Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга*

А. А. Адуенко
aduenko1@gmail.com
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Работа посвящена выбору оптимального набора признаков для определения уровня надежности заемщиков, подавших заявку на банковский кредит. Для ответа на поставленный вопрос оценивается вероятность невозврата кредита. Для отбора признаков используется шаговая регрессия, исследуется зависимость информативности отобранных признаков от параметров шаговой регрессии. В вычислительном эксперименте алгоритм тестируется на данных потребителей, подававших заявки на кредиты в определенный банк, а также на данных об отклике клиентов на маркетинговую кампанию банка.

Ключевые слова: *банковский кредит, логистическая регрессия, выбор признаков, функция эмпирического риска, вероятность невозврата.*

Feature selection and stepwise logistic regression for credit scoring*

A. A. Aduenko
Moscow Institute of Physics and Technology

The article is dedicated to the selection of the optimal set of features for determining the quality of bank loans' requests. The default probability is estimated to answer this question. The stepwise regression is used for the feature selection. The dependency of an informativity of the selected features on the stepwise regression parameters is studied. In the computational experiment the algorithm described in the paper is tested on the data of consumers who applied for loans in a certain bank and also on data about clients' response to bank's marketing campaign.

Keywords: *bank loan, logistic regression, feature selection, empirical risk function, default probability.*

Введение

В работе рассматривается задача кредитного скоринга [1]. По данным заемщиком ответам на фиксированный набор вопросов анкеты требуется определить, в состоянии ли тот вернуть кредит банку. Кроме того, одной из основных задач является выделение некоторого небольшого набора признаков, по которому наиболее точно можно будет судить о кредитоспособности. Основным источником алгоритмов и способов отбора признаков служили [1, 2]. Модифицированные версии алгоритмов, представленных там и применяются в работе. Они основаны на подсчёте WOE (англ. weight of evidence), меры информативности соответствующего значения признака, для каждого из признаков в отдельности. Для поиска весов признаков из найденного оптимального набора используется логистическая регрессия [1, 2, 3, 4], а для их отбора— шаговая логистическая регрессия [1, 2]. Для контроля качества на тестовой выборке рассчитывается функция эмпирического риска [5, 6]. В вычислительном эксперименте представлены результаты работы построенного алгоритма

Научный руководитель В. В. Стрижов

на данных об отклике клиентов на маркетинговую кампанию банка [7]. Также рассмотрены свойства алгоритма при работе с данными анкет по потребительским кредитам [8].

Постановка задачи

Имеются исходные данные – выборка $D = \{(x_i, y_i)\}$, $i \in \mathcal{I} = \mathcal{S} \sqcup \mathcal{T}$: матрица признаков $X \in \mathbb{R}^{m \times n}$ (m–число записей данных, а n–количество признаков) и вектор ответов $\mathbf{y}$, $y_i \in \{-1, 1\}$. Здесь -1 означает, что заемщик кредит вернул (класс Y_{-1}), а 1— не вернул (класс Y_1). Разбиение на обучающую выборку $S\{(x_i, y_i)\}$, $i \in \mathcal{S}$ и тестовую $T\{(x_i, y_i)\}$, $i \in \mathcal{T}$ осуществляется случайно. Предполагается, что $x_{ij} \in \mathbb{Z}$.

Для определения уровня кредитоспособности заемщиков используется модель логистической регрессии

$$f(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp\left(-\langle \mathbf{x}, \mathbf{w} \rangle\right)}. \tag{1}$$

Здесь $\mathbf{w} \in \mathcal{W} = \mathbb{R}^n$ вектор параметров модели, а $\mathbf{x} \in \mathbb{Z}^n$ — вектор значений признаков объекта. $f(\mathbf{x}, \mathbf{w})$ задает оценочную вероятность того, что рассматриваемый объект принадлежит классу Y_{-1}.

Требуется по обучающей выборке S оценить параметр $\mathbf{w}^*$ модели (1), чтобы далее классифицировать объекты в предположении, что из исходного множества признаков $\{\boldsymbol{\chi}_j\}$, $j \in \mathcal{J} = \{1, \ldots, n\}$ отобрано некоторое подмножество $\{\boldsymbol{\chi}_j\}$, $j \in \mathcal{A}$ оптимальных согласно (3) признаков, $|\mathcal{A}| = n^* \leqslant n$. Параметр находится путем максимизации качества модели на обучающей выборке S.

В качестве меры качества используется функция эмпирического риска

$$R(\mathbf{w}, \mathcal{X}, \mathcal{A}) = \sum_{i=1}^{|\mathcal{X}|} \ln\left(1 + \exp\left(-y_i \cdot \langle \mathbf{x}_i, \mathbf{w} \rangle\right)\right), \tag{2}$$

где $\mathcal{X}$ — некоторая выборка объектов, $(\mathbf{x}_i,\, y_i) \in \mathcal{X}$, y_i задает класс объекта $\mathbf{x}_i$. $\mathcal{A}$ —набор индексов используемых признаков. Поиск оптимального набора параметров в соответствии с (2) осуществляется следующим образом:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathcal{W} = \mathbb{R}^n}{\arg\min} R(\mathbf{w}, S, \mathcal{A}) \tag{3}$$

Задачу поиска оптимального набора признаков $\{\boldsymbol{\chi}_j\}$, $j \in \mathcal{A}$ можно записать в виде

$$\mathcal{A} = \underset{\mathbf{w} \in \mathcal{W} = \mathbb{R}^n, \mathcal{A} \subseteq \mathcal{J}}{\arg\min} R(\mathbf{w}, \mathcal{X}, \mathcal{A}) \tag{4}$$

Задача нахождения оптимального набора признаков решается в работе с помощью шаговой логистической регрессии.

Нахождение весов признаков

Перейдем к нахождению весов признаков. Эта задача является одной из основных, поскольку от того, с каким весом тот или иной признак $\boldsymbol{\chi}_j$ войдет в модель, существенно зависит поведение классификатора (7).

Присоединим каждому вектору $\mathbf{x}_i$ в качестве первого элемента -1. Заменим $\mathcal{A}$ на $\mathcal{A} \cup \{1\}$, сохраняя обозначение $\mathcal{A}$. При исключении из $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]$ элементов x_{ij}, $j \notin \mathcal{A}$ сохраним для полученного вектора обозначение $\mathbf{x}_i$.

Для оценки $\mathbf{w}^*$, вектора параметров модели (1), пользуясь формулой (2), запишем производную функции эмпирического риска:

$$\frac{\partial R(\mathbf{w}, \mathcal{X}, \mathcal{A})}{\partial \mathbf{w}} = -\sum_{i=1}^{|\mathcal{X}|} \frac{\exp\left(-\mathbf{y}_i \cdot \langle \mathbf{x}_i, \mathbf{w} \rangle\right) \mathbf{y}_i}{1 + \exp\left(-\mathbf{y}_i \cdot \langle \mathbf{x}_i, \mathbf{w} \rangle\right)} \cdot \mathbf{x}_i. \tag{5}$$

Минимум эмпирического риска достигается в точке $\mathbf{w}^*$, определяемой из соотношения:

$$\frac{\partial R(\mathbf{w}^*, \mathcal{X}, \mathcal{A})}{\partial \mathbf{w}} = \mathbf{0} \tag{6}$$

Точку, удовлетворяющую (6), найдем методом градиентного спуска [9].

По полученным весам $\mathbf{w}^*$ строим классификатор:

$$\psi(\mathbf{x}_i) = \operatorname{sign} \langle \mathbf{w}^*, \mathbf{x}_i \rangle, \tag{7}$$

где $\mathbf{x}_i$ произвольный объект. Вероятности попасть в соответствующие классы определяется сигмоидной функцией:

$$P(Y_{-1}|\mathbf{x}_i) = f(\mathbf{x}_i, \mathbf{w}^*) = \frac{1}{1 + \exp(- < \mathbf{w}^*, \mathbf{x}_i >)}, \tag{8}$$

$$P(Y_1|\mathbf{x}_i) = 1 - f(\mathbf{x}_i, \mathbf{w}^*) = \frac{\exp(- < \mathbf{w}^*, \mathbf{x}_i >)}{1 + \exp(- < \mathbf{w}^*, \mathbf{x}_i >)}. \tag{9}$$

Теперь вернемся к отбору и порождению признаков.

Порождение признаков

Рассмотрим подробнее процесс порождения признаков. Порождение признаков необходимо, поскольку возможно часть исходных признаков мало информативна для линейного классификатора (7). Примером такого признака может служить возраст. Банковская статистика подтверждает, что лучше всего кредиты возвращают люди среднего возраста, в то время как молодые и пожилые чаще их не возвращают. Ясно, что никакой вес $\mathbf{w}_{age}$ возраста не позволит учесть эту особенность, поскольку классификатор линейный (7). Проведем процесс порождения признаков так, чтобы избежать подобных проблем.

Пусть $\boldsymbol{\chi}_j$ произвольный признак. Пусть его значения в порядке возрастания на обучающей выборке S есть $v_1, \ldots, v_k$, $v_i < v_j \, \forall i, j : i < j$. Определим для каждого значения рассматриваемого признака WOE (англ. weight of evidence) по формуле:

$$WOE_j(v_q) = \log \frac{[y_i = -1 \,\&\, \boldsymbol{\chi}_j(\mathbf{x}_i) = v_q] + 1}{[y_i = 1 \,\&\, \boldsymbol{\chi}_j(\mathbf{x}_i) = v_q] + 1}, \tag{10}$$

где $(\mathbf{x}_i, y_i) \in S$. [условие]—количество элементов выборки, на которых условие выполнено.

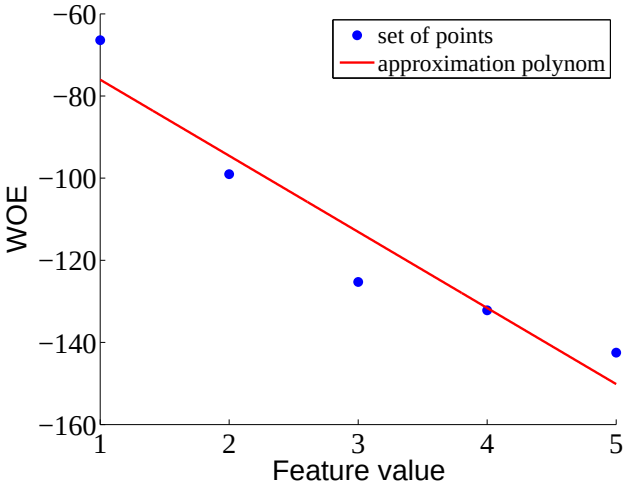
Пусть $e_q = WOE_j(v_q) \, \forall q \in \{1, \ldots, k\}$. Пусть также $\{e_1, \ldots, e_{i_1}\}$, $\{e_{i_1+1}, \ldots, e_{i_2}\}, \ldots$, $\{e_{i_{L-1}+1}, \ldots, e_{i_L} = e_k\}$ есть монотонные последовательности. Причем соседние последовательности обладают противоположным направлением роста. Назовем L числом монотонных компонент. Так как построенный классификатор линейный, то можно предположить, что наивысшее качество классификации в терминах (2) и (4) наблюдается по признакам, у которых L мало, а лучше $L = 1$.

Рассматриваемый признак $\boldsymbol{\chi}_j = [x_{1j}, \ldots, x_{Mj}]$, $M = |S|$, $x_{qj} \in \mathbb{Z}$. Разобьем числовую ось $\mathbb{R}$ на несколько полуинтервалов, а именно на L. Для этого определим $L - 1$ место разбиения $\mathbb{R}$: $d_1 < \ldots < d_{L-1}$. Положим также, что $d_0 = -\infty$, а $d_l = \infty$.

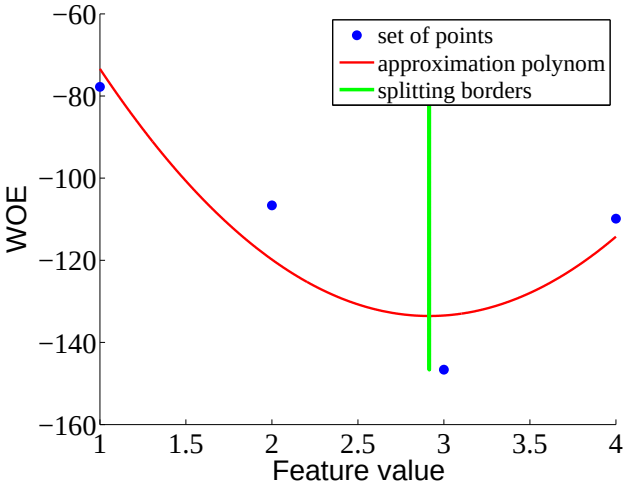
Заменяем исходный вектор признаков $\boldsymbol{\chi}_j$ на L новых: $\boldsymbol{\chi}_j^1, \ldots, \boldsymbol{\chi}_j^L$ по такому правилу: если для объекта $\mathbf{x}_q$ выборки D $\boldsymbol{\chi}_j(\mathbf{x}_q) = v_c$ и $v_c \in (d_i, d_{i+1}]$ для некоторого $i \in \{0 \ldots L-1\}$, то $\boldsymbol{\chi}_j^s(\mathbf{x}_q) = 0 \, \forall s \neq i + 1$ и $\boldsymbol{\chi}_j^{i+1}(\mathbf{x}_q) = \boldsymbol{\chi}_j(\mathbf{x}_q)$.

Теперь определим как найти $d_1, \ldots, d_{L-1}$. Для этого найдем полином степени L, наименее уклоняющийся в среднеквадратическом от точек $\{v_i, e_i\}_{i=1}^k$, то есть

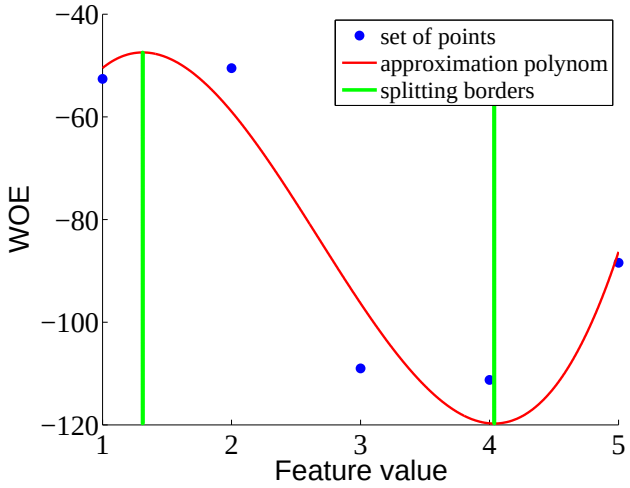
$$\mathbf{c}^* = \arg\min \|A\mathbf{c} - \mathbf{e}\|^2, \tag{11}$$



(a) Число монотонных компонент $L = 1$



(b) Число монотонных компонент $L = 2$



(c) Число монотонных компонент $L = 3$

Рис. 1. Приближение полиномом зависимости WOE от значения признаков и иллюстрация порождения признаков

где матрица A имеет следующий вид

$$\begin{pmatrix} a_{11} & \dots & a_{1L} \\ \vdots & \ddots & \vdots \\ a_{M1} & \dots & a_{ML} \end{pmatrix},$$

где $a_{kl} = (v_k)^l$. Осталось найти набор нулей B производной полученного многочлена с коэффициентами $\mathbf{c}^*$, определяемыми из (11). Именно нули производной и задают границы полуинтервалов $d_1, \dots, d_{L-1}$. На рис.1 приведены примеры разбиений действительной оси $\mathbb{R}$ для признаков с разным числом монотонных компонент L.

После выполнения процедуры для каждого признака получаем новую матрицу плана $\mathcal{X}$. Далее применяем алгоритм отбора признаков.

Отбор признаков

Зачастую то, вернет или не вернет заемщик кредит, заметно зависит не от всего набора признаков, а лишь от их части. Для удаления из множества признаков $\boldsymbol{\chi}_j$, $j \in \mathcal{J}$ таких неинформативных признаков и применим их отбор.

Для отбора признаков воспользуемся шаговой логистической регрессией. Подробно алгоритм шаговой логистической регрессии изложен в [2], в работе же опишем лишь общую его идею.

Для поиска оптимального набора признаков будем пользоваться следующим жадным алгоритмом: изначально имеем модель f, в которой ровно один признак: $\mathcal{A} = \{1\}$. Первый признак есть константа, а именно: $\boldsymbol{\chi}_1(\mathbf{x}_i) = -1\,\forall \mathbf{x}_i$.

Этот признак всегда будет в модели. Все дальнейшие шаги относятся ко всем признакам, кроме этого.

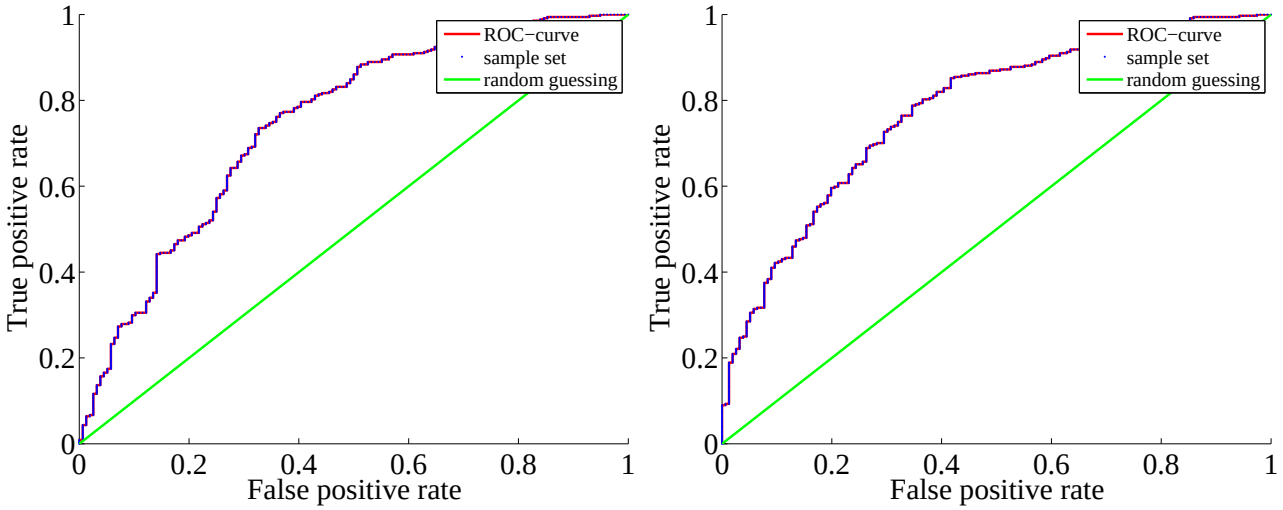
На каждом следующем шаге проверяем сначала возможность добавить новый признак в модель f, а затем возможность удалить в соответствии со следующими правилами.

Добавление признака

Пусть в модели f уже есть признаки $\boldsymbol{\chi}_{j_1}, \ldots, \boldsymbol{\chi}_{j_k}$, $j_1 = 1$, то есть $\mathcal{A} = \{j_1, \ldots, j_k\}$. Пусть также признаки $\boldsymbol{\chi}_{j_{k+1}}, \ldots, \boldsymbol{\chi}_{j_n}$ не находятся в модели f. Обозначим $\tilde{\mathbf{w}}(\mathcal{A})$ значение вектора весов признаков $\{\boldsymbol{\chi}_j\}$, $j \in \mathcal{A}$, определяемое из (6).

Эмпирический риск для модели f в соответствии с (2) обозначим $R_0 = R(\tilde{\mathbf{w}}(\mathcal{A}), S, \mathcal{A})$. Для каждого $s \in \{k+1, \ldots, n\}$ обозначим $\mathcal{A}' = \mathcal{A} \cup \{j_s\}$. Пусть эмпирический риск получаемой модели f' для каждого $s \in \{k+1, \ldots, n\}$ есть R_s. Среди всех R_s выбираем наименьший:

$$s^* = \underset{s \in \{k+1, \ldots, n\}}{\arg\min} R_s$$



(a) Исходные признаки (b) Разделенные признаки

Рис. 2. ROC-кривые для исходных и разделенных признаков

Это же в терминах (2) и введенного $\tilde{\mathbf{w}}(\mathcal{A})$ выглядит следующим образом:

$$j_{s^*} = \underset{j_s \in \mathcal{J} \setminus \mathcal{A}}{\arg\min} R(\tilde{\mathbf{w}}(\mathcal{A}'), S, \mathcal{A}'). \tag{12}$$

Обозначим $j_{s^*} = j^*$. Далее считаем рисковую разницу G_{j^*} и вероятность p, отражающую значимость признака

$$G_{j^*} = 2 \cdot (R_0 - R_{s^*}), \tag{13}$$

$$p = Pr\left[\chi^2(\nu) > G_{j^*}\right]. \tag{14}$$

Здесь $\chi^2(n)$ имеет функцию плотности вероятности

$$f_{\chi^2}(x) = \begin{cases} 0, & \text{если } x \leqslant 0; \\ \dfrac{x^{\frac{n}{2}-1} \cdot \exp\left(-\frac{x}{2}\right)}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)}, & \text{если } x > 0. \end{cases}$$

Для дискретного признака в качестве ν используем число разных значений признака. Если признак непрерывный, полагаем $\nu = 1$. При этом чем больше значение p для признака, тем менее признак значимый. Поэтому установим границу отсечения P_E. Если для признака $\boldsymbol{\chi}_{j^*}$ $p < P_E$, этот признак $\boldsymbol{\chi}_{j^*}$ добавляется в модель f, то есть $\mathcal{A} \to \mathcal{A} \cup \{j^*\}$, иначе модель f остается неизменной, то есть $\mathcal{A}$ не изменяется.

Удаление признака

После выполнения первой части шага проверяем в полученной модели все признаки на значимость. Пусть в модели f уже есть признаки $\boldsymbol{\chi}_{j_1}, \ldots, \boldsymbol{\chi}_{j_{k'}}$, $j_1 = 1$, то есть $\mathcal{A} = \{j_1, \ldots, j_{k'}\}$, где $k' = k$, если на шаге добавления признаки не добавлялись и $k' = k + 1$ иначе. Пусть также признаки $\boldsymbol{\chi}_{j_{k'+1}}, \ldots, \boldsymbol{\chi}_{j_n}$ не находятся в модели f.

Эмпирический риск для модели f в соответствии с (2) обозначим $R_0 = R(\tilde{\mathbf{w}}(\mathcal{A}), S, \mathcal{A})$. Для каждого $s \in \{1, \ldots, k'\}$ обозначим $\mathcal{A}' = \mathcal{A} \backslash \{j_s\}$. Пусть эмпирический риск получаемой модели f' для каждого $s \in \{1, \ldots, k'\}$ есть R_s. Среди всех R_s выбираем наименьший:

$$s^* = \underset{s \in \{1, \ldots, k'\}}{\arg\min} R_s$$

Это же в терминах (2) и введенного $\tilde{\mathbf{w}}(\mathcal{A})$ выглядит следующим образом:

$$j_{s^*} = \underset{j_s \in \mathcal{A}}{\arg\min} R(\tilde{\mathbf{w}}(\mathcal{A}'), S, \mathcal{A}'). \tag{15}$$

Обозначим $j_{s^*} = j^*$. Далее считаем рисковую разницу G_{j^*} и вероятность p, отражающую значимость признака

$$G_{j^*} = 2 \cdot (R_0 - R_{s^*}), \tag{16}$$

$$p = Pr\left[\chi^2(\nu) > G_{j^*}\right]. \tag{17}$$

Руководствуясь значением вероятности p как оценкой значимости признака, устанавливаем границу отсечения $P_R > P_E$. Если окажется, что $p > P_E$, то признак $\boldsymbol{\chi}_{j^*}$ удаляется из модели, то есть $\mathcal{A} \to \mathcal{A} \backslash \{j^*\}$. Иначе модель не изменяется, то есть $\mathcal{A}$ остается прежним.

Переход на следующий шаг происходит, если было совершено или удаление, или добавление, иначе алгоритм заканчивает свою работу. По окончании работы алгоритма получаем некоторый набор признаков, по которому можно классифицировать объекты.

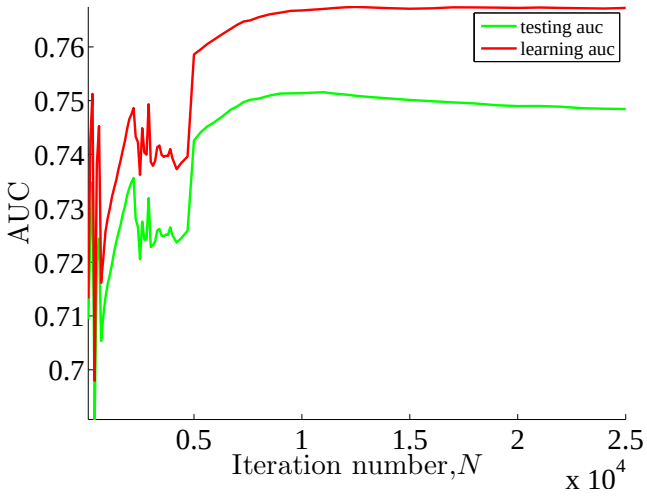
Вычислительный эксперимент

В вычислительном эксперименте продемонстрируем работу приведенных алгоритмов на следующих данных:

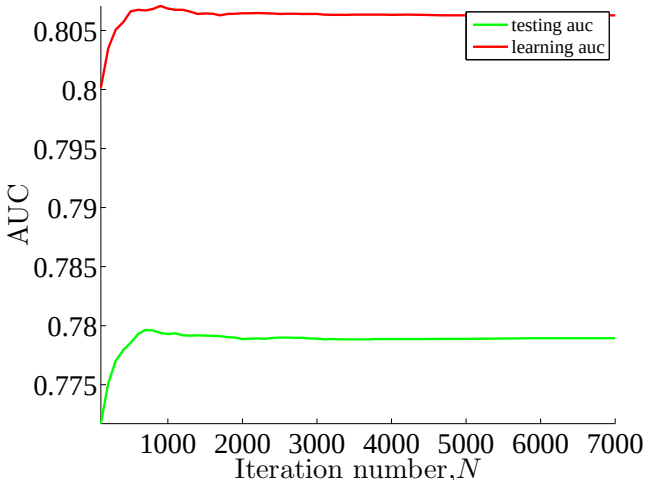
— Данные анкет по потребительским кредитам [8] (1000 объектов, 24 признака)

— Данные отклика клиентов на маркетинговую кампанию OTP-банка [7] (15223 объекта, 36 признаков)

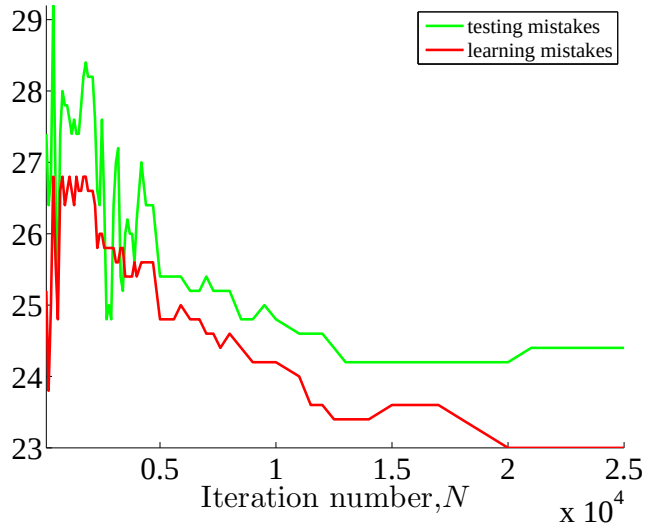
В данных по потребительским кредитам пропусков и данных в текстовом виде не было, поэтому эти данные использовались без обработки. В данных же OTP-банка были пропуски, а также признаки в текстовом виде. Пропуски были заполнены нулями, а признаки в текстовом виде не учитывались.



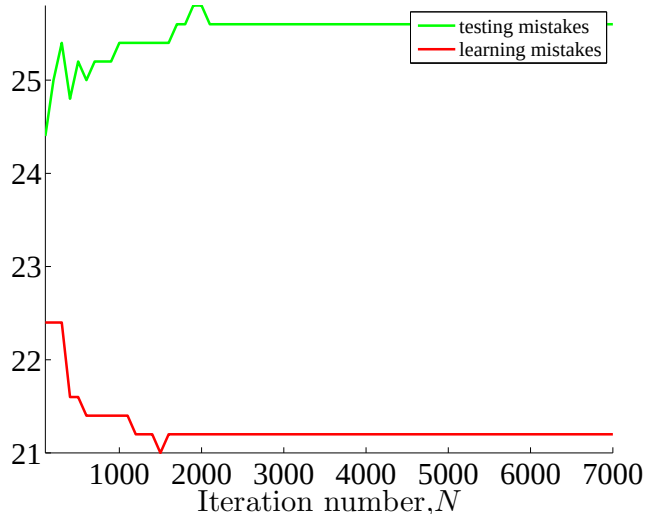
(a) Исходные признаки



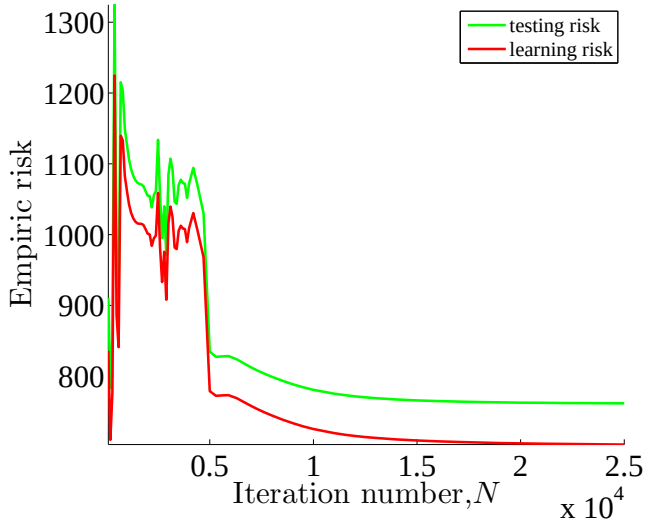
(b) Разделенные признаки



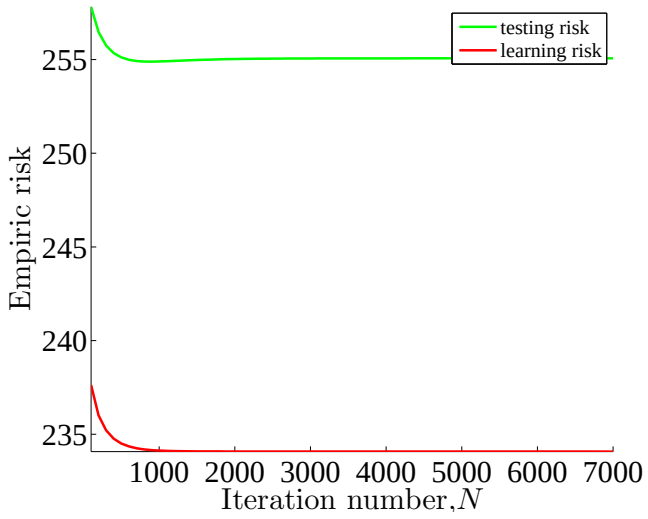
(c) Исходные признаки



(d) Разделенные признаки

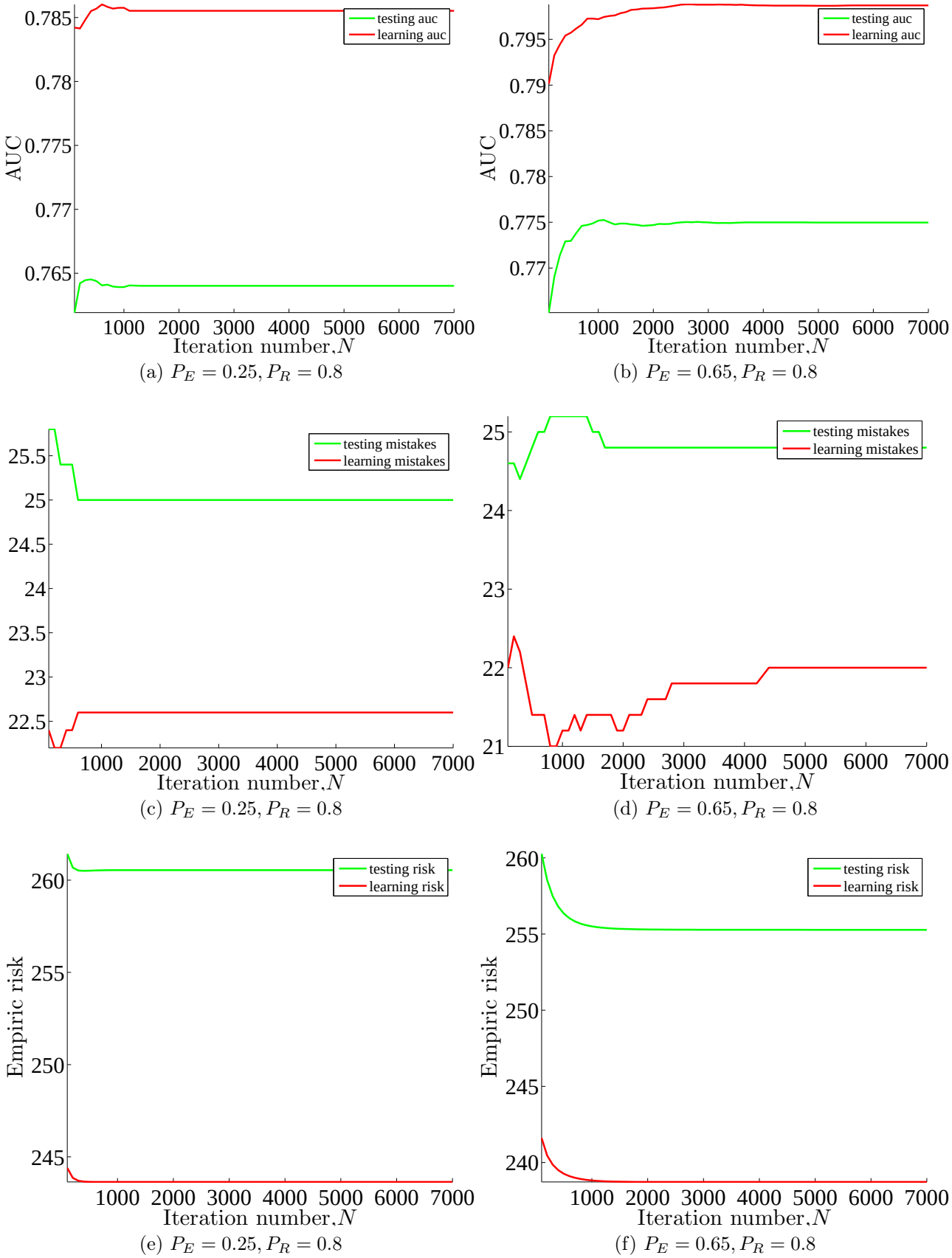


(e) Исходные признаки



(f) Разделенные признаки

Рис. 3. Зависимость площади AUC под ROC-кривой, процента ошибок и эмпирического риска для исходных и разделенных признаков от числа итераций метода градиентного спуска

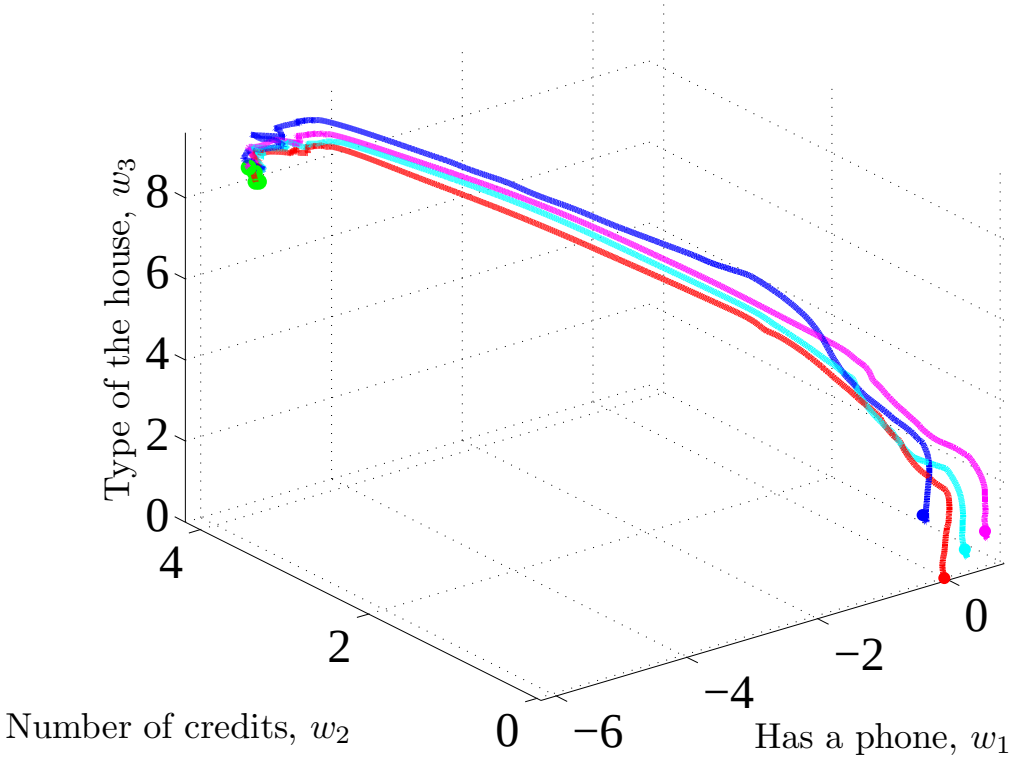


(a) $P_E = 0.25, P_R = 0.8$ (b) $P_E = 0.65, P_R = 0.8$

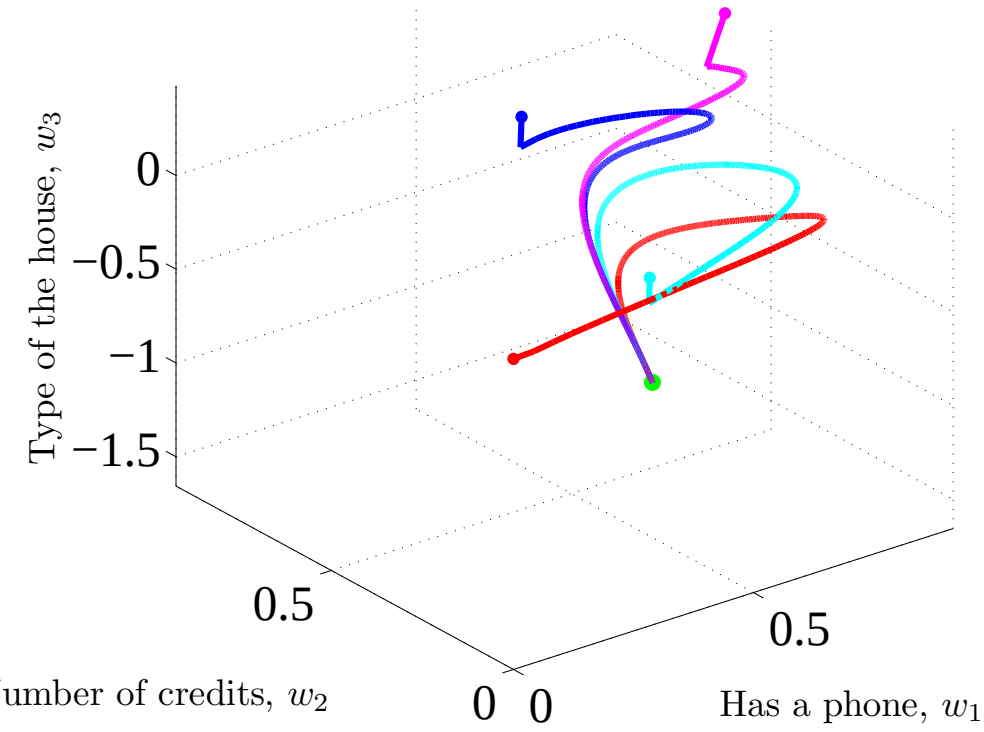
(c) $P_E = 0.25, P_R = 0.8$ (d) $P_E = 0.65, P_R = 0.8$

(e) $P_E = 0.25, P_R = 0.8$ (f) $P_E = 0.65, P_R = 0.8$

Рис. 4. Зависимость площади под *ROC*—кривой, процента ошибок и эмпирического риска от числа итераций градиентного спуска для $P_E = 0.25$, $P_R = 0.8$ и $P_E = 0.65$, $P_R = 0.8$



(a) Исходные признаки

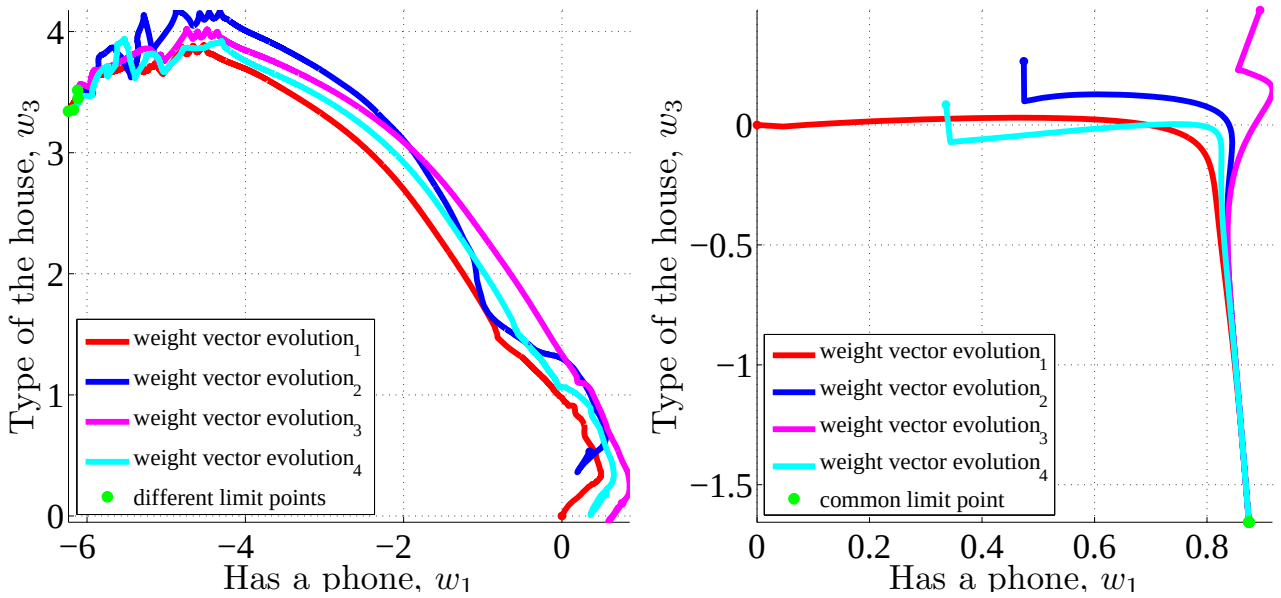


(b) Отобранные признаки, $P_E = 0.65$, $P_R = 0.8$

Рис. 5. Сходимость к оптимальному вектору весов для исходных и отобранных признаков из двух разных начальных приближений в пространстве трех признаков

Целью вычислительного эксперимента было проверить предположения о том, что построенные алгоритмы порождения и отбора признаков позволяют повысить качество классификации в терминах (2), выделить значимые признаки (4) и тем самым сократить объем обрабатываемой информации. Также требовалось оценить как применение алгоритмов сказывается на требуемых для обработки вычислительных ресурсах.

Иллюстрация порождения признаков



(a) Исходные признаки

(b) Отобранные признаки, $P_E = 0.65$, $P_R = 0.8$

Рис. 6. Сходимость к оптимальному вектору весов для исходных и отобранных признаков из двух разных начальных приближений в плоскости двух признаков

Проиллюстрируем описанное выше на реальных данных о потребительских кредитах [8]. Начнем с примеров признаков, обладающих разным числом монотонных компонент L, выделяемых в отдельные признаки. Приведем иллюстрации приближения полиномом (11) с коэффициентами $\mathbf{c}^*$ зависимости $WOE_j(v_q)$ для трех разных признаков. На графиках на рис. 1 также изобразим найденные границы разбиения действительной оси $\mathbb{R}$ $d_1, \ldots, d_{L-1}$ как нулей производной полученного полинома с коээфициентами $\mathbf{c}^*$. При этом на рис. 1(а) границы разбиения не показаны, так как для этого признака $L = 1$. Точки на графиках соответствуют всем представленным в обучающей выборке S значениям рассматриваемого признака $\boldsymbol{\chi}_j$ $v_1, \ldots, v_K$ и посчитанным для них по обучающей выборке S в соответствии с (10) значениям $WOE_j(v_q)$, $q \in \{1, \ldots, K\}$. Линией на графиках рис. 1 показан полином степени L с коэффициентами $\mathbf{c}^*$, наименее уклоняющийся в среднеквадратическом от этих точек.

Теперь построим логистическую регрессию на исходных признаках. Приводим ROC-кривые (на рис. 2) [5], график площади под ними AUC, график процента ошибок и график эмпирического риска (рис. 3) в зависимости от числа итераций алгоритма градиентного спуска, то есть фактически от того, насколько точно найден оптимальный для обучающей выборки T вектор весов признаков $\mathbf{w}^*$. Прямой линией на рис. 2 показан худший классификатор, основанный на случайном угадывании. На рис. 2 и рис. 3 слева приведены графики для исходных признаков, а справа – для разделенных.

Графики на рис. 2 и рис. 3 иллюстрируют тот факт, что в случае разделенных признаков наступает заметно более ранняя сходимость. Возможно, это объясняется оптимальной структурой построенных признаков для линейного классификатора. Более того, экспери-

мент показывает, что вектор весов $\mathbf{w}$ в случае с неразделенными признаками очень быстро растет по норме, а при регуляризации эмпирического риска падает эффективность. Напротив, $\|\mathbf{w}\|$ слабо растет на разделенных признаках.

Отбор признаков и сравнение результатов

В качестве данных будем использовать данные клиентов, подававших заявки на потребительские кредиты, а также данные об отклике людей на маркетинговую компанию банка.

Так как людей, подающих заявки на кредиты в банк обычно заметно больше, чем компаний, особенно важно выбрать некоторый небольшой набор признаков для идентификации надежного заемщика, чтобы работать с меньшими массивами данных.

Именно для выделения наиболее информативных признаков и будет использоваться описанный алгоритм шаговой логистической регрессии.

Для реализации алгоритма требуется задать границ отсечения для шагов добавления и удаления признаков P_E и P_R. В зависимости от выбора P_E и P_R алгоритм будет выделять в общем случае разные признаки и разное количество таковых. На опыте оказалось, что рекомендованное в [2] значение $P_E = 0.15 - 0.25$ для рассматриваемых задач слишком мало, а также, что от P_R почти ничего не зависит.

При указанном значении $P_E = 0.25$ алгоритм отбирает очень узкий набор признаков, которого не вполне хватает. Приведём графики зависимости площади под ROC-кривой, эмпирического риска и процента ошибок от числа итераций алгоритма градиентного спуска для следующих значений P_E и P_R: 0.25 и 0.8 (рекомендация [2]), 0.5 и 0.8 (для маркетинговой компании), 0.65 и 0.8 (для потребительских кредитов). Для потребительских кредитов графики приведены на рис. 4, для маркетинговой кампании — на рис. 7. На этих рисунках (рис.4 и рис. 7) слева приведены графики для $P_E = 0.25$, $P_R = 0.8$, а справа — для $P_E = 0.65$, $P_R = 0.8$ и $P_E = 0.5$, $P_R = 0.8$ соответственно. Две последних пары значений оптимальны для соответствующих задач как показывает эксперимент.

Для данных по потребительским кредитам при $P_E = 0.25$ и $P_R = 0.8$ было отобрано 9 признаков из 24, при $P_E = 0.65$ и $P_R = 0.8$ – 16 признаков. Хотя формально качество классификации после отбора возросло несильно (AUC возросло лишь на 1%), в действительности после отбора алгоритм требует не только меньше входных данных, но и меньше вычислительного времени, что демонстрирует следующая серия графиков. На них показана работа алгоритма на протяжении 5000 итераций из двух начальных приближений к оптимальному вектору весов. В случае с отобранными признаками (при $P_E = 0.65$, $P_R = 0.8$) сходимость заметно более быстрая и монотонная.

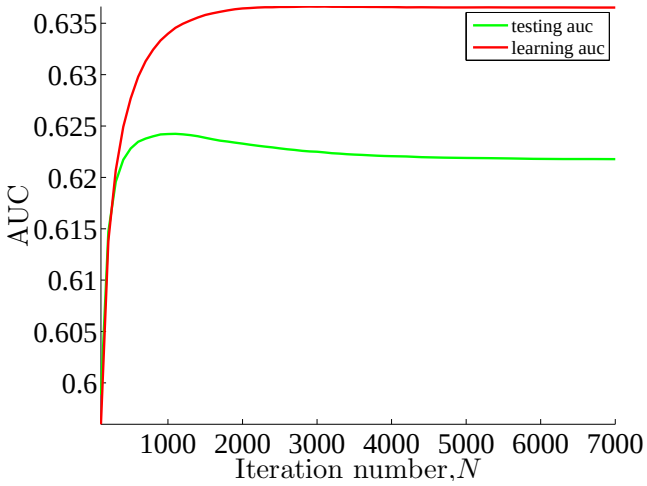
На рис. 5 и рис. 6 первым приводится график для исходных признаков, затем для отобранных с помощью шаговой регрессии при $P_E = 0.65$, $P_R = 0.8$.

Рис. 5 и рис. 6 демонстрируют, что после 5000 итераций градиентного спуска в случае исходных признаков полученные $\mathbf{w}$ из разных начальных приближений ещё значительно отличаются, а также, что сходимость немонотонна. Напротив, в случае отобранных признаков сходимость монотонная и после 5000 итераций заметных отличий траекторий эволюции $\mathbf{w}$ не наблюдается.

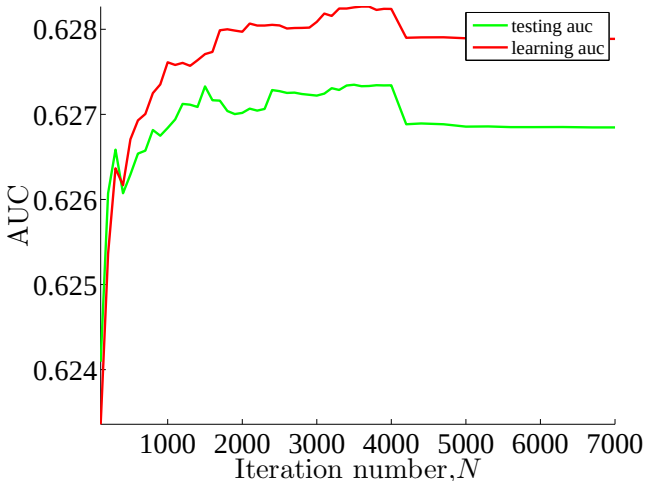
Маркетинговая кампания

Применим алгоритм отбора признаков к данным об отклике клиентов на маркетинговую кампанию банка [7]. Сравним результат работы алгоритма отбора на исходных признаках и признаках, полученных после работы алгоритма порождения признаков.

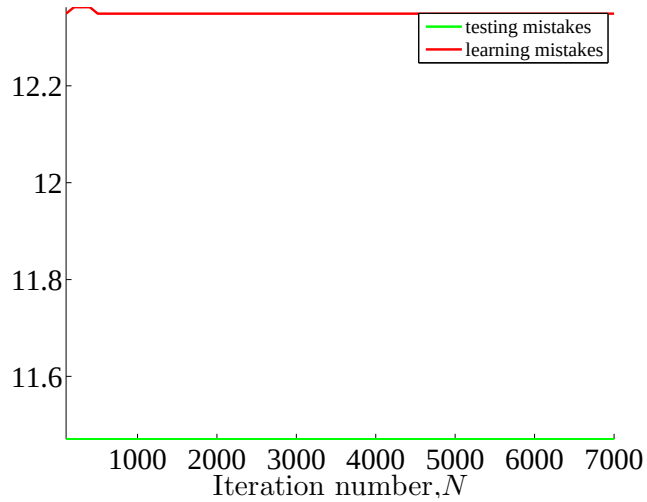
Априори можно предположить, что на этих данных качество классификации, выраженное через площадь под ROC-кривой будет ниже, чем для данных о потребительских кредитах, поскольку то, примет ли человек участие в промо-акции банка зависит во многом не от его дохода, места работы и пр., а от того, пожелает ли он того в конкретный момент времени. Однако полученные результаты говорят о практической применимости алгоритма.



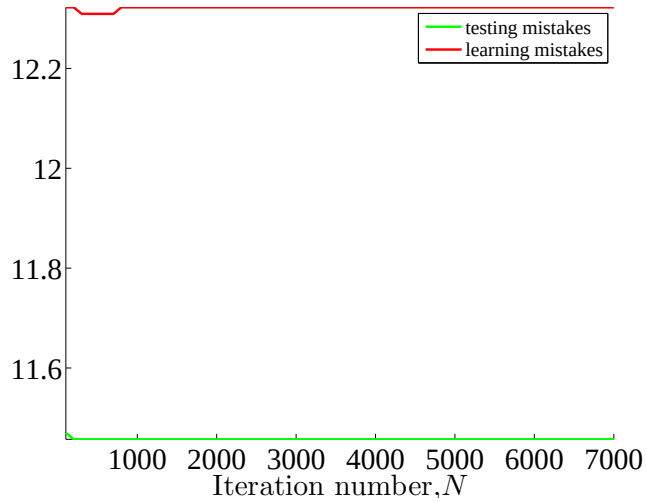
(a) Исходные признаки, $P_E = 0.5$, $P_R = 0.8$



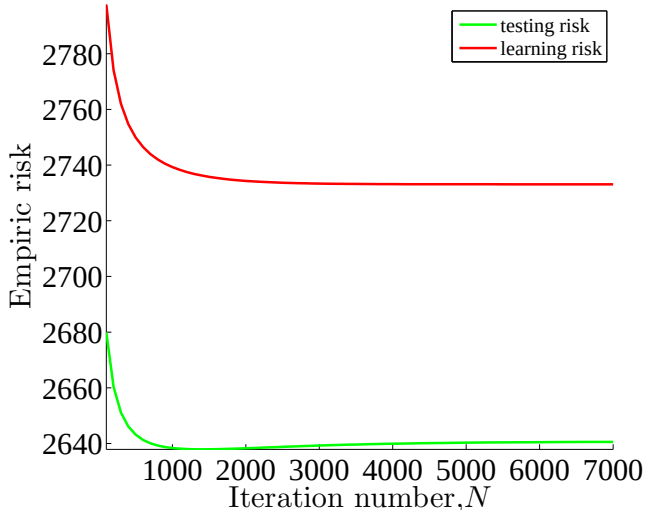
(b) Разделенные признаки, $P_E = 0.5$, $P_R = 0.8$



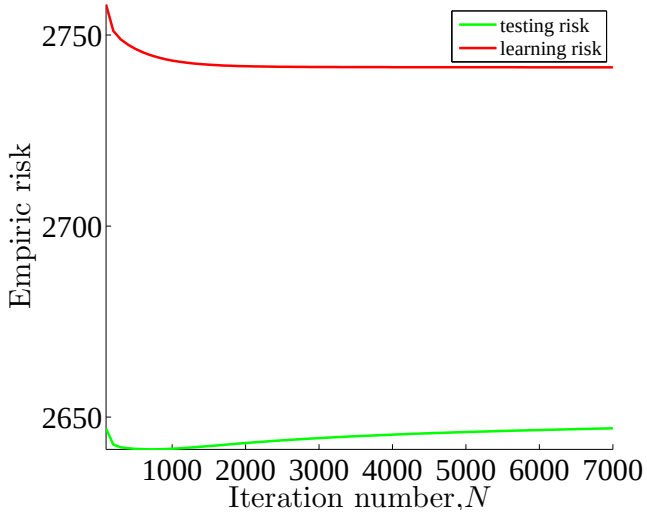
(c) Исходные признаки, $P_E = 0.5$, $P_R = 0.8$



(d) Разделенные признаки, $P_E = 0.5$, $P_R = 0.8$



(e) Исходные признаки, $P_E = 0.5$, $P_R = 0.8$



(f) Разделенные признаки, $P_E = 0.5$, $P_R = 0.8$

Рис. 7. Зависимость площади под *ROC*—кривой, процента ошибок и эмпирического риска от числа итераций градиентного спуска для $P_E = 0.5$, $P_R = 0.8$

Приведем результаты его работы для исходных и разделенных признаков(рис. 7). В последнем случае число отобранных признаков мало отличается от отобранных по исходным, однако качество классификации несколько выше.

Переобучение в обоих случаях наступает примерно после 1000 итераций, когда эмпирический риск $R(\mathbf{w}, T, \mathcal{A})$ для тестовой выборки T начинает расти при продолжающемся снижении эмпирического риска $R(\mathbf{w}, S, \mathcal{A})$ для обучающей выборки.

Заключение

В данной работе рассматривалась задача порождения признаков и выбора оптимального их набора, а также определения весов признаков с целью оценки качества заемщиков. Результаты вычислительного эксперимента показали, что после порождения признаков по описанному в работе алгоритму заметно возрастает по сравнению с исходными признаками скорость сходимости, а также заметно слабее растет норма вектора весов $\mathbf{w}$.

Также была исследована зависимость величины эмпирического риска от параметров шаговой регрессии при отборе признаков. Оказалось, что в рассмотренных примерах рекомендованные в [2] значения параметров не являются оптимальными. Отбор признаков еще более ускоряет сходимость, а она приобретает более монотонный характер.

Литература

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

[2] D.W. Hosmer and Stanley Lemeshow. *Applied logistic regression.* A Wiley-Interscience Publication, 2000.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[4] C.M. Bishop, N.M. Nasrabadi. Pattern recognition and machine learning. *J. Electronic Imaging*, 16(4):049901, 2007.

[5] К.В. Воронцов. *Линейные методы классификации.* MachineLearning.Ru, февраль 2010.

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

[7] Данные об отклике клиентов otp-банка на маркетинговую кампанию. http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/, 2011.

[8] Данные о немецких потребительских кредитах. http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/, 2000.

[9] А.Г. Сухарев, А.Г. Тимохов, В.В. Федоров. *Курс методов оптимизации.* Физматлит, 2005.