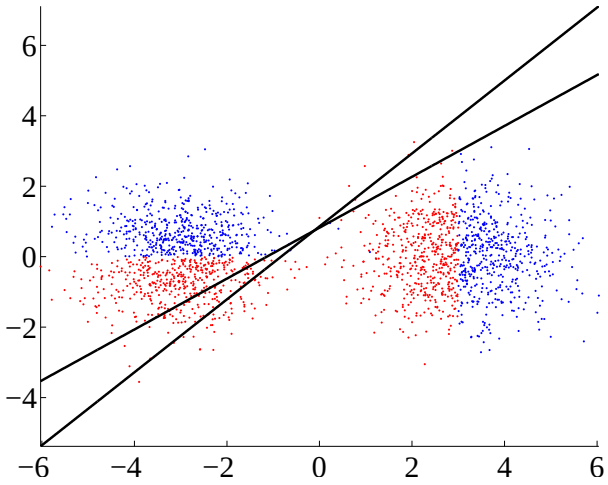


Рис. 4: Иллюстрация переобучения двух многоуровневых моделей.

модели приводит к частой неправильной классификации, так как построенные модели чрезмерно подобраны под обучающие данные с учетом их вектора ответов. Далее используем предложенную для устранения указанного недостатка модификацию метода многоуровневых моделей. Кроме того, добавим также нормировку весов на 1 на каждом шаге оптимизации в EM-алгоритме. Построенные модели, разделяющие синтетическую выборку приведены на рис. 5.

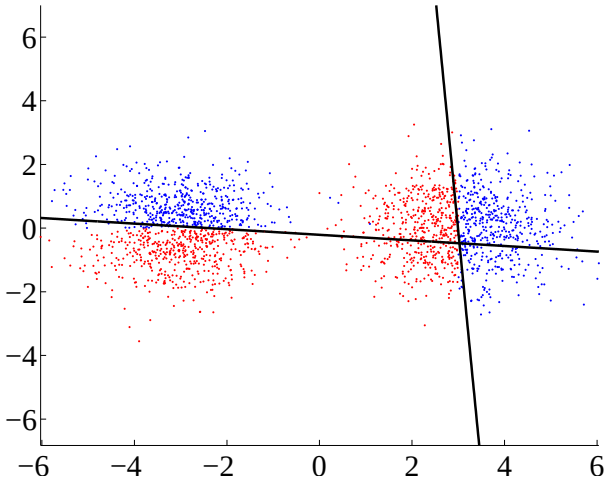
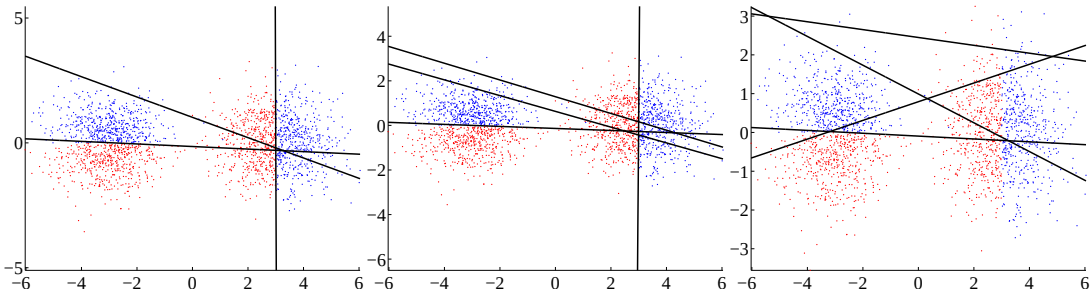


Рис. 5: Иллюстрация устранения переобучения двух многоуровневых моделей.

Приведенный рисунок показывает, что предложенная модификация метода выбора модели на обучении позволяет избежать переобучения. Значения функционала качества AUC на обучении и контроле при этом равно 0.9696 и 0.9460 соответственно. При $K = 3$ значение AUC достигает 0.9771 и 0.9757 на обучении и контроле соответственно. При дальнейшем увеличении числа моделей увеличения качества не происходит, а исходные две модели заменяются на множество более сложных, происходит переусложнение модели. При этом качество близко к значению для двух моделей (см.рис. 6).

Применим на тех же данных смесь моделей и приведем результаты для AUC на обучении и контроле в табл. 2. При этом в соответствующей колонке через точку с



(а) $K = 3$ (б) $K = 4$ (в) $K = 5$

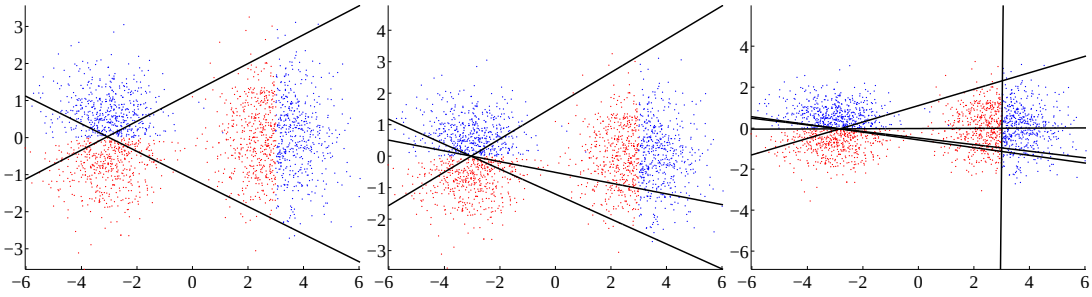
Рис. 6: Классификация выборки с помощью многоуровневых моделей при разном числе моделей

запятой сначала приведены данные на обучении, а затем на контроле.

Таблица 2: Площадь под ROC-кривой при разном числе моделей в смеси моделей.

Количество моделей	AUC
2	0.8357; 0.8349
3	0.8687; 0.8400
5	0.9744; 0.9723

Модели, построенные с помощью смеси моделей отличаются от тех, что построены с помощью многоуровневых моделей. В частности, среди них может не быть моделей, разделяющих каждую из компонент сгенерированной выборки (см. рис. 7).



(а) $K = 2$ (б) $K = 3$ (в) $K = 5$

Рис. 7: Классификация выборки с помощью смеси логистических моделей при разном числе моделей

Приведем далее ROC-кривые для SVM, одной логистической модели, смеси 2, 3 и 5 логистических моделей, а также 2 многоуровневых моделей (см. рис. 8). Соответствующие значения площади под ROC-кривой AUC приведены в табл. 3.

Из табл. 3 получаем, что для достижения сопоставимого качества в смеси моделей требуется их большее число, чем в многоуровневых моделях.

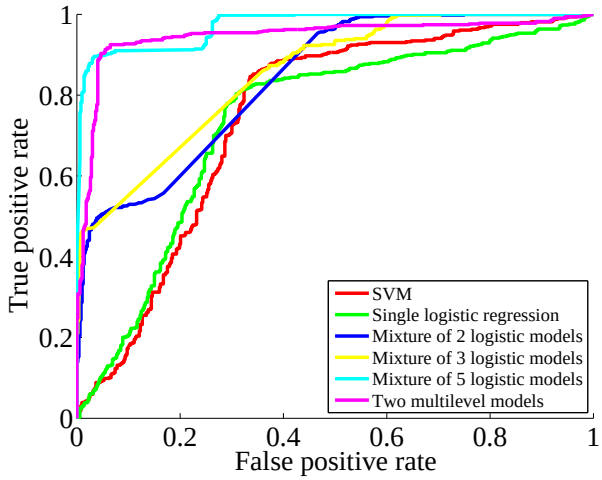


Рис. 8: AUC для разных моделей.

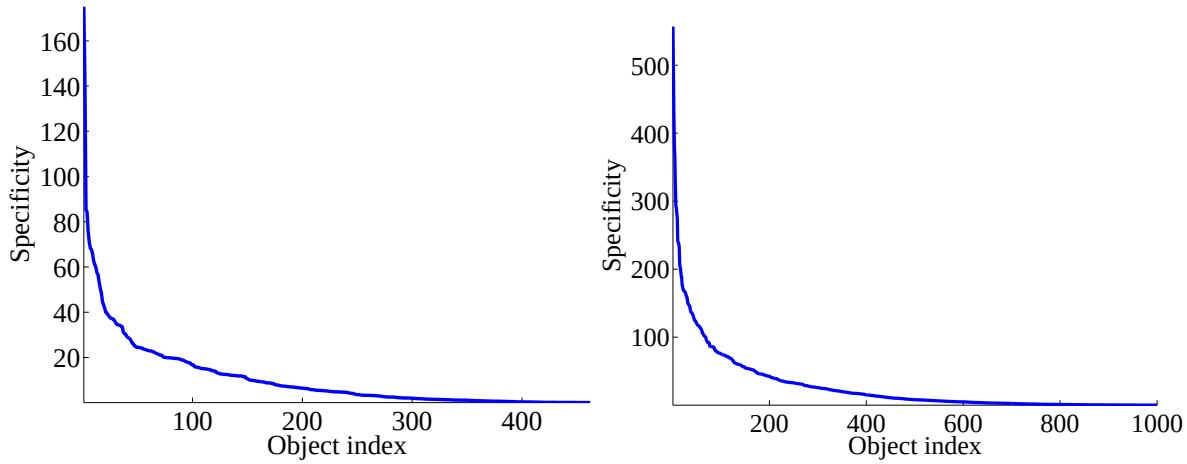
Таблица 3: Площадь под ROC-кривой на обучении и контроле для разных моделей.

Модель	$\mathrm{AUC}_{\mathrm{learn}}$	$\mathrm{AUC}_{\mathrm{test}}$
1 логистическая модель	0.7707	0.7346
SVM	0.7512	0.7496
Смесь двух логистических моделей	0.8357	0.8349
Смесь трех логистических моделей	0.8687	0.8400
2 многоуровневые модели	0.9696	0.9460
Смесь пяти логистических моделей	0.9744	0.9723
3 многоуровневые модели	0.9771	0.9757

5.2 Проверка значимости повышения качества при фильтрации выбросов

Произведем отбор выбросов с помощью описанного в работе алгоритма, основанного на специфичности объектов (16) и (20). Сравним результаты расчетов с помощью формулы (16) и упрощенной формулы, неиспользующей ковариационную матрицу оценок параметров (20). Для указанных целей будем использовать три выборки реальных данных: данные по заболеваниям сердца (SAHD) [24], данные по немецким потребительским кредитам [21] и данные по венгерским потребительским кредитам. Последние в силу их зашумленности будем использовать для демонстрации применимости предложенного метода для отделения шумовой компоненты данных.

Приведем графики зависимости специфичности от номера объекта после упорядочения объектов по убыванию специфичности (см. рис. 9). В обоих случаях специфичность меняется без скачков и лишь небольшая доля выборки обладает высокой специфичностью. Графики приведены для удобства дальнейшего сравнения с заменой $Sp(\mathbf{x}_i)$ на $mSp(\mathbf{x}_i)$, где m – число объектов, для специфичности, посчитанной по формуле (16). Приведем далее зависимости специфичностей для обоих выборок от



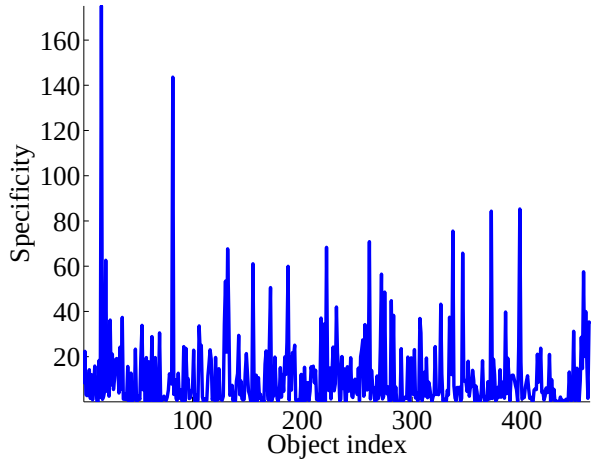
(а) Данные по SAHD.

(б) Данные немецких потребительских кредитов.

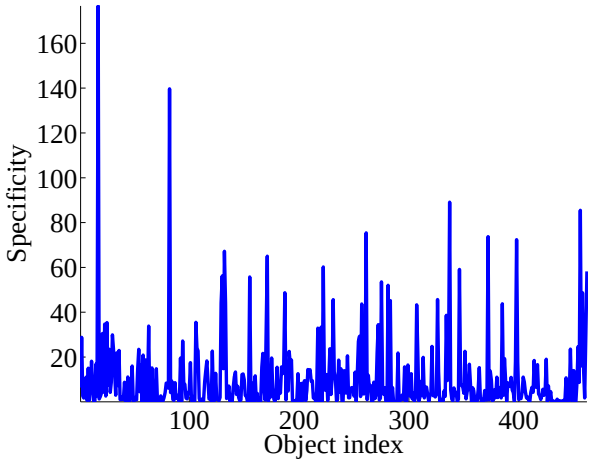
Рис. 9: Зависимость специфичности объектов от номера объекта.

номера объекта без упорядочения по величине специфичности для специфичности, посчитанной по формуле (16) с учетом ковариационной матрицы оценок параметров $\mathbf{H}^{-1}$ и по формуле (20), не использующей этой матрицы (см. рис. 10). В обоих случаях виды зависимости специфичности от номера объекта для формул (16) и (20) мало отличаются, а потому для практики применима формула (20), не использующая плохо обусловленную матрицу $\mathbf{H}$. Приведем далее таблицу изменения площади под ROC-кривой, сосчитанной по всей выборке при удалении небольшой части объектов с максимальной специфичностью, для данных по SAHD и немецким потребительским кредитам.

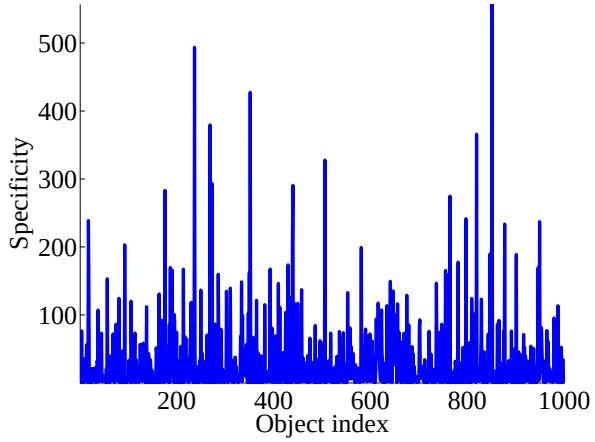
Данные в табл. 4 демонстрируют значительное увеличение качества одной логистической модели по данным в обоих случаях. Покажем, однако, что это увеличение вызвано не только уменьшением объектов в модели, но и качественным изменением их состава, то есть удалением выбросов. Для этого на данных по немецким потребительским кредитам и на данных по заболеваниям сердца в Южной Африке проведем



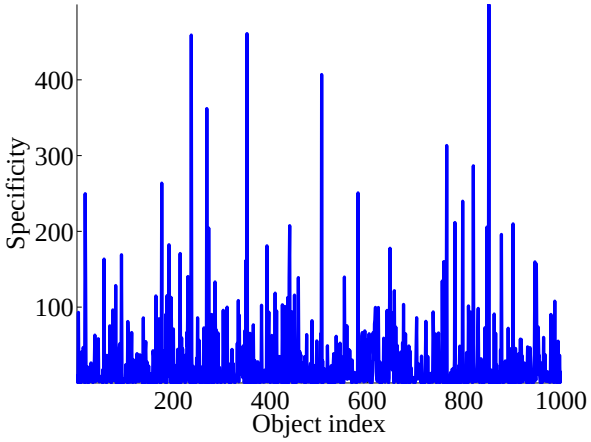
(а) Данные по SAHD. Формула (16)



(б) Данные по SAHD. Формула (20).



(в) Данные кредитов. Формула (16)



(г) Данные кредитов. Формула (20).

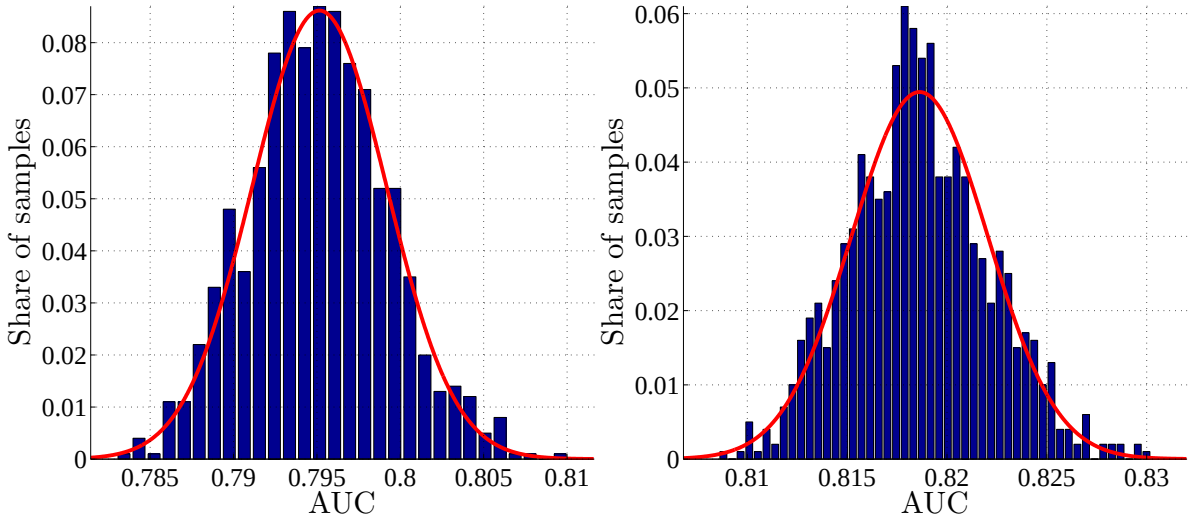
Рис. 10: Зависимость специфичности объектов от номера объекта без упорядочения по специфичности.

Таблица 4: Изменение площади под ROC-кривой при отборе объектов.

Данные	AUC до отбора	AUC после отбора	Количество удаленных объектов
SAHD	0.7948	0.8275	15 из 462
Кредиты	0.8179	0.8779	50 из 1000

следующий эксперимент. Будем сэмплировать равновероятно выборки из 950 и 447 элементов соответственно для указанных данных. По полученной выборке будем искать наилучшую модель и соответствующую ей AUC и проверять гипотезу о том, что полученное улучшение случайно. Достигаемый уровень значимости будем рассчитывать по доле сэмплов, для которых полученное AUC превышает, полученное после удаления выбросов. Однако такое определение достигаемого уровня значимости дает значение p = 0, так как даже при количестве сэмплов $N = 1000$ нет ни одного сэмпла, качества на котором было бы близко к качеству у построенной модели.

Поэтому воспользуемся нормальной аппроксимацией построенного эмпирического распределения. Для проверки нормальности будем использовать критерий Шапиро-Уилка. Гистрограммы эмпирической плотности распределения и их нормальные аппроксимации приведены на рис. 11. Далее приведем следующие харак-



(а) Данные по SAHD.

(б) Данные кредитов.

Рис. 11: Эмпирическое распределение AUC и его нормальная аппроксимация для данных кредитов и сердечных заболеваний.

теристики построенного эмпирического распределения и его нормальной аппроксимации для данных по немецким потребительским кредитам и данным по сердечным заболеваниям (см. табл. 5):

- Достигаемый уровень значимости в критерии Шапиро-Уилка для проверки нормальности p_{SW},
- Оценка математического ожидания для AUC $\hat{m}$,
- Оценка дисперсии для AUC $\hat{\sigma}^2$,
- Оценка стандартного отклонения для AUC $\hat{\sigma}$,
- Отклонение AUC модели без выбросов от математического ожидания, в стандартных отклонениях отклонениях, $\mathrm{M} = \dfrac{AUC - \hat{m}}{\hat{\sigma}}$,

- Достигаемый уровень значимости p_0 для проверки гипотезы случайности полученного улучшения качества.

Таблица 5: Характеристики эмпирического распределения и его нормальной аппроксимации.

Характеристики \ Данные	Немецкие кредиты	Сердечные заболевания
p_{SW}	0.0317	0.2035
$\hat{m}$	0.8186	0.7951
$\hat{\sigma}^2$	$1.17 \cdot 10^{-5}$	$1.63 \cdot 10^{-5}$
$\hat{\sigma}$	0.00342	0.00404
M	17.33	7.39
p_0	0	0

Из приведенной таблицы получаем, что на уровне значимости $\alpha = 0.05$ гипотеза нормальности для данных по кредитам отвергается, однако построенные оценки тем не менее будем использовать. Гипотеза нормальности для данных по сердечным заболеваниям не отвергается. Достигаемый уровень значимости при проверки гипотезы о случайности полученного улучшения качества равен 0 с машинной точностью, а потому полученное улучшение качества статистически значимо. В таблице для иллюстрации также приведено отклонение в большую сторону от математического ожидания для качества модели с отобранными выбросами в стандартных отклонениях. То, что эти значения очень велики, как раз и иллюстрирует неслучайность полученного повышения качества.

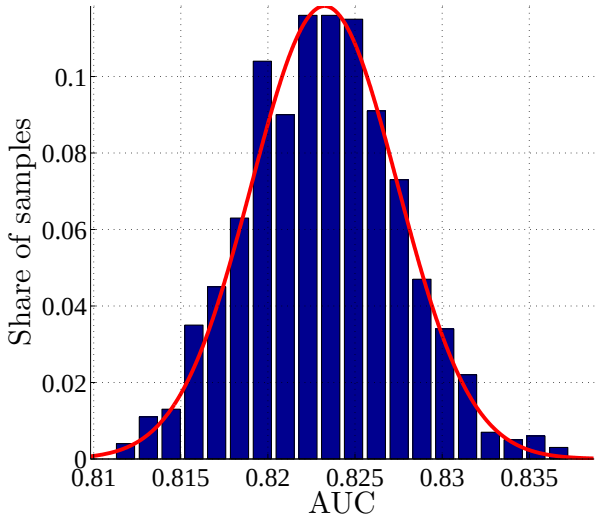
Приведем далее результаты подобных экспериментов, когда проводилось разбиение на обучающую и тестовую выборку выборок полученных удалением заданного числа объектов из всей выборки (50 объектов для данных по немецким потребительским кредитам и 15 для данных по сердечным заболеваниям). Для данных по немецким потребительским кредитам размер обучения равнялся 690, для данных по сердечным заболеваниям – 300 объектов. При этом генерировалось 1000 сэмплов и для каждого сэмпла проводилось 50 случайных разбиений объектов сгенерированного сэмпла на обучение и контроль. Полученные AUC для каждого сэмпла усреднялись и снова строилось эмпирическое распределение, которое аппроксимировалось нормальным. Приведем усредненные значения AUC на обучении и контроле для каждой из выборок при использовании фильтрации выбросов (см. табл. 6).

Приведем гистрограммы эмпирической плотности распределения и их нормальные аппроксимации для обучения и контроля (см. рис. 12). Далее укажем, как и ранее, характеристики построенных эмпирических распределений и их нормальных аппроксимаций для данных по немецким потребительским кредитам и данным по сердечным заболеваниям (см. табл. 7). При этом в каждой ячейке значения на обучении и контроле укажем через точку с запятой.

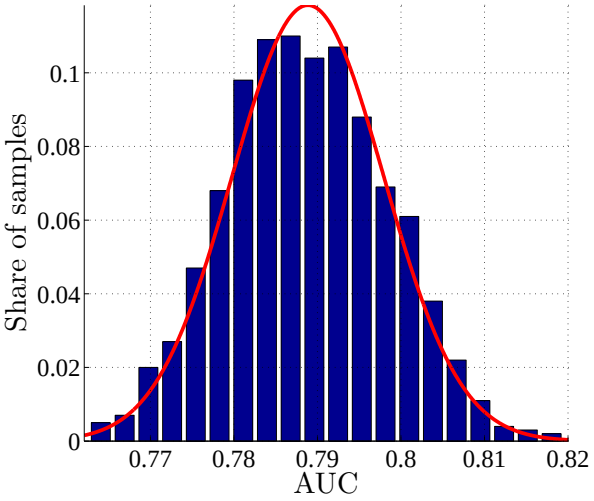
Из приведенной таблицы получаем, что на уровне значимости $\alpha = 0.05$ гипотеза нормальности AUC для обоих выборок данных не отвергается. Достигаемый уровень значимости при проверки гипотезы о случайности полученного улучшения качества

Таблица 6: Полученные значения AUC для выборок на обучении и контроле после отбора выбросов.

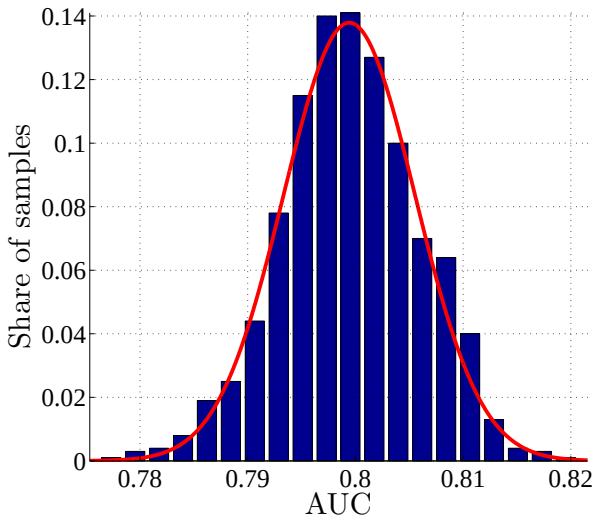
Данные / Характеристики	Немецкие кредиты	Сердечные заболевания
AUC_{learn}	0.8819	0.8507
AUC_{test}	0.8308	0.8093



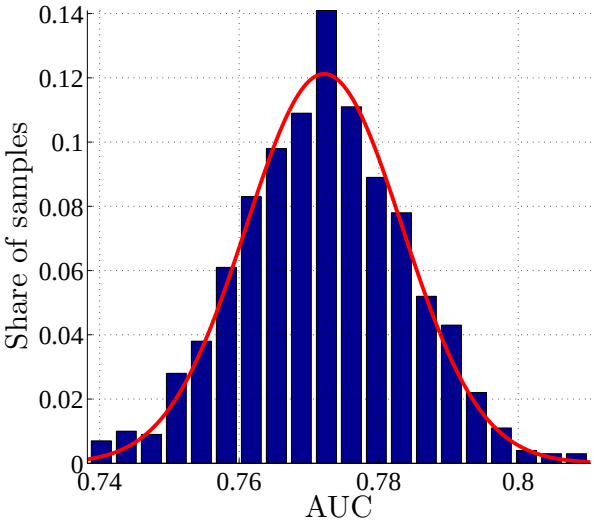
(а) Данные кредитов. Обучение.



(б) Данные кредитов. Контроль.



(в) Данные по SAHD. Обучение.



(г) Данные по SAHD. Контроль.

Рис. 12: Эмпирическое распределение AUC на обучении и контроле и его нормальная аппроксимация для данных кредитов и сердечных заболеваний.

Таблица 7: Характеристики эмпирических распределений AUC и их нормальных аппроксимаций на обучении и контроле.

Характеристики \ Данные	Немецкие кредиты	Сердечные заболевания
p_{SW}	0.2655; 0.2364	0.2786; 0.7879
$\hat{m}$	0.8233; 0.7889	0.7994; 0.7722
$\hat{\sigma}^2$	$1.75 \cdot 10^{-5}$; $8.27 \cdot 10^{-5}$	$3.73 \cdot 10^{-5}$; $1.26 \cdot 10^{-4}$
$\hat{\sigma}$	0.0042; 0.0091	0.0061; 0.011
M	14.0; 6.8	5.15; 3.32
p_0	0; $5.3 \cdot 10^{-12}$	$1.33 \cdot 10^{-7}$; $4.55 \cdot 10^{-4}$

близок к нулю, а потому полученное улучшение качества статистически значимо. В таблице для иллюстрации также приведено отклонение в большую сторону от математического ожидания для качества модели с отобранными выбросами в стандартных отклонениях. То, что эти значения очень велики, как раз и иллюстрирует неслучайность полученного повышения качества.

Приведем зависимость специфичности от номера для данных по венгерским потребительским кредитам, которые оказались зашумлены. Для расчета по этим данным понадобилась формула (20), поскольку матрица $\mathbf{H}$ оказалась очень плохо обусловленной, что в рабочем приближении являлась вырожденной.

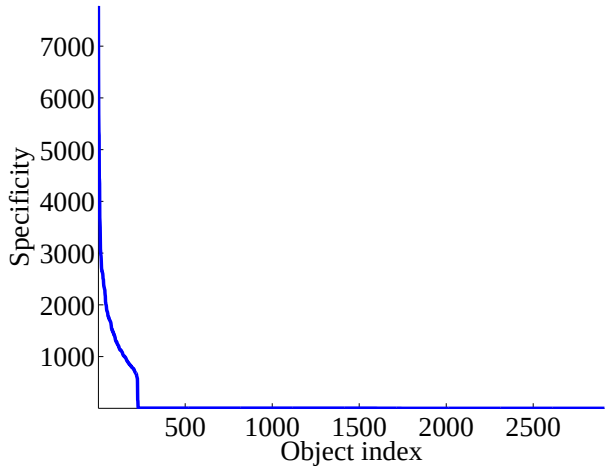


Рис. 13: Зависимость специфичности объектов от номера объекта по данным венгерских потребительских кредитов.

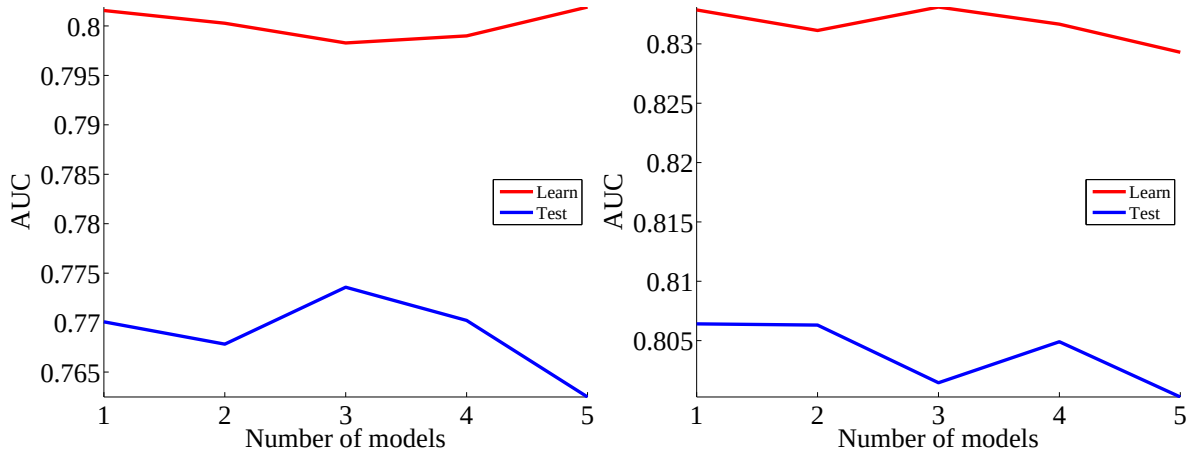
5.3 Тестирование мультимодельного подхода на реальных данных

Применим многоуровневые модели и смеси логистических моделей на данных по немецким потребительским кредитам и данных по сердечным заболеваниям.

Рассмотрим результаты применения многоуровневых моделей к данным по сердечным заболеваниям. Если не производить нормировку векторов весов моделей $\mathbf{w}_k$, $k = 1, \ldots, K$, то среди моделей выделяется одна, к которой относятся все объекты и контроля, и обучения. Это связано с тем, что объект при осторожном выборе модели относится к той модели, разделяющая гиперплоскость которой ближе с точностью до нормы вектора весов. Таким образом, если норма $\|\mathbf{w}_k\| \gg \|\mathbf{w}_j\| \, \forall \, j \neq k$, то все объекты будут отнесены к модели с номером k. Значительное же различие в нормах весов моделей связано с плохой обусловленностью матрицы $\mathbf{H}$. Поэтому будем нормировать вектора весов на одну и ту же постоянную C в соответствии с (45).

$$\mathbf{w}_k \mapsto \frac{C\mathbf{w}_k}{\max(C, \sqrt{\mathbf{w}_k^\intercal \mathbf{w}_k})}. \tag{45}$$

Зависимость площади под ROC-кривой от числа моделей на обучении и контроле приведена на рис. 14. Рис. 14 a) соответствует исходным данным, а рис. 14 b) – данным с удаленными выбросами. В обоих случаях значительного улучшения на обучении при увеличении числа многоуровневых моделей не наблюдается. На исходных данных наблюдается рост AUC для трех многоуровневых моделей по сравнению с одной на 0.005. Подобный результат получаем и на данных по немецким потребительским кредитам (см. рис. 15).

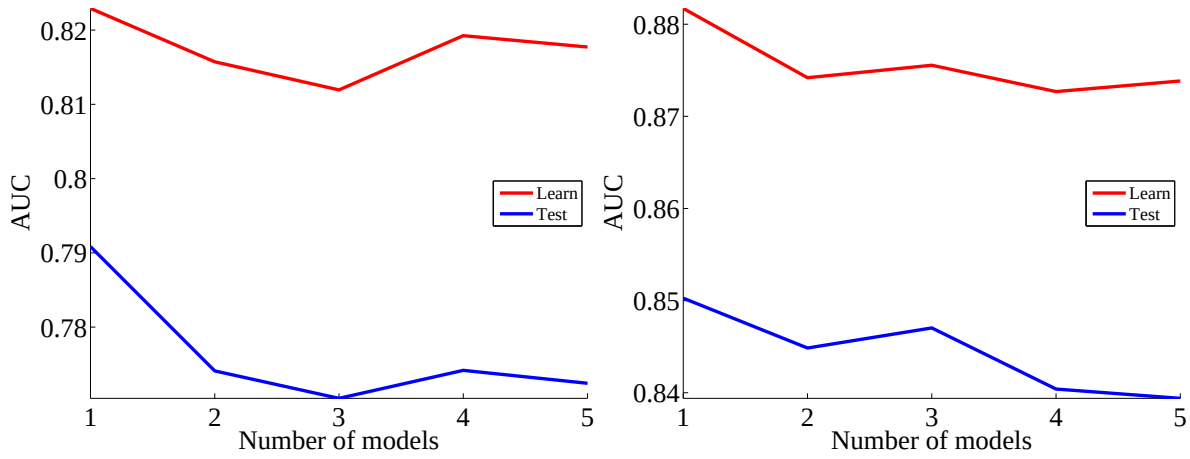


(а) Исходные данные.

(б) Удалены выбросы.

Рис. 14: Зависимость AUC от числа многоуровневых моделей для данных SAHD

Воспользуемся смесью логистических моделей для классификации тех же данных. Как и в случае с многоуровневыми моделями будем нормировать параметры моделей на константу. Будем пользоваться предложенным пошаговым алгоритмом увеличения числа моделей. Эксперимент проводился при максимальном числе моделей 10. Результат на данных по сердечным заболеваниям следующий: все построенные модели совпадают. Этот вывод был сделан при константе нормировки $C = 100$, поскольку при максимальной норме вектора весов 100, максимальное отличие в параметрах моделей имеет порядок 10^{-14}. Ясно, что значения $\mathbf{w}^\intercal \mathbf{x}$ поэтому для этих моделей почти не отличаются, так как $|x_{ij}| \leq 1 \, \forall i, j$. Однако указанный вывод можно сделать и статистически. Для этого воспользуемся локальной нормальностью оценок



(а) Исходные данные. (б) Удалены выбросы.

Рис. 15: Зависимость AUC от числа многоуровневых моделей для данных по немецким потребительским кредитам

параметров, то есть $\hat{\mathbf{w}} \sim N(\mathbf{w}_0,\, \Sigma^{-1})$, где

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \mathbf{H}_k$$

в силу вида функции правдоподобия (25), а $\mathbf{H}_k$ определено формулой (27). Тогда если $\hat{\mathbf{w}}_1$, $\hat{\mathbf{w}}_2$ есть оценки параметров одной модели, получим

$$\hat{\mathbf{w}}_1 - \hat{\mathbf{w}}_2 \sim \mathrm{N}(\mathbf{0},\, 2\boldsymbol{\Sigma}^{-1}),$$

откуда

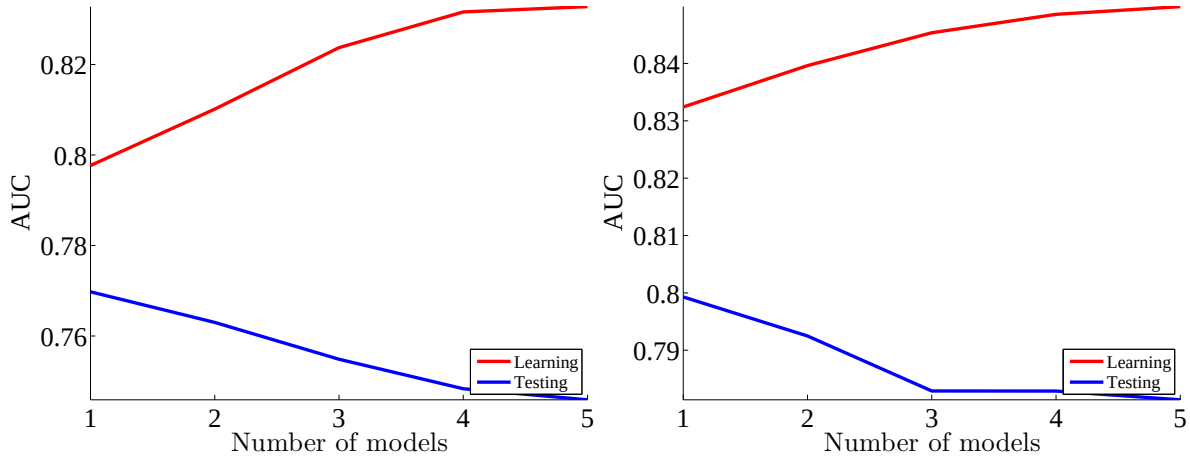
$$\frac{1}{2}\,(\hat{\mathbf{w}}_1 - \hat{\mathbf{w}}_2)^{\mathsf{T}}\,\boldsymbol{\Sigma}^{-1}\,(\hat{\mathbf{w}}_1 - \hat{\mathbf{w}}_2) \sim \chi^2(n).$$

При увеличении константы нормировки C до 1000 модели становятся существенно отличными друг от друга, но качество в терминах AUC растет только на обучении. На контроле происходит снижение качества (см. рис. 16). На рис. 16 a) приведена зависимость AUC на обучении и контроле от числа моделей для исходных данных, а на рис. 16 b) та же зависимость для данных с удаленными выбросами.

Для данных по немецким потребительским кредитам построенные модели при $C = 300$ тоже почти совпадают (хотя порядок различий в параметрах уже 10^{-3}). При дальнейшем сильном увеличении C метод IRLS перестает сходиться. Поэтому далее приведем результаты для $C = 300$. При этом значении C значительного улучшения качества не происходит, как и его ухудшения (см. рис. 17). Такие же результаты имеют место и при меньших значениях C.

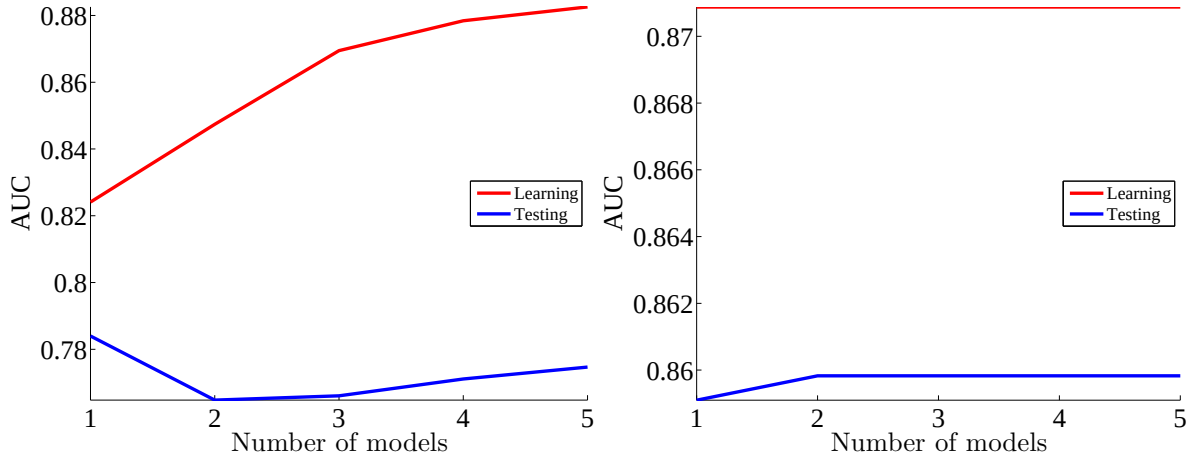
Полученные результаты можно объяснить тем, что в рассматриваемые данные не описываются двумя или более моделями. Это косвенно подтверждает вид проекции данных на две первые главные компоненты (см. рис. 18).

Синим на рис. 18 обозначены точки класса 1, а красным – класса 0.



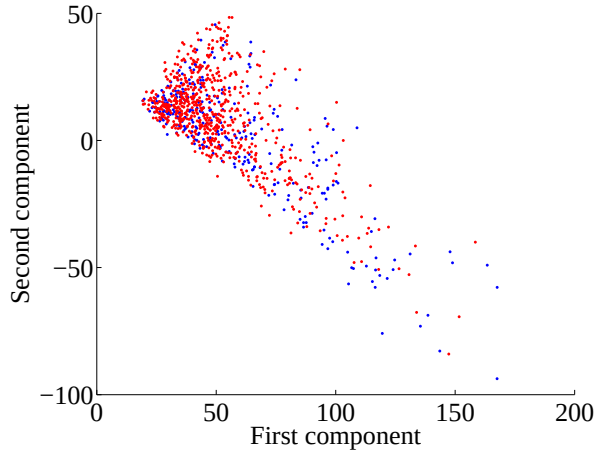
(а) Исходные данные. (б) Удалены выбросы.

Рис. 16: Зависимость AUC от числа моделей в смеси для данных SAHD

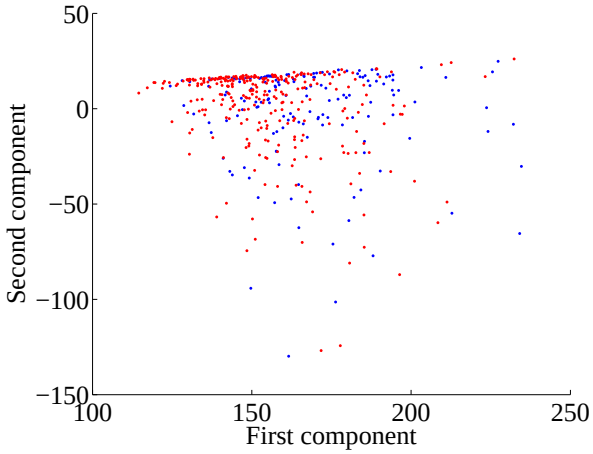


(а) Исходные данные. (б) Удалены выбросы.

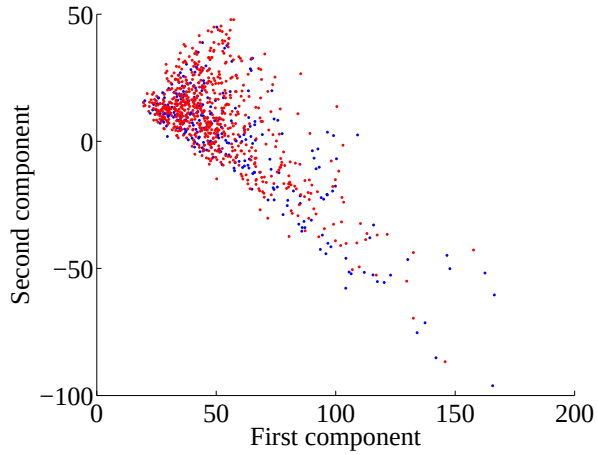
Рис. 17: Зависимость AUC от числа моделей в смеси для данных немецким потребительским кредитам.



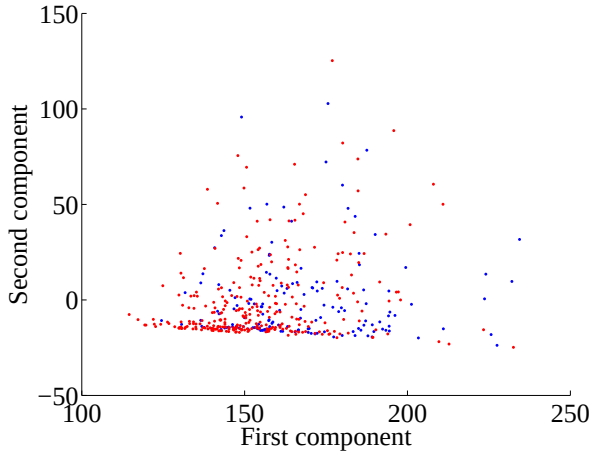
(а) Данные по немецким потребительским кредитам.



(б) Данные по сердечным заболеваниям.



(в) Данные по немецким потребительским кредитам. Отобраны выбросы.



(г) Данные по сердечным заболеваниям. Отобраны выбросы.

Рис. 18: Две первые главные компоненты данных.

5.4 Результаты отбора признаков с помощью предложенной модификации метода Белсли

Проиллюстрируем работу предложенной модификации метода Белсли на данных венгерских потребительских кредитов и данных немецких потребительских кредитов. Для выборки немецких потребительских кредитов воспользуемся предложенным в работе пошаговым алгоритмом отбора признаков, в этапе удаления которого и используется предложенная модификация метода Белсли. Зависимость отрицательного логарифма правдоподобия от номера шага алгоритма приведена на рис. 19. При этом в предложенном алгоритме в качестве параметра Z_1, определяющего момент остановки на этапе добавления признаков, было взято значение -0.5. Этап удаления признаков продолжался до тех пор, пока полученное увеличение значения $-\log L$ не превысит половину его уменьшения на предыдущем шаге.

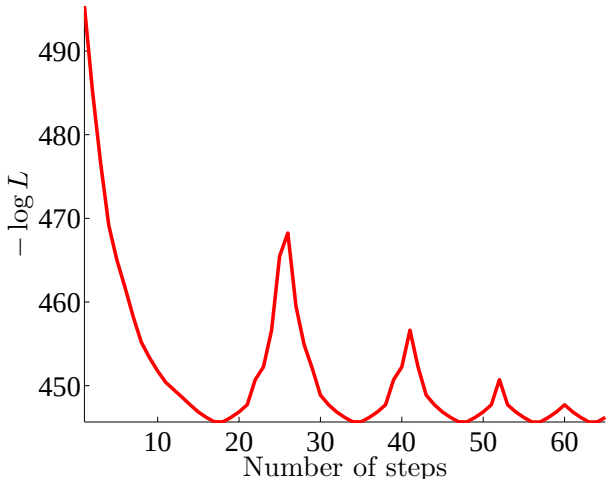
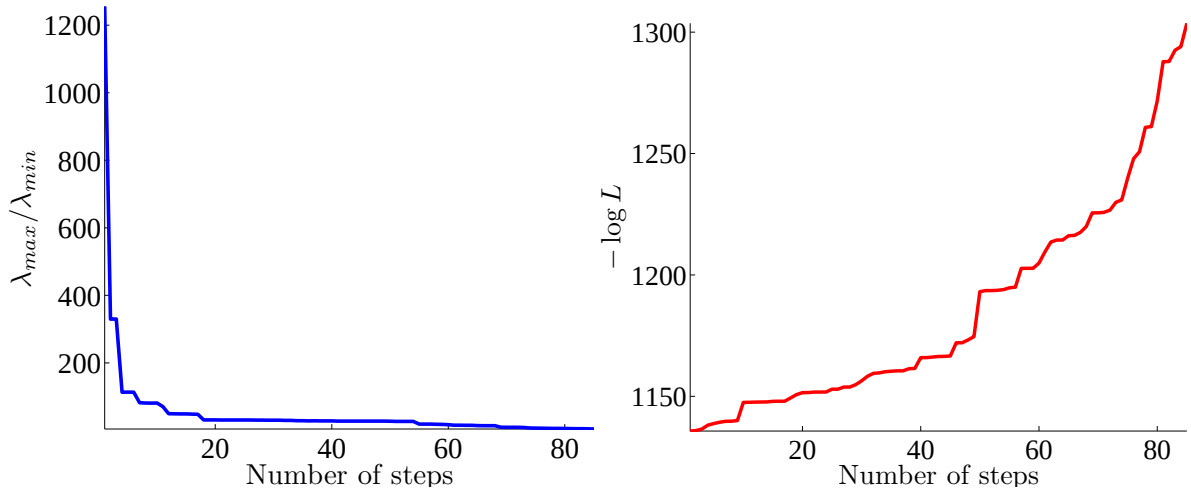


Рис. 19: Зависимость $-\log L$ от номера шага в алгоритме отбора признаков.

Из рис. 19 заключаем, что модификация метода Белсли не пригодна для применения в качества стратегии удаления признаков в алгоритме отбора признаков типа Add-Del. Покажем, однако, что предложенная модификация метода Белсли позволяет повысить устойчивость, что в терминах задачи означает снизить число обусловленности матрицы $\mathbf{H}$, обратная к которой является оценкой матрицы ковариаций параметров модели и используется в методе IRLS, сходимость в котором при плохой обусловленности этой матрицы может не наблюдаться или зависеть от начальной точки. Применим предложенную модификацию метода Белсли для удаления коррелированных признаков из данных по венгерским потребительским кредитам. На рис. 20 приведены зависимости число обусловленности матрицы $\mathbf{H}$ и значения отрицательного логарифма правдоподобия $-\log L$ от количества удаленных с помощью модификации метода Белсли признаков. При этом общее число признаков до удаления равно 90.

Из рис. 20 заключаем, что удаление 10 признаков мало изменяет правдоподобие модели, то есть слабо уменьшает качество, но число обусловленности снижается в 10 раз. Это и говорит о применимости метода Белсли для повышения устойчивости модели.



(а) Число обусловленности η матрицы $\mathbf{H}$ (б) $-\log L$

Рис. 20: Зависимость числа обусловленности η матрицы $\mathbf{H}$ и $-\log L$ от количества удаленных признаков.

5.5 Результаты на реальных данных Яндекса

Реальные данные представляют собой выборку объемом 97290 объектов, которые относятся к пяти линейно упорядоченным классам $\{0, 1, 2, 3, 4\}$. Объекты представляют собой выдачи Яндекса на поисковые запросы. Все признаки нормированы на отрезок $[0, 1]$, номера классов соответствуют релевантности полученной выдачи соответствующему запросу. Общее число признаков—245.

Кратко опишем функционал качества, используемый Яндексом для оценки качества решения задачи ранжирования документов. Рассмотрим произвольный запрос $q_j \in Q$, где Q – множество всех запросов, и соответствующие ему документы и оценки их релевантности $\Omega_j = \{\mathbf{x}_i, \hat{y}_i\}$, $i \in \mathcal{I}_j$. Здесь $\mathcal{I}_j$ задает набор индексов документов, соответствующих запросу q_j. Для каждого q_j отсортируем документы внутри Ω_j по убыванию их оценок релевантности $\mathbf{y}$, получим множество Ω_j^*. При этом документы $\mathbf{x}_i$, $\mathbf{x}_j$ с одинаковыми оценками релевантности $\hat{y}_i = \hat{y}_j$ располагаются в порядке убывания их реальных релевантностей y_i и y_j. Обозначим $\text{ind}(\mathbf{x}_i)$ –номер документа $\mathbf{x}_i$ в $\mathcal{I}$.

В качестве функционала качества будем использовать DCG (англ. Discounted Cumulative Gain), усредненный по запросам:

$$Q_2(\hat{y}) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} DCG_j, \tag{46}$$

где

$$DCG_j = \sum_{i=1}^{|\Omega_j|} \frac{y_{\text{ind}(\mathbf{x}_i)}}{\log_2 i + 1}. \tag{47}$$

Подготовка данных. Особенностью представленной выборки является малое число объектов классов 3 и 4, менее 3% объектов каждого из классов и наличие почти

постоянных признаков. Для устранения мультиколлинеарности воспользуемся методом главных компонент [30]. В данной работе были взяты 63 главных компоненты из условия $\frac{\lambda}{\lambda_{max}} > \beta = 3 \cdot 10^{-3}$, где λ_{max} — максимальное собственное число матрицы $\mathbf{X}^\intercal\mathbf{X}$, а λ — собственное число, соответствующее рассматриваемой главной компоненте. Кроме того, с учетом малого числа объектов классов 3 и 4 обучающая выборка была дополнительно сбалансирована и содержала примерно одинаковое количество объектов из каждого класса.

Сравнение логистической регрессии и SVM. Приведем значения функционала Q_2 при классификации объектов с помощью многоклассовой логистической регрессии и с помощью базового алгоритма SVM (см. табл. 8). Алгоритм логистической

Таблица 8: Сравнение логистической регрессии и базового алгоритма SVM .

Алгоритм	Значение Q_2
SVM	3.520
Логистическая регрессия	3.639

регрессии оказывается более предпочтительным в терминах функционала Q_2 (46).

Далее по отобранным наборам признаков решалась задача нахождения векторов весов $\mathbf{w}_1, \ldots, \mathbf{w}_K$ в соответствии с (33). Затем применялась предложенная модификация алгоритма логистической регрессии. Результаты в терминах Q_2 (46) приведены в табл. 9. В ней Q_2 обозначает значение Q_2 для логистической модели без предложенной модификации, а $\hat{Q}_2$ – с модификацией. Полученное значение Q_2 4.058 превосходит baseline, предложенный Яндексом [23]. Это позволяет говорить о перспективности предложенного алгоритма для ранжирования документов. Для сравнения качества с существующими алгоритмами был использован пакет SVM^{light} в режиме построения регрессии. Полученное значение функционала качества DCG равно 4.234 при обучении по всей обучающей выборке. Это на 4.5% выше, чем получено предложенным алгоритмом, потому предложенный алгоритм еще, видимо, можно улучшать.

Таблица 9: Сравнение качества Q_2 для двух алгоритмов отбора признаков.

Алгоритм отбора	Число признаков	Q_2	$\hat{Q}_2$
Пошаговый	12	3.612	4.028
Генетический	18	3.639	4.058

6 Заключение

Предложен алгоритм отбора объектов и признаков. Отбор признаков основан на предложенной модификации метода Белсли отбора признаков для логистической

регрессии и существенно повышает устойчивость построенных моделей. Отбор объектов основан на введенной функции специфичности объектов. Показано, что связанное с отбором объектов повышение качества классификации значимо. Для многоуровневых моделей предложено правило выбора модели для объектов обучения, которое снижает переобучение. Предложена модификация алгоритма многоклассовой логистической регрессии для ранжирования объектов внутри классов, которая позволила значительно увеличить качество прогноза релевантности для данных Яндекса.

7 Публикации по теме

1. Адуенко А. А. Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга // Машинное обучение и анализ данных, 2012. № 3. С. 279–291.

2. Адуенко А. А., Стрижов В. В. Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов // Инфокоммуникационные технологии, 2013. № 2.

3. Адуенко А. А., Кузьмин А. А., Стрижов В. В. Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ, 2012. № 3. С. 119–131.

4. Адуенко А. А., Стрижов В. В. Алгоритм оптимального расположения названий коллекции документов // Программная инженерия, 2013. № 3. С. 21–25.

5. Иванова А. В., Адуенко А. А., Стрижов В. В. Алгоритм построения логических правил при разметке текстов // Программная инженерия, 2013. № 6.

Список литературы

[1] Устное сообщение В. В. Стрижова. 14.06.2013.

[2] *Bishop C. M.* Pattern recognition and machine learning. // Springer, 2006.

[3] *Bishop C. M., Nasrabadi N. M.* Pattern recognition and machine learning. // Journal of electronic imaging, 2007. Vol. 16. No. 4.

[4] *Verlinde P., Cholet G.* Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application // Proc. Int. Conf. Audio and Video-Based Biometric Person Authentication (AVBPA),1999. Pp. 188–193.

[5] *Gelman A., Hill J.* Data analysis using regression and multilevel/hierarchical models // Cambridge University Press, 2006.

[6] *Oh I. S., Lee J. S., Moon B. R.* Hybrid genetic algorithms for feature selection. // IEEE transactions on pattern analysis and machine intelligence, 2004. Vol. 26. No. 11. Pp. 1424–1437.

[7] *Leardi R., Boggia R., Terrile M.* Genetic algorithms as a strategy for feature selection. // Journal of chemometrics, 1992. Vol. 6. No. 5. Pp. 267–281.

[8] *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring // Wiley, 2006.

[9] *Hosmer D. W., Lemeshow S.* Applied logistic regression // A Wiley-Interscience Publication, 2000.

[10] *Hastie T., Tibshirani R., Friedman J. H.* The Elements of Statistical Learning // Springer, 2001.

[11] *Weston J. et al.* Feature selection for SVMs // Advances in neural information processing systems, 2001. Pp. 668-674.

[12] *Chapelle O. et al.* Choosing multiple parameters for support vector machines // Machine learning, 2002. Vol. 46. No. 1. Pp. 131–159.

[13] *Khalili A.* An Overview of the New Feature Selection Methods in Finite Mixture of Regression Models // Journal of Iranian Statistical Society, 2011. Vol. 10. No. 2. Pp. 201–235.

[14] *Huang C. L., Wang C. J.* A GA-based feature selection and parameters optimization for support vector machines // Expert Systems with applications, 2006. Vol. 31. No. 2. Pp. 231–240.

[15] *Chen Y. W., Lin C. J.* Combining SVMs with various feature selection strategies // Feature Extraction, 2006. Pp. 315–324.

[16] *Neumann J., Schnörr C., Steidl G.* Combined SVM-based feature selection and classification // Machine Learning, 2005. Vol. 61. No. 1. Pp. 129–150.

[17] *Krishnapuram B. et al.* Sparse multinomial logistic regression: Fast algorithms and generalization bounds // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005. Vol. 27. No. 6. Pp. 957–968.

[18] *Lee Y., Nelder J. A., Pawitan Y.* Generalized linear models with random effects: unified analysis via H-likelihood // Chapman&Hall/CRC, 2006. Vol. 106.

[19] *Леонтьева Л. Н.* Последовательный выбор признаков при восстановлении регрессии // Машинное обучение и анализ данных, 2012. Т. 1. № 3. С. 335–346.

[20] *Motrenko A., Strijov V., Weber G. W.* Bayesian sample size estimation for logistic regression.

[21] Данные по немецким потребительским кредитам. http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/, 2000.

[22] Данные по венгерским потребительским кредитам. http://svn.code.sf.net/p/mlalgorithms/code/Aduenko2013BsThesis/data/cleared_data.csv.

[23] Данные конкурса Интернет-математика 2009. http://imat2009.yandex.ru/.

[24] Данные по сердечным заболеваниям в Южной Африке. http://svn.code.sf.net/p/mlalgorithms/code/Aduenko2013BsThesis/data/SAHD.csv

[25] *Malkovich J. F., Afifi A. A.* On tests for multivariate normality // Journal of the American Statistical Association, 1973. Vol. 68. No. 341. Pp. 176–179.

[26] *Agresti A.* An introduction to categorical data analysis // Wiley-Interscience, 2007. 423 p.

[27] *Ling C. X., Huang J., Zhang H.* AUC: a statistically consistent and more discriminating measure than accuracy // International joint Conference on artificial intelligence, 2003. Vol. 18. Pp. 519–526.

[28] *Van den Noortgate W., De Boeck P., Meulders M.* Cross-classification multilevel logistic models in psychometrics // Journal of Educational and Behavioral Statistics, 2003. Vol. 28. No. 4. Pp. 369–386.

[29] *Moerbeek M., Van Breukelen G. J. P., Berger M. P. F.* Optimal experimental designs for multilevel logistic models // Journal of the Royal Statistical Society: Series D (The Statistician), 2001. Vol. 50. No. 1. Pp. 17–30.

[30] *Jolliffe I. T.* Principle Component Analysis. // New York: Springer, 2002.