Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Митяшов Андрей Андреевич

Предметно-экспертные ограничения для штрафной функции elastic-net в случае логистической регрессии

010656 - Математические и информационные технологии

Выпускная квалификационная работа бакалавра

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

Москва
2014

Содержание

1 Введение **4**

2 Постановка задачи **8**

3 Предметно-экспертные ограничения **10**

4 Базовый алгоритм **11**

5 Метод внешних штрафных функций **13**

6 Модификация базового алгоритма **14**

7 Вычислительный эксперимент **16**

8 Заключение **16**

Аннотация

Работа посвящена созданию метода выбора оптимальной скоринговой модели. Для оценки вероятности невозврата кредита используется логистическая регрессия. С целью учета экспретного мнения вводятся предметно-экспертные ограничения на параметры модели. Также предлагается отбирать признаки с помощью штрафа elastic-net.
Ключевые слова: *логистическая регрессия, elastic-net, предметно-экспертные ограничения, банковский скоринг, отбор признаков.*

1 Введение

Актуальность темы. Задача банковского кредитного скоринга актуальна в связи с тем, что банки выдают большое количество потребительских кредитов. Так, на 2013 год в России выдано кредитов более чем на 8,8 триллионов рублей. Решение о выдаче кредита заемщика принимается либо экспертом-аналитиком, либо скоринговой системой: автоматизированной системой, оценивающей вероятность невозврата кредита по предоставленным заемщиком данным).
С ростом количества потребительских кредитов экспертные оценки теряют актуальность, поскольку ограниченное число аналитиков не может обработать все поступающие заявки. Один кредитный инспектор обрабатывает, как правило, не более 12-15 заявок в день. Это приводит к тому, что большинство банков использует скоринговые системы. Тем не менее, мнение эксперта необходимо учесть при построении скоринговой модели. Назовем скоринговой моделью отображение из признакового пространства в отрезок [0; 1], значения из этого отрезка - оценка вероятности невозврата кредита. Значения признаков формируются на основе предоставленных заемщиков данных.
Для учета экспертного мнения предлагается ввести *предметно-экспертные ограничения* на множестве допустимых значений весов признаков. Эти ограничения будут отражать мнение эксперта-аналитика о том, какие именно должны быть получены веса признаков в построенной модели. Другой особенностью данных ограничений является то, что результаты, получаемые при выборе оптимальной модели, будут заведомо интерпертируемы экспертом. Более подробно примеры таких ограничений и их возможные применения будут описаны в данной работе в разделе "Практическое применение".
Для создания скоринговой модели необходимо составить список признаков-данных, которые нужно получить от заемщика. Так, в большинство скоринговых картах входят следующие поля:

- проживание;
- срок проживания в регионе;
- стаж работы на последнем месте;
- возраст;
- заработная плата;
- семейное положение;
- образование и т.д.

Не все полученные от заемщика данные могут быть достоверны, кроме того, наличие избыточных признаков может привести к переобучению модели. Предлагается отбирать некоторое множество информативных признаков – данных, которые проверяются на истинность и которые коррелируют с вероятностью невозврата кредита. Для этого будем использовать штрафную функцию elastic-net.

Цель работы. Построить алгоритм выбора оптимальной банковской скоринговой модели, отбирающий признаки и учитывающий экспертное мнение (выраженное в виде предметно-экспертных ограничений); решить задачу оценки вероятности невозврата кредита с учетом предметно-экспертных ограничений.

Методы исследований. При построении алгоритма использовались принцип наибольшего правдоподобия, необходимое и достаточное условие глобального экстремума для выпуклой функции и метод внешних штрафных функций.

Научная новизна.

- Решена задача оценки вероятности невозврата кредита с учетом предметно-экспертных ограничений.
- Разработана и обоснована модификация алгоритма, предложенного в [1].

Практическая ценность. Разработан программный модуль, который находит оптимальные структурные параметры модели, находит веса признаков модели, отбирает признаки, учитывает введенные ограничения, визуализирует результаты.

Положения, выносимые на защиту.

- Введение предметно-экспертных ограничений.
- Решение задачи оценки вероятности невозврата кредита с учетом предметно-экспертных ограничений.
- Модификация алгоритма поиска оптимальных параметров скоринговой модели.

Испытания на реальных данных. Результаты квалификационной работы бакалавра были использованы для решения конкурсной задачи ОТП Банка по классификации заемщиков. Данные были предоставлены на конференции ММРО-15.

Обзор литературы. Введем сначала необходимые определения. Штрафная функция (регуляризатор) $P(\boldsymbol{\beta})$ – функция, зависящая от параметров модели и не зависящая от данных, ограничивающая вектор параметров модели. В нашем случае параметрами модели являются коэффициенты $\boldsymbol{\beta}$ признаков, всего их p. Регуляризатор складывается с функцией отрицательного лог-правдоподобия с коэффициентом-структурным параметром (обозначается обычно λ). После этого решается задача поиска глобального минимума получившейся целевой функции и оптимальных праметров модели, доставляющих это минимум:

$$[\beta_0, \boldsymbol{\beta}] = \arg \min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} (-L(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})).$$

В случае, если регуляризатор состоит из нескольких слагаемых, домноженных на коэффициенты, то эти коэффициенты будем называть также структурными параметрами.

Регуляризаторы служат, как прваило для уменьшения размерности признакового пространства. Некоторые из регулязитаров обладают селективными способностями, т.е. с их помощью обнуляются некоторые коэффициенты признаков.
Впервые штрафная функция elastic-net была представлена в [2] для случая линейной регрессии. До этого существовали такие типы регуляризаторов, как:

- лассо (L_1–регуляризация) [3]

$$\lambda P_(\boldsymbol{\beta}) = \lambda ||\boldsymbol{\beta}||_1 = \lambda \sum_{j=1}^{p} |\beta_j|;$$

- регуляризатор Тихонова (L_2–регуляризация, можно встретить название ридж) [4]

$$\lambda P(\boldsymbol{\beta}) = \lambda ||\boldsymbol{\beta}||_2^2 = \lambda \sum_{j=1}^{p} \beta_j^2.$$

Опишем свойства этих регуляризаторов. Лассо обладает свойством отбора признаков, но в случае сильно коррелированных признаков оно отбирает всего один из них [5]. Кроме того, данный штраф не является непрерывно дифференцируемым. С помощью регуляризатора Тихонова сжимаются, но не обнуляются коэффициенты признаков. Elastic-net – это линейная комбинация лассо и регуляризатора Тихонова [2]:

$$\lambda P_\alpha(\boldsymbol{\beta}) = \lambda \left(\frac{1-\alpha}{2} ||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1 \right) = \lambda \sum_{j=1}^{p} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right).$$

С помощью elastic-net можно отбирать все коррелированные признаки и обнулять коэффициенты. При этом сохраняется проблема недифференцируемости в нуле, а также количество структурных параметров увеличилось до двух (у лассо и регуляризатора Тихонова он всего один). В [1] был представлен алгоритм для нахождения коэффициентов признаков в случае обобщенно-линейных регрессионных моделей в случае регуляризатора elastic-net. Данный алгоритм является базовым в данной работе.
В [6] был представлен так называемый адаптивный elastic-net для линейной регрессии. В нем коэффициенты параметров взвешиваются:

$$P(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} w_j |\beta_j|.$$

Адаптивный elastic-net используется для уменьшения размерности признакового пространства.
Как показано в [1, 7], алгоритм elastic-net является наиболее быстродействующим из большинства известных алгоритмов и доставляет наибольшее значение AUC. Поэтому в данной работе предлагается задавать дополнительные экспертно-интерпретируемые ограничения на коэффициенты признаков логистической регрессии в случае штрафной функции elastic-net.

Поскольку для elastic-net подобные исследования ранее не проводились, следует обратить внимание на работы, посвященные дополнительным ограничениям в других случаях.
Например, ограничением являются равенства на линейные комбинации коэффициентов:

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{b}, \text{ где}$$

$$\mathbf{C} \in \mathbb{R}^{n \times p}, \mathbf{b} \in \mathbb{R}^n.$$

Исследования на данную тему проведены в [8]. Данные ограничения могут быть использованы, если эксперты считают, что коэффициенты некоторых признаков равны по модулю и различны по знаку, либо, например, если мы предполагаем, что следующий по индексу признак в некоторое количество раз больше предыдущего(или нескольких предыдущих).
Проблема определения знака признака коэффициента исследуется в работе [9]. Здесь предлагается определить знак признака при помощи удаления и внесения признака в модель. Также в [10] описывается так называемая "постмодельная селекция". В случае, если предполагался, например, положительный коэффициент, но был получен отрицательный, эксперты предлагают удалить данный признак из модели. Однако, это может привести к ухудшению качества модели.
Еще одно ограничение, которое возникает при экспертных оценках – упорядочение признаков по абсолютному значению. Так, например, эксперты предполагают, чтобы в некотором подвекторе следующий коэффициент был больше предыдущего. С этой целью была предложена штрафная функция fused-lasso [11]. Кроме суммы модулей, добавляется сумма модулей разностей между коэффициентами:

$$P(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

Эти дополнения позволяют ограничить возрастание весов друг относительно друга и упорядочить их. В [12] показано, как можно добавить веса для увелечения селективных способностей, а также дано расширение модели на обобщенно-линейную регрессию.
Регуляризатор HORSES, описанный в [13], напоминает fused-lasso, но, отличие от последнего, берется разность не между двумя последовательными коэффициентами:

$$\lambda P(\boldsymbol{\beta}) = \lambda \left(\sum_{j=1}^{p} \alpha |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \right).$$

Данный метод позволяет собирать группами коэффициенты, коррелированные между собой.
Также упорядочить коэффициенты и сжать их позволяет метод OSCAR, представленный в [14] для линейной регрессии и распространенный на обощенно-линейные модели в [15]. В данном методе помимо суммы модулей также ограничивает сумма максимумов коэффициентов до k-го:

$$\lambda P(\boldsymbol{\beta}) = \lambda \left(\sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \right).$$

Кроме того, данный метод позволяет сделать коэффициенты корреллированных признаков равными.
В статье [16] приведен алгоритм для штрафной функции лассо, который позволяет ограничить не только вектор коэффициентов, но и любой из его подвекторов:

$$||\boldsymbol{\beta}_{\mathcal{A}}|| \leq t_{\mathcal{A}}$$

В данном алгоритме задача переформулируется эквивалентным образом: вместо добавления штрафа к правдоподобию появляется ограничение на сумму модулей – она должна быть меньше некого параметра. Далее показывается, что решение всегда находится на границе и рассматривается случай, когда мы аналогичным образом ограничиваем любой из подвекторов. Это позволяет нам ограничить веса искусственно построенных признаков (например, с помощью решающих деревьев), обладающих небольшим покрытием выборки.

2 Постановка задачи

Имеются исходные данные – выборка

$$D = \{(\mathbf{x}_1, y_1),\ \ldots,\ (\mathbf{x}_i, y_i),\ \ldots,\ (\mathbf{x}_m, y_m)\},$$

$$\mathbf{x}_i \in \mathbb{R}^p, y_i \in \{1, 0\}\ \text{(возврат-невозврат денег)}.$$

Далее для удобства введем множество индексов объектов: $i \in I = \{1,\ \ldots,\ m\}$. Векторы признаков $\mathbf{x}_i$ можно записать как матрицу признаков

$$\mathbf{X} \in \mathbb{R}^{m \times p}, \mathbf{X} = [\mathbf{x}_1^\intercal,\ \ldots,\ \mathbf{x}_i^\intercal,\ \ldots,\ \mathbf{x}_m^\intercal]^\intercal.$$

Аналогично матрице признаков введем вектор ответов

$$\mathbf{y} = (y_1,\ \ldots, y_i,\ \ldots, y_m)^\intercal.$$

Предлагается разделить исходную выборку D на на обучающую

$$L = \{(x_i, y_i)\},\ i \in \mathcal{L}$$

и контрольную

$$T = \{(x_i, y_i)\},\ i \in \mathcal{T},$$
$$I = \mathcal{L} \sqcup \mathcal{T}.$$

На выборке L будет проходить обучение алгоритма, а на T – его проверка. Разбиение предполагается делать случайно. Пусть $|L| = n$. Перенумеруем для удобства объекты и ответы таким образом, чтобы

$$L = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

Стандартизируем данные, т.е.

$$\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1.$$

Решается задача классификации с помощью логистической регрессии. Предположим, что y_i – случайная величина, имеющая распределение Бернулли, при этом все y_i независимы в совокупности:

$$y_i \sim Be(\pi(\mathbf{x}_i)),$$

$$\pi(\mathbf{x}_i) = \frac{1}{1 + exp(-\beta_0 - x_i^\intercal \boldsymbol{\beta})}.$$

Модель имеет следующий вид:

$$y_i = \pi(\mathbf{x}_i) + \varepsilon.$$

Здесь $\boldsymbol{\beta}$-вектор коэффициентов признаков, его значение определяется по обучающей выборке с помощью принципа наибольшего правдоподобия:

$$[\beta_0, \boldsymbol{\beta}] = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max}\, L(\beta_0, \boldsymbol{\beta}) = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max}\, \frac{1}{n} \sum_{i=1}^{n} \left(y_i(\beta_0 + x_i^\intercal \boldsymbol{\beta}) - \log(1 + e^{\beta_0 + x_i^\intercal \boldsymbol{\beta}}) \right).$$

Здесь функция лог-правдоподобия:

$$L(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i(\beta_0 + x_i^\intercal \boldsymbol{\beta}) - \log(1 + e^{\beta_0 + x_i^\intercal \boldsymbol{\beta}}) \right).$$

Для отбора признаков предлагается использовать штрафную функцию elastic-net, объединяющую L_1– и L_2–регуляризации(лассо и регуляризатор Тихонова, соответственно). При использовании данной штрафной функции задача будет выглядит следующим образом:

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \big(-L(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta}) \big),$$

$$\text{где } P_\alpha(\boldsymbol{\beta}) = \frac{1 - \alpha}{2} ||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} \left(\frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha |\beta_j| \right).$$

Предлагается рассмотреть дополнительные ограничения типа

$$\beta_j \geq 0, \ (*)$$

$$\beta_{j+1} \geq \beta_j, \ (**)$$

$$||\boldsymbol{\beta}_{\mathcal{A}}|| \leq t_{\mathcal{A}}, (***)$$

где $\boldsymbol{\beta}_{\mathcal{A}}$– подвектор $\boldsymbol{\beta}$.

3 Предметно-экспертные ограничения

Назовем ограничения типа $(*)$,$(**)$ и $(***)$ *предметно-экспертными*, поскольку данные ограничения вводятся решающими задачу кредитного скоринга аналитиками для повышения интерпретируемости результата.
В общем случае данные ограничения не улучшают качество модели (критерием служит AUC – площадь под ROC-кривой). Тем не менее, приведем несколько примеров, когда данные ограничения все же повышают обобщающую способность или качество алгоритма. Подобные случаи возможны в первую очередь из-за неправильно составленной (неслучайной) обучающей выборки.

Как правило, одним из признаков является размер кредита. Экспертно предполагается, что с увеличением суммы кредита риски увеличиваются, т.е. данный признак должен войти в модель с положительным коэффициентом. Если обучающая выборка составлена неправильно, возможна обратная ситуация. В данном случае неправильно составленной обучающей выборкой служит, например, следующая: с увеличением суммы кредита количество объектов целевого класса увеличивается. В таком случае следует поставить дополнительное ограничение:

$$\beta_{\text{sum of credit}} \geq 0$$

На этом же примере покажем возможное примение другого ограничения. В некоторых случаях количественные признаки (например, сумма кредита) градуируются, т.е. разбиваются на интервалы определенным образом. После этого вводятся «искусственные» признаки для каждого из интервалов – попадание в интервал обозначается как 1, непопадание – как 0. Экспертная оценка в таком случае предполагает, что чем большей сумме соответствует данный признак-интервал, тем больший коэффициент должен перед ним стоять. В подобном случае можно ввести ограничение следующего рода:

$$\beta_{j+1} \geq \beta j$$

Приведем также пример для использования ограничения типа

$$||\boldsymbol{\beta}_{\mathcal{A}}|| \leq t_{\mathcal{A}}, \text{ где}$$

$$\boldsymbol{\beta}_{\mathcal{A}} \text{ – подвектор } \boldsymbol{\beta}.$$

Пусть в исходных данных существуют, например, бинарные признаки, которые принимают значение 1 только на объектах целевого класса. Пусть также каждый из них при этом покрывает лишь малую часть целевого класса. В таком случае возможно переобучение, обусловленное как раз этими признаками (особенно если обучающая выборка составлена некорректно). Эффект переобучения можно уменьшить, если ввести ограничение сверху на норму подвектора, составленного из данных признаков. В этом случае предлагается сначала обучиться без каких-либо ограничений и найти норму посчитанного вектора коэффициентов. После этого можно взять в качестве ограничения сверху, например, одну сотую от полученной нормы.
Обобщим предыдущие ограничения:

$$\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b},$$

$$\mathbf{C} \in \mathbb{R}^{n \times p},$$

$$\mathbf{b} \in \mathbb{R}^{n}.$$

Если заменить неравенство на равенство, то получим ограничение на линейные комбинации.
Сведем данные примеры в одну таблицу.

Таблица 1: Примеры предметно-эксперных ограничений

Ограничение	Пример использования	Пример поля скоринговой карты
$\beta_j \geq 0$	Экспертно предполагается, что с увеличением признака (количественного) увеличиваются риски.	Сумма кредита.
$\beta_{j+1} \geq \beta j$	Аналогично предыдущему ограничению, если количественный признак градуировать.	Сумма кредита, возраст
$\|\|\boldsymbol{\beta}_{\mathcal{A}}\|\| \leq t_{\mathcal{A}}$, $\boldsymbol{\beta}_{\mathcal{A}}$– это подвектор $\boldsymbol{\beta}$.	Позволяет снизить эффект переобучения, ограничив веса определенных признаков.	Наличие домашнего телефона
$C\boldsymbol{\beta} = \mathbf{b}$ $C \in \mathbb{R}^{n \times p}$ $\mathbf{b} \in \mathbb{R}^{n}$	Позволяет ввести линейные ограничения на коэффициенты признаков. Если заменить знак = в ограничении на $\leq$, то является обобщением предыдущих ограничений.	Выше перечисленные поля

Таким образом, на основе данных примеров показано, что различные предметно-экспертные ограничения повышают интерпретируемость конечного результата, а также позволяют уменьшить переобучение и сделать некоторые выводы относительно корректности составления обучающей выборки.

4 Базовый алгоритм

Опишем алгоритм, представленный в [1]. Для упрощения задачи предлагается использовать квадратичную аппроксимацию функции лог-правдоподобия

$$L(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i(\beta_0 + x_i^{\mathsf{T}}\boldsymbol{\beta}) - \log(1 + e^{\beta_0 + x_i^{\mathsf{T}}\boldsymbol{\beta}}) \right),$$

получаемую с помощью использования формулы Тейлора:

$$L_Q(\beta_0, \boldsymbol{\beta}) = -\frac{1}{2n} \sum_{i=1}^{n} w_i(z_i - \beta_0 - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2 + C,$$

где $z_i = \hat{\beta}_0 + \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}} + \dfrac{y_i - \hat{\pi}(\mathbf{x}_i)}{\hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i))}$ и $w_i = \hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i))$

Используя необходимое и достаточное условие глобального экстремума для выпуклой функции, получаем выражение для β_j.

Теорема 1. *Пусть решается задача*

$$[\beta_0, \boldsymbol{\beta}] = \arg \min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} (-L_Q + \lambda P_\alpha(\boldsymbol{\beta}))$$

без предметно-экспертных ограничений, тогда оценка оптимальных параметров:

$$\hat{\beta}_j = \frac{S(\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \hat{y}_i^{(j)}), \lambda\alpha)}{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}^2 + \lambda(1-\alpha)}.$$

Здесь S – оператор сжатия,

$$S(a, b) = sign(a)(|a| - b)_+ \ u$$

$$\hat{y}_i^{(j)} = \hat{\beta}_0 + \sum_{l \neq j} x_{il}\hat{\beta}_l.$$

Доказательство. Так как в данном случае функция $R(\beta_0, \boldsymbol{\beta}) = -L_Q(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})$ определена и выпукла на всем пространстве, то необходимое и достаточное условие минимума представляется в следующем виде:

$$0 \in \partial\{R(\beta_0, \boldsymbol{\beta})\} \ (****)$$

Распишем это условие для определенного β_j:

1. Случай $\beta_j > 0$.

   В данном случае функция $R(\beta_0, \boldsymbol{\beta})$ дифференцируема, тогда из условия $(****)$

$$-\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \beta_0 - x_i^\intercal \boldsymbol{\beta}) + \lambda(1-\alpha)\beta_j + \lambda\alpha = 0,$$

$$\beta_j = \frac{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \hat{y}_i^{(j)}) - \lambda\alpha}{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}^2 + \lambda(1-\alpha)}.$$

2. Случай $\beta_j < 0$. В данном случае функция $R(\beta_0, \boldsymbol{\beta})$ дифференцируема, тогда из условия $(****)$ аналогично

$$-\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \beta_0 - x_i^\intercal \boldsymbol{\beta}) + \lambda(1-\alpha)\beta_j - \lambda\alpha = 0,$$

$$\beta_j = \frac{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \hat{y}_i^{(j)}) + \lambda\alpha}{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}^2 + \lambda(1-\alpha)}.$$

3. Случай $\beta_j = 0$: Из условия $(****)$:

$$0 \in \left\{ -\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \beta_0 - x_i^\intercal \boldsymbol{\beta}) + \lambda(1-\alpha)\beta_j + \lambda\alpha\partial\{|\beta_j|\} \right\},$$

$$\text{где } \partial\{|\beta_j|\} = [-1;1]$$

$$\beta_j \in \left\{ \frac{\frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}(z_i - \hat{y}_i^{(j)}) + \lambda\alpha\partial|\beta_j|}{\frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}^2 + \lambda(1-\alpha)} \right\}.$$

Таким образом, объединяя все три случая получаем следующее выражение для оптимальных параметров модели:

$$\beta_j = \frac{S(\frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}(z_i - \hat{y}_i^{(j)}), \lambda\alpha)}{\frac{1}{n}\sum_{i=1}^{n} w_i x_{ij}^2 + \lambda(1-\alpha)},$$

где S – оператор сжатия,

$$S(a,b) = sign(a)(|a| - b)_+.$$

□

Данный алгоритм не учитывает предметно-экспертных ограничений. Дальнейшее исследование посвящено модернизации базового алгоритма с учетом возникающих ограничений.

5 Метод внешних штрафных функций

Для модификации алгоритма предлагается воспользоваться методом внешних штрафных функций. Поэтому опишем его подробнение в данном разделе.
Расссмотрим задачу

$$f^* = f(x^*) = \min_{x\in\mathbb{R}^n} f(x), \text{ при ограничениях}$$

$$\varphi_i(x) \le 0, i = 1, \dots, m,$$
$$\varphi_i(x) = 0, i = m+1, \dots, l,$$
$$x \in \mathbb{R}^n.$$

Тогда x^* – решение данной задачи.
Пусть все эти ограничения задают некоторое множество $G \in \mathbb{R}^n$. Тогда определим индикаторную функцию множества G следующим образом:

$$\delta(x|G) = \begin{cases} 0, x \in G, \\ +\infty, x \notin G. \end{cases}$$

Пусть

$$F(x) = f(x) + \delta(x|G).$$

Тогда рассматриваемая задача эквивалентна следующей задаче безусловной минимизации:

$$\min_{x\in\mathbb{R}^n} F(x).$$

Для решения подобной задачи удобно воспользоваться следующим подходом. Пусть

$$\delta(x|G) = \lim_{k\to\infty} \delta_k(x|G).$$

Тогда назовем функции $\delta_k(x|G)$ штрафными и будем решать задачи

$$\min_{x\in\mathbb{R}^n} F_k(x) = \min_{x\in\mathbb{R}^n} f(x) + \delta_k(x|G),$$

$$x_k^* = \arg\min_{x\in\mathbb{R}^n} F_k(x).$$

Рассмотрим внешние штрафные функции, т.е. штрафные функции, удовлетворяющие следующим свойствам:

1. $$\delta_k(x|G) = 0, x \in G$$
2. $$\delta_k(x|G) > 0, x \notin G$$
3. $$\delta_{k+1}(x|G) > \delta_k(x|G), x \notin G$$
4. $$\lim_{k\to\infty} \delta_k(x|G) = \delta(x|G)$$

Тогда доказана следующая теорема:

Теорема 2. *Пусть $f(x)$ – непрерывна на $\mathbb{R}^n$, пусть также заданы внешние штрафные функции $\delta_k(x|G)$, $x_k^* = \arg\min_{x\in\mathbb{R}^n} F_k(x)$, тогда существует предельная точка x^*:* $x^* = \lim_{k\to\infty} x_k^*$, $\lim_{k\to\infty} f(x_k^*) = f(x^*) = f^*$.

6 Модификация базового алгоритма

Воспользуемся методом штрафных функций для модификации базового алгоритма. Рассмотрим, например, ограничение типа

$$\varphi(\boldsymbol{\beta}) \le 0,$$

где $\varphi(\boldsymbol{\beta})$ – выпуклая функция.
Предлагается вместо исходной задачи

$$\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}(-L(\beta_0,\boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})),$$

$$\varphi(\boldsymbol{\beta}) \le 0,$$

решать следующие задачи:

$$\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}(-L(\beta_0,\boldsymbol{\beta})+\lambda P_\alpha(\boldsymbol{\beta}))+\delta_k(\boldsymbol{\beta}),$$

$$\text{где } \delta_k(\boldsymbol{\beta})=r_k(\varphi(\boldsymbol{\beta}))_+^2,\ k=1,2,....$$

Для упрощения расчетов предлагается, как и в базовом алгоритме, воспользоваться разложением по формуле Тейлора и решать следующие задачи:

$$[\beta_0,\boldsymbol{\beta}]=\arg\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}(-L_Q(\beta_0,\boldsymbol{\beta})+\lambda P_\alpha(\boldsymbol{\beta}))+r_k(\varphi(\boldsymbol{\beta}))_+^2,$$

$$r_{k+1}>r_k,\ ,k=1,2,....$$

Если найденные в $(k+1)$-й задачи коэффициенты отличаются по норме от найденных в k-й не более, чем на ε, то алгоритм останавливается.
Воспользуемся необходимым и достаточным условием минимума для выпуклой функции и приходим к следующей теореме:

Теорема 3. *Пусть решается задача*

$$[\beta_0,\boldsymbol{\beta}]=\arg\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}(-L_Q(\beta_0,\boldsymbol{\beta})+\lambda P_\alpha(\boldsymbol{\beta})),\ \textit{при условии}$$

$$\varphi(\boldsymbol{\beta})\le 0,\varphi(\boldsymbol{\beta})\ \textit{– выпуклая функция,} \tag{1}$$

тогда вместо данной задачи можно решать следующие:

$$[\beta_0,\boldsymbol{\beta}]=\arg\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}(-L_Q(\beta_0,\boldsymbol{\beta})+\lambda P_\alpha(\boldsymbol{\beta}))+r_k(\varphi(\boldsymbol{\beta}))_+^2,\ r_{k+1}>r_k>0, \tag{2}$$

и для каждого из $\beta_j, j=1,...,p$ выполняется условие:

$$0\in-\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i-\beta_0-x_i^\intercal\boldsymbol{\beta})+\lambda(1-\alpha)\beta_j+\lambda\alpha\partial|\beta_j|+2r_k(\varphi(\boldsymbol{\beta}))_+\partial\varphi(\boldsymbol{\beta}). \tag{3}$$

Доказательство. Функция

$$R(\beta_0,\boldsymbol{\beta})=-L_Q(\beta_0,\boldsymbol{\beta})+\lambda P_\alpha(\boldsymbol{\beta})$$

является непрерывной и выпуклой, т.к. непрерывны и выпуклы функции $-L_Q(\beta_0,\boldsymbol{\beta})$ и $P_\alpha(\boldsymbol{\beta})$. Пусть множество G задается следующим образом:

$$G=\boldsymbol{\beta}:\varphi(\boldsymbol{\beta})\le 0.$$

Функции $\delta_k(\boldsymbol{\beta}|G)=r_k(\varphi(\boldsymbol{\beta}))_+^2$ удовлетворяют свойствам внешних штрафных функций, если $r_{k+1}>r_k>0$. Тогда, воспользовавшись теоремой 2, получаем, что вместо задачи (1) мы можем решать задачи (2).
Т.к. $\varphi(\boldsymbol{\beta})$ – выпуклая функция, то $\delta_k(\boldsymbol{\beta}|G)$ и $R(\beta_0,\boldsymbol{\beta})+\delta_k(\boldsymbol{\beta}|G)$ – также выпуклые функции. Тогда применим необходимое и достаточное условие глобального минимума для выпуклой функции, и для каждой из компонент вектора $\boldsymbol{\beta}$ получаем:

$$0\in\partial(R(\beta_0,\boldsymbol{\beta})+\delta_k(\boldsymbol{\beta}|G)).$$

Выписав теперь субдифференциал функции $R(\beta_0,\boldsymbol{\beta})+\delta_k(\boldsymbol{\beta}|G)$ по каждой из компонент приходим к условию:

$$0\in-\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i-\beta_0-x_i^\intercal\boldsymbol{\beta})+\lambda(1-\alpha)\beta_j+\lambda\alpha\partial|\beta_j|+2r_k(\varphi(\boldsymbol{\beta}))_+\partial\varphi(\boldsymbol{\beta}).$$

□

Приведем пример применения теоремы. Рассмотрим, например, ограничение типа

$$\beta_{j+1} \geq \beta_j.$$

Предлагается вместо исходной задачи решать следующие:

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} (-L_Q(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})) + \delta_k(\boldsymbol{\beta}),$$

$$\text{где } \delta_k(\boldsymbol{\beta}) = r_k(\beta_j - \beta_{j+1})_+^2.$$

Используя условие (3), получаем выражения для β_l:

$$\beta_l = \frac{S(\frac{1}{n}\sum_{i=1}^n w_i x_{il}(z_i - \hat{y}_i^{(k)}), \lambda\alpha)}{\frac{1}{n}\sum_{i=1}^n w_i x_{il}^2 + \lambda(1-\alpha)}, \; l \neq j,$$

$$\beta_j = \frac{S(\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \hat{y}_i^{(j)}) + r_k\beta_{j+1}[\beta_j > \beta_{j+1}], \lambda\alpha)}{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}^2 + \lambda(1-\alpha) + r_k[\beta_j > \beta_{j+1}]},$$

$$\beta_{j+1} = \frac{S(\frac{1}{n}\sum_{i=1}^n w_i x_{ij}(z_i - \hat{y}_i^{(j)}) + r_k\beta_j[\beta_j > \beta_{j+1}], \lambda\alpha)}{\frac{1}{n}\sum_{i=1}^n w_i x_{ij}^2 + \lambda(1-\alpha) + r_k[\beta_j > \beta_{j+1}]}.$$

Аналогичным образом выписываются выражения, которые получаются для остальных ограничений.

7 Вычислительный эксперимент

Вычислительный эксперимент проводился на конкурсных данных, предоставленных ОТП банком в рамках конференции ММРО-15.
Данные представляли собой обучающую и контрольную выборки с известными ответами. Имелись как количественные, так и номинальные признаки (всего 50 признаков, после градуирования – 101).
Размер исходной обучающей выборки – 15 223 объекта, контрольной – 14 910.
Были проверены гипотезы, представленные вданной работе в разделе 3 "Практическое применение". Составлялись некорректные обучающие выборки и на них строились модели с предметно-экспертными ограничениями и без них. Как показали вычислительные эксперименты, гипотезы об улучшении качества в данной ситуации были оправданы.
Критерием качества служил AUC.

В общем случае качество модели на контроле не улучшается, т.к. мы ищем условный минимум (см. рис 1).
Если составить несбалансированную некорректную обучающую выборку (т.е., например, с увеличением суммы кредита увеличивается число объектов целевого класса), то модель с ограничениями показывает лучший результат на контроле (см. рис 2).

8 Заключение

В работе был предложен метод, позволяющий эксперту-аналитику внести дополнительные ограничения на признаки, используемые в модели. Это позволяет улучшить

интерпретируемость полученных результатов, а также повысить качество классификации, в случае, если обучающая выборка составлена некорректно. Кроме того, данный метод позволяет отбирать наиболее информативные признаки.

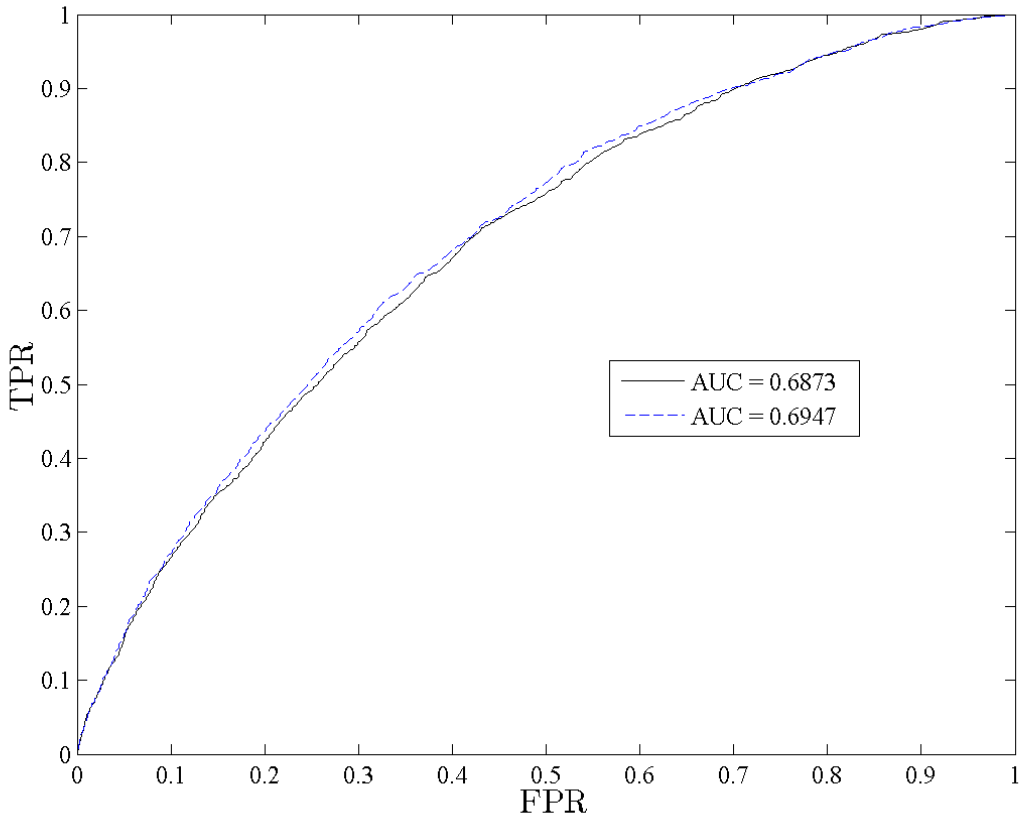

Рис. 1: Общий случай

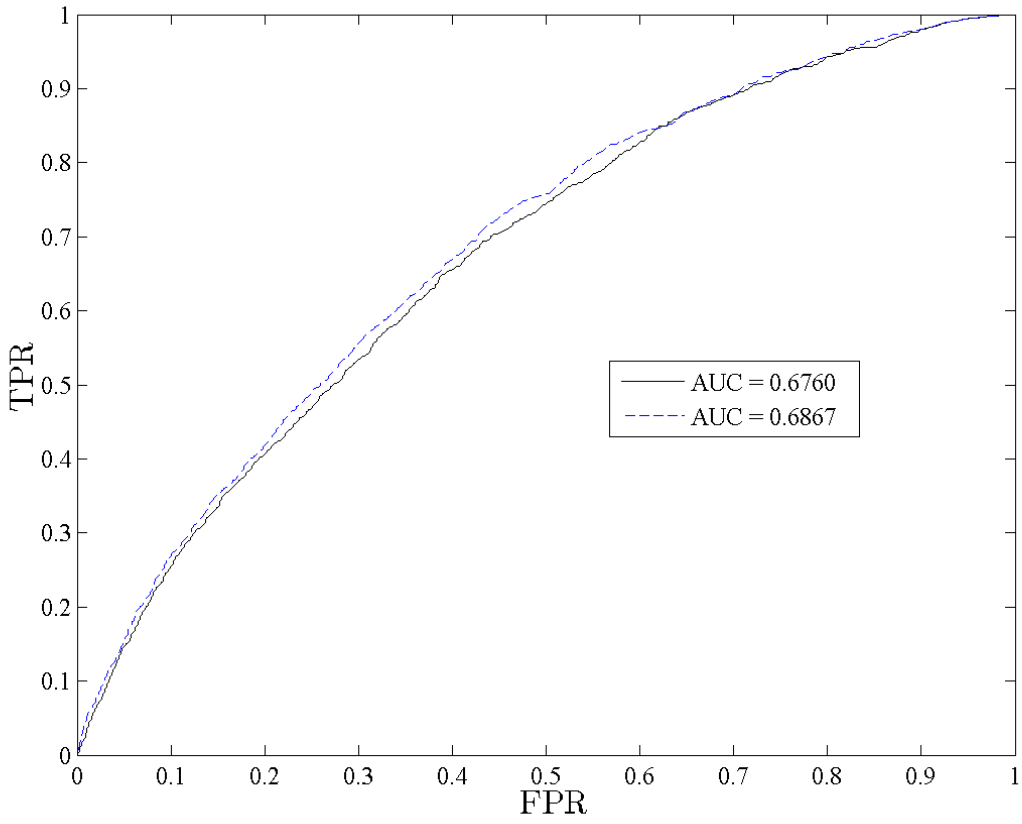

Рис. 2: Некорректная обучающая выборка

Список литературы

[1] Tibshirani R. Friedman J., Hastie T. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

[2] Hastie T. Zou H. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67:301–320, 2005.

[3] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[4] Lemeshow S. Hosmer D. W. *Applied Logistic Regression*. John Wiley & Sons, 2000.

[5] Стрижов В. В Катруца А. М. Проблема мультиколлинеарности при выбора признаков в регрессионных задачах. *Информационные технологии*, (9), 2014.

[6] Zhang H. H. Zou H. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733–1751, 2009.

[7] Makalic E. Schmidt D. F. Review of modern logistic regression methods with application to small and medium sample size problems. Technical report, The University of Melbourne.

[8] Sivaraname N. Jaggi S. Restrictions in regression model.

[9] Visco I. On obtaining the right sign of a coefficient estimate by omitting a variable from the regression. *Journal of Econometrics*, 7:115–117, 1978.

[10] Hassler U. Testing regression coefficients after model selection through sign restrictions. *Economics Letters*, 107:220–223, 2010.

[11] Rosset S. Zhu J. Knight K. Tibshirani R., Saunders M. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 1(67):91–108, 2005.

[12] Flexeder C. Generalized lasso regularization for regression models. Technical report, Ludwig-Maximilians-Universitet at Munchen Institut fur Statistik.

[13] Lazar N. A. Loh J. M. Yu D. Jang W., Lim J. Regression shrinkage and grouping of highly correlated predictors with horses. Technical report, 2013.

[14] Bondell H. D. Reich B. J. Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics*, 64:115–123, (2008).

[15] Petry S. Tutz G. The oscar for generalized linear models. Technical report, Ludwig-Maximilians-Universitet Munchen, 2011.

[16] Kim Y. Kim J., Kim Y. Gradient lasso algorithm.