Выбор признаков в задачах логистической регрессии

К. С. Скипор
skiporkonstantin@mail.ru
Московский физико-технический институт

Предлагается и исследуется алгоритм отбора признаков для решения задач восстановления логистической регрессии. Алгоритм основан на методе наименьших углов для модели линейной регрессии с использованием дополнительно линеаризации функционала качества. Приводится математическое обоснование предложенного алгоритма. Работа алгоритма проиллюстрирована задачей изучения факторов риска ишемических заболеваний сердца.

Ключевые слова: *логистическая регрессия, выбор признаков, метод наименьших углов, линейное программирование*

Введение

В работе рассматривается отыскание из множества признаков такого его подмножества, для которого их линейная комбинация наиболее точно описывает данные. В 1966 году Дрейпером был предложен ступенчатый алгоритм выбора признаков (Forward Stagewise) [1, 2, 3]. На каждой итерации алгоритма выбирается признак, имеющий наибольшую проекцию на вектор ответов, после этого делается небольшое смещение текущего приближения функции регрессии в направлении выбранного признака. Среди полученных на каждой итерации моделей находится оптимальная, тем самым производится отбор признаков. Алгоритм Forward Selection [4] представляет собой модифицированную версию Forward Stagewise. Основное отличие заключается в выборе величины смещения. Смещение выбирается таким, чтобы максимизировать приращение функционала качества для выбранного признака.

В 1970 году Хоэрл и Кеннард предложили метод гребневой регрессии (Ridge Regression) [5], в котором использовался метод регуляризации [6]. Было введено дополнительное регуляризующее слагаемое в минимизируемый функционал; стало возможным улучшить устойчивость решения [7]. Еще один метод регуляризации, Лассо (The Lasso), был предложен Тибширани в 1996 году [8]. В нем вводится ограничение на L_1-норму вектора параметров модели, что приводит к обнулению части параметров модели и улучшению устойчивости решения. В модели логистической регрессии этот метод также называется L_1-regularized Logistic Regression [2].

В 2002 году Эфрон, Хасти, Джонстон и Тибширани предложили метод наименьших углов (Least Angle Regression) [9]. Изначально медод был предложен для линейных моделей, его реализацией является алгоритм последовательного добавления признаков LARS. На каждом шаге алгоритма признак выбирается таким образом, что вектор регрессионных остатков равноуголен [10] добавленным в модель признакам. Данный метод был предложен авторами для разрешения проблемы слишком быстрой сходимости к локальному оптимуму в многоэкстремальных задачах выбора признаков [11, 12, 13]. В 2004 году Мадиган и Ридгевэй предложили идею применения данного метода при использовании линеаризации для обобщенных линейных моделей, в частности, для модели логистической регрессии [14]. Реализация этой идеи лежит в основе написания данной работы.

Данная работа состоит из пяти частей. В разделе «Постановка задачи отбора признака» ставится задача отбора признаков в модели логистической регрессии, решаемая в

даной работе. Раздел «Описание алгоритма» разделен на три сегмента. Вначале кратко реферируются основные принципы работы алгоритма LARS для линейных моделей. Далее предлагается алгоритм последовательного добавления признаков в модели логистической регрессии LALR, решающий поставленную задачу. Отличие алгоритмов состоит в используемых функционалах качества. Предложенный алгоритм использует функционал качества, соответствующий бернуллиевской гипотезе порождения данных. После формального описания дается математическое обоснование предложенного алгоритма. Доказательство основных утверждений приводится в разделе «Приложение». В разделе «Вычислительные эксперименты» иллюстрируется работа предложенного алгоритма на модельных данных и на реальных данных «SAHD». Также работа предложенного алгоритма сравнивается с работой алгоритма Forward Stagewise.

Постановка задачи отбора признаков

Дана выборка $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^m$, в которой i-й объект описывается строкой из n числовых признаков, $\mathbf{x}^i = (x_j^i)_{j=1}^n \in \mathbb{R}^n$ и метки класса $y^i \in \{0, 1\}$. Верхний индекс i указывает порядковый номер объекта выборки, нижний индекс j — порядковый номер признака. Векторы признаков $\mathbf{x}_j = (x_j^1, \ldots, x_j^i, \ldots, x_j^m)^T$ являются линейно независимыми свободными переменными, а вектор $\mathbf{y} = (y^1, \ldots, y^i, \ldots, y^m)^T$ является зависимой переменной. Без ограничения общности будем считать, что признаки $\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n$ стандартизованы

$$\|\mathbf{x}_j\|_1 = \sum_{i=1}^m x_j^i = 0, \quad \|\mathbf{x}_j\|_2 = \sum_{i=1}^m (x_j^i)^2 = 1, \quad j = 1, \ldots, n. \tag{1}$$

Предполагается, что зависимая переменная y^i имеет распределение Бернулли. Для удобства описания алгоритма обозначим матрицу признаков $X = (\mathbf{x}_1 \ldots \mathbf{x}_j \ldots \mathbf{x}_n)$ и вектор параметров $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j, \ldots, \beta_n)^T$. Принята модель логистической регрессии, согласно которой

$$\mathbf{y} = \boldsymbol{\sigma}(X, \boldsymbol{\beta}) + \varepsilon, \tag{2}$$

где $\boldsymbol{\sigma}(X, \boldsymbol{\beta})$ — сигмоидная функция

$$\boldsymbol{\sigma}(X, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-X\boldsymbol{\beta})}. \tag{3}$$

Критерием качества модели назначен функционал логарифма правдоподобия

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \big(y^i \mathbf{x}^i \boldsymbol{\beta} - \ln(1 + \exp(\mathbf{x}^i \boldsymbol{\beta}))\big). \tag{4}$$

Требуется построить такой алгоритм последовательного добавления признаков, что на каждом шаге:

— определяются набор *активных признаков* с *активным множеством* индексов $\mathcal{A}$ и соответствующий набору ненулевой вектор параметров $\boldsymbol{\beta}_{\mathcal{A}}$, такой что $\boldsymbol{\beta}_{\mathcal{A}^c} = \mathbf{0}$, $\mathcal{A} \sqcup \mathcal{A}^c = \{1, \ldots, n\}$;
— набор *активных признаков* и вектор параметров $\boldsymbol{\beta}_{\mathcal{A}}$ доставляют максимум приращению логарифма правдоподобия ℓ;
— скорость роста функционала ℓ по любому активному признаку не меньше скорости роста по любому неактивному признаку.

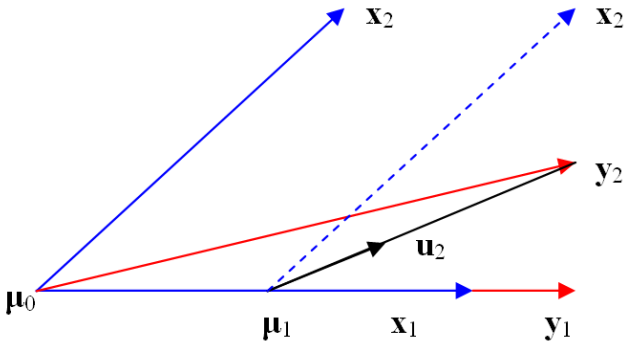

Рис. 1. *Пример работы алгоритма LARS в случае двух признаков* $\mathbf{x}_1$ *и* $\mathbf{x}_2$*. Пусть вектор* $\mathbf{y}_2$ *является проекцией вектора* $\mathbf{y}$ *на линейное подпространство* $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$*. Назначим начальное приближение* $\boldsymbol{\mu}_0 = \mathbf{0}$*. Вектор регрессионных остатков* $\mathbf{y}_2 - \boldsymbol{\mu}_0$ *коррелирует с вектором* $\mathbf{x}_1$ *больше, чем с вектором* $\mathbf{x}_2$*. Первый шаг заключается в оценке* $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \gamma_1 \mathbf{x}_1$*. Скаляр* γ_1 *выбирается таким образом, что вектор остатков* $\mathbf{y}_2 - \boldsymbol{\mu}_1$ *делит пополам угол между векторами* $\mathbf{x}_1$ *и* $\mathbf{x}_2$*. Далее получаем значение* $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \gamma_2 \mathbf{u}_2$*, где* $\mathbf{u}_2$ *- нормированный вектор, делящий этот угол пополам. Так как мы рассматриваем случай двух переменных, то* $\boldsymbol{\mu}_2 = \mathbf{y}_2$*.*

Описание алгоритма

Метод наименьших углов.

В данном подразделе предлагается краткое описание метода наименьших углов для задач линейной регрессии, см. [9]. Будем считать, что принята линейная модель

$$\mathbf{y} = \boldsymbol{\mu}(X, \boldsymbol{\beta}) + \varepsilon,$$

где функция регрессии $\boldsymbol{\mu}(X, \boldsymbol{\beta})$, представляющая собой приближение вектора y, имеет вид

$$\boldsymbol{\mu}(X, \boldsymbol{\beta}) = \sum_{j=1}^{n} \mathbf{x}_j \beta_j = X\boldsymbol{\beta}, \tag{5}$$

Критерием качества назначена среднеквадратичная ошибка

$$S(X, \boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{\mu}(X, \boldsymbol{\beta})\|^2.$$

Требуется построить такой алгоритм последовательного добавления признаков, что на каждом шаге:

— определяются набор активных признаков с активным множеством индексов $\mathcal{A}$ и соответствующий набору ненулевой вектор параметров $\boldsymbol{\beta}_{\mathcal{A}}$, такой что $\boldsymbol{\beta}_{\mathcal{A}^c} = \mathbf{0}$, $\mathcal{A} \sqcup \mathcal{A}^c = \{1, \ldots, n\}$;

— набор активных признаков и вектор параметров $\boldsymbol{\beta}_{\mathcal{A}}$ доставляют наибольшую корреляцию векторов $\mathbf{y}$ и $\boldsymbol{\mu}$;

— абсолютная корреляция вектора регрессионных остатков $\mathbf{y} - \boldsymbol{\mu}$ с любым активным признаком не меньше абсолютной корреляции вектора остатков с любым неактивным признаком.

Для решения этой задачи был предложен метод наименьших углов, реализацией которого является алгоритм LARS [9]. Рассмотрим некоторый шаг алгоритма. Пусть на этом шаге определено множество индексов $\mathcal{A}$, которое соответствует выбранным до этого шага признакам, и некоторое приближение функции регрессии $\boldsymbol{\mu}_{\mathcal{A}}$. Корреляция c_j вектора остатков

$\mathbf{y} - \boldsymbol{\mu}_{\mathcal{A}}$ на некоторый признак $\mathbf{x}_j$ вычисляется как

$$c_j = \mathbf{x}_j^T(\mathbf{y} - \boldsymbol{\mu}_{\mathcal{A}}).$$

На первом шаге выбирается признак, имеющий наибольшую абсолютную корреляцию с вектором $\mathbf{y}$.

Далее вычисляется единичный вектор $\mathbf{u}$, лежащий на биссекторе выбранных признаков. Алгоритм смещает текущее приближение $\boldsymbol{\mu}_{\mathcal{A}}$ в направлении вектора $\mathbf{u}$,

$$\boldsymbol{\mu}_{\mathcal{A}_+} = \boldsymbol{\mu}_{\mathcal{A}} + \gamma\mathbf{u},$$

где γ — коэффициент смещения, который определяется из условия, что корреляция нового вектора остатков $\mathbf{y} - \boldsymbol{\mu}_{\mathcal{A}_+}$ на некоторый неактивный признак $\mathbf{x}_d$ будет равна корреляции на все активные признаки. Здесь $\mathcal{A}_+$ есть новое активное множество индексов $\mathcal{A} \cup \{d\}$. Смещение в направлении вектора $\mathbf{u}$ обеспечивает равенство корреляций вектора остатков $\mathbf{y} - \boldsymbol{\mu}_{\mathcal{A}_+}$ на выбранные признаки, или другими словами, обеспечивает равенство углов между вектором остатков и выбранными признаками.

Рис. 1 иллюстрирует работу алгоритма в случае $n = 2$ признаков, $X = (\mathbf{x}_1, \mathbf{x}_2)$.

В следующем подразделе описывается алгоритм отбора признаков для модели логистической регрессии, после чего будет дано математическое обоснование приведенного алгоритма.

Алгоритм LALR.

В настоящей работе предлагается новый алгоритм выбора признаков при восстановлении логистической регрессии — «Least Angle Logistic Regression (LALR)». Принята модель логистической регрессии (2), (3). Обозначим множество индексов параметров $\mathcal{I} = \{1, 2, \ldots, n\}$. Для некоторого подмножества индексов $\mathcal{A} \subseteq \mathcal{I}$, назовем его *активным множеством*, определим матрицу *активных признаков*

$$X_{\mathcal{A}} = (\cdots\ s_j\mathbf{x}_j\ \cdots)_{j\in\mathcal{A}}, \tag{6}$$

где s_j, назовем его *знаком корреляции*, принимает значения ± 1. Определим также матрицы разностей и сумм между активными признаками и некоторым фиксированным неактивным признаком $\mathbf{x}_d$, где $d \in \mathcal{A}^c$, в разбиении $\mathcal{A} \sqcup \mathcal{A}^c = \mathcal{I}$,

$$\begin{aligned} M_{d-} &= (\ \cdots\ \ s_j\mathbf{x}_j - s_d\mathbf{x}_d\ \ \cdots\ )_{j\in\mathcal{A}}, \\ M_{d+} &= (\ \cdots\ \ s_j\mathbf{x}_j + s_d\mathbf{x}_d\ \ \cdots\ )_{j\in\mathcal{A}}. \end{aligned} \tag{7}$$

Опишем алгоритм последовательного добавления признаков. Начальные значения положим

$$\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\beta} = \mathbf{0}, \quad \mathcal{A} = \varnothing. \tag{8}$$

Рассмотрим некоторый шаг алгоритма. Пусть $\boldsymbol{\mu}_{\mathcal{A}}$ есть текущее приближение функции регрессии на этом шаге. Тогда вектор текущих корреляций между признаками и вектором регрессионных остатков $\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})$ имеет вид:

$$\mathbf{c} = X^T\left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})\right). \tag{9}$$

Положим знак корреляции

$$s_j = \operatorname{sign}(c_j) \qquad j \in \mathcal{I}. \tag{10}$$

Вычисляем матрицы $X_{\mathcal{A}}$, M_{d-} и M_{d-}, согласно (6) и (7), для $d \in \mathcal{A}^c$. Для удобства изложения введем матрицу весов объектов W, матрицы A_{d-}, A_{d+} и векторы $\mathbf{b}_{d-}$, $\mathbf{b}_{d+}$. Обозначим диагональную $m \times m$ матрицу W с элементами

$$W_{ii} = \sigma_i(\boldsymbol{\mu}_{\mathcal{A}})(1 - \sigma_i(\boldsymbol{\mu}_{\mathcal{A}})), \tag{11}$$

где i — номер объекта. Также обозначим матрицы A_{d-}, A_{d+} и векторы $\mathbf{b}_{d-}$, $\mathbf{b}_{d+}$

$$A_{d\pm} = M_{d\pm}^T W X_{\mathcal{A}}, \tag{12}$$

$$\mathbf{b}_{d\pm} = M_{d\pm}^T(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})), \tag{13}$$

для всех $d \in \mathcal{A}^c$. Двойной знак «$\pm$» используется для компактной записи двух выражений с «$+$» и «$-$».

Далее, используя введенные обозначения, вычисляем множество векторов Υ, которое, как будет доказано ниже, содержит оптимальный вектор коэффициентов,

$$\Upsilon = \{A_{d-}^{-1}\mathbf{b}_{d-}, A_{d+}^{-1}\mathbf{b}_{d+}\}_{d\in\mathcal{A}^c}. \tag{14}$$

В приложении показано, что в предположениях поставленной задачи матрица A всегда имеет обратную матрицу A^{-1}. Алгоритм обновляет текущее приближение функции регрессии $\boldsymbol{\mu}_{\mathcal{A}}$

$$\boldsymbol{\mu}_{\mathcal{A}_+} = \boldsymbol{\mu}_{\mathcal{A}} + X_{\mathcal{A}}\boldsymbol{\gamma}_{\mathcal{A}}, \tag{15}$$

где оптимальный вектор коэффициентов $\boldsymbol{\gamma}_{\mathcal{A}}$ определяется из условия

$$\boldsymbol{\gamma}_{\mathcal{A}} = \arg\min_{\boldsymbol{\gamma}\in\Upsilon}{}^{+}\big((\mathbf{s}_{\mathcal{A}} \circ \mathbf{c}_{\mathcal{A}})^T\boldsymbol{\gamma}\big), \tag{16}$$

"$\min^{+}$" означает, что минимум берется только из положительных значений минимизируемой функции. Операция «$\circ$» означает поэлементное (адамарово) умножение векторов. В другой интерпретации вектор $\mathbf{s}_{\mathcal{A}} \circ \mathbf{c}_{\mathcal{A}}$ есть вектор абсолютных корреляций с компонентами $|c_j|$.

Найденное решение $\boldsymbol{\gamma}_{\mathcal{A}}$ принадлежит множеству Υ, поэтому для некоторого индекса параметров $d^* \in \mathcal{A}^c$ выполнено либо $\boldsymbol{\gamma}_{\mathcal{A}} = A_{d^*-}^{-1}\mathbf{b}_{d^*-}$, либо $\boldsymbol{\gamma}_{\mathcal{A}} = A_{d^*+}^{-1}\mathbf{b}_{d^*+}$. Так определяется оптимальный индекс d^* соответствует найденному решению $\boldsymbol{\gamma}_{\mathcal{A}}$,

$$d^* = \arg\boldsymbol{\gamma}_{\mathcal{A}}. \tag{17}$$

В случае, когда $\mathcal{A} = \varnothing$, что соответствует первому шагу, d^* находится из условия максимума абсолютной корреляции:

$$d^* = \arg\max_{d\in\mathcal{I}}|c_d|. \tag{18}$$

Таким образом, определяется индекс d^*, соответствующий оптимальному добавляемому признаку $\mathbf{x}_{d^*}$, и обновляется активное множество индексов $\mathcal{A}$, путем добавления к нему d^*:

$$\mathcal{A}_+ = \mathcal{A} \cup \{d^*\}. \tag{19}$$

Также обновляется вектор параметров $\boldsymbol{\beta}$, с учетом знака корреляции (10):

$$\boldsymbol{\beta}_{\mathcal{A}} = \boldsymbol{\beta}_{\mathcal{A}} + \mathbf{s}_{\mathcal{A}} \circ \boldsymbol{\gamma}_{\mathcal{A}}.$$

Нижний индекс $\boldsymbol{\beta}_{\mathcal{A}}$ указывает, что изменяются только компоненты, соответствующие активным признакам. Этим завершается шаг алгоритма. Формула (16) дает приближенное значение вектора коэффициентов $\boldsymbol{\gamma}_{\mathcal{A}}$, поэтому алгоритм можно проитерировать для получения точного значения.

На последнем шаге, когда активный набор индексов соответствует полному, т.е $\mathcal{A} = \mathcal{I}$, все дополнительные условия на скорость роста функционала ℓ выполнены автоматически. Поэтому оптимальный вектор параметров находится из условия максимизации логарифма правдоподобия ℓ, с помощью итерационного метода наименьших квадратов с перевзвешиванием элементов (IRLS) [15].

Далее приводится обоснование используемых выше формул.

Обоснование алгоритма.

Стратегия построения метода наименьших углов, которая была использована для выбора признаков в линейной регрессии, применяется также и в логистической регрессии, но с использованием дополнительно линеаризации.

Пусть имеется некоторое активное множество $\mathcal{A}$ и пусть к тому же известно текущее приближение функции регрессии $\boldsymbol{\mu}_{\mathcal{A}}$.

Запишем логарифм правдоподобия (4) через функцию регрессии $\boldsymbol{\mu}_{\mathcal{A}}$, (5):

$$\ell(\boldsymbol{\mu}_{\mathcal{A}}) = \sum_{i=1}^{m} \left(y^i \mu_{\mathcal{A}}(\mathbf{x}^i) - \ln(1 + \exp(\mu_{\mathcal{A}}(\mathbf{x}^i))) \right). \tag{20}$$

Рассмотрим производную логарифма правдоподобия по некоторому признаку $\mathbf{x}_j$, обозначим ее c_j:

$$c_j = \frac{d}{d\gamma} \ell(\boldsymbol{\mu}_{\mathcal{A}} + \mathbf{x}_j \gamma)\Big|_{\gamma=0}, \tag{21}$$

откуда, пользуясь (3), получим:

$$c_j = \mathbf{x}_j^T \left(\mathbf{y} - \frac{\exp(\boldsymbol{\mu}_{\mathcal{A}})}{1 + \exp(\boldsymbol{\mu}_{\mathcal{A}}))} \right) = \mathbf{x}_j^T \left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}}) \right), \tag{22}$$

В матричном виде (22) принимает следующий вид

$$\mathbf{c} = X^T \left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}}) \right). \tag{23}$$

Замечание 1. Как и в случае LARS, вектор $\mathbf{c}$ есть вектор текущих корреляций векторов признаков и вектора остатков $\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})$. Поэтому далее под вектором корреляций будем понимать вектор производных по направлению.

Обозначим знак корреляции, как это было сделано в (10),

$$s_j = \operatorname{sign}(c_j), \quad j \in \mathcal{I}. \tag{24}$$

Таким образом, определим матрицу активных признаков $X_{\mathcal{A}}$, согласно (6). Выразим новое приближение функции регрессии (15) через неизвестные коэффициенты $\boldsymbol{\gamma}$:

$$\boldsymbol{\mu}_{\mathcal{A}_+} = \boldsymbol{\mu}_{\mathcal{A}} + X_{\mathcal{A}} \boldsymbol{\gamma}. \tag{25}$$

Основная цель алгоритма заключается в поиске оптимального вектора коэффициентов $\boldsymbol{\gamma}$ и нового активного множества индексов $\mathcal{A}_+$ следующего шага.

Перейдем теперь к формальной интерпретации решаемой задачи. Под скоростью роста функционала $\ell(\boldsymbol{\mu}_{\mathcal{A}_+})$ по некоторому признаку понимается абсолютное значение производной функционала по этому признаку. Поэтому решаемая задача заключается в максимизации приращения логарифма правдоподобия (20)

$$\ell(\boldsymbol{\mu}_{\mathcal{A}_+}) - \ell(\boldsymbol{\mu}_{\mathcal{A}}) \to \max_{\boldsymbol{\gamma}}, \tag{26}$$

при условии, что абсолютная корреляция нового вектора остатков $\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})$ на любой активный признак $\mathbf{x}_j$, $j \in \mathcal{A}$ не меньше абсолютной корреляции на любой неактивный признак $\mathbf{x}_d$, $d \in \mathcal{A}^c$, см. замечание 1. Запишем это условие через производную по направлению (21):

$$\Big|\frac{d}{d\alpha}\ell(\boldsymbol{\mu}_{\mathcal{A}_+} + \mathbf{x}_j\alpha)\Big|_{\alpha=0} \geqslant \Big|\frac{d}{d\alpha}\ell(\mathbf{x}_{\mathcal{A}_+} + \mathbf{x}_d\alpha)\Big|_{\alpha=0}, \tag{27}$$

для любых $j \in \mathcal{A}$ и $d \in \mathcal{A}^c$. Пользуясь обозначениями (12), (13) сформулируем лемму о линеаризации решаемой задачи.

Лемма 1. *Задача (26), (27) при линеаризации эквивалентна задаче линейного программирования:*

$$\begin{gathered}(\mathbf{s}_{\mathcal{A}} \circ \mathbf{c}_{\mathcal{A}})^T\boldsymbol{\gamma} \to \max_{\boldsymbol{\gamma}}, \\ \begin{cases} A_{d-}\boldsymbol{\gamma} \leqslant \mathbf{b}_{d-}, \\ A_{d+}\boldsymbol{\gamma} \leqslant \mathbf{b}_{d+}, \\ \forall d \in \mathcal{A}^c. \end{cases}\end{gathered} \tag{28}$$

Задачу линейного программирования (28) можно решать обычным симплекс-методом [16, 17], но с этим возрастает трудоемкость. Следующая теорема 4 позволяет существенно сократить количество опорных точек, которые могут являться решением задачи (28). Для доказательства теоремы 4 сформулируем некоторые вспомогательные утверждения.

Лемма 2. *Пусть векторы* $\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1} \in \mathbb{R}^n$ *линейно независимы. Тогда векторы* $(\mathbf{a}_1 + \mathbf{a}_{k+1}), \ldots, (\mathbf{a}_k + \mathbf{a}_{k+1}), \mathbf{a}_{k+1}$ *также линейно независимы.*

Для использования леммы 3 определим понятие афинной зависимости векторов [16].

Определение 1. *Точки* $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \mathbb{R}^n$ *называются аффинно зависимыми, если существуют* $\lambda_1, \ldots, \lambda_k$*, не равные нулю одновременно и такие, что*

$$\sum_{i=1}^{k} \lambda_i \mathbf{a}_i = 0, \qquad \sum_{i=1}^{k} \lambda_i = 0.$$

Лемма 3. *Пусть векторы* $\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1} \in \mathbb{R}^n$ *линейно независимы. Обозначим матрицы*

$$A = (\mathbf{a}_1, \ldots, \mathbf{a}_k),\ A_+ = (\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}),\ C = (\mathbf{a}_1 - \mathbf{a}_{k+1}, \ldots, \mathbf{a}_k - \mathbf{a}_{k+1}).$$

Матрица A^TC *имеет полный ранг тогда и только тогда, когда столбцы матрицы* A^TA_+ *аффинно независимы.*

Лемма 3 используется при доказательстве существования множества Υ, определенного в (14).

С помощью следующей теоремы формулируется утверждение о решении задачи линейного программирования (28).

Теорема 4. *Если ЗЛП (28) имеет решение* $\boldsymbol{\gamma}^*$, *то*

$$\boldsymbol{\gamma}^* \in \Upsilon, \tag{29}$$

причем

$$\boldsymbol{\gamma}^* = \arg\min\nolimits^{+}_{\gamma\in\Upsilon}\{(\mathbf{s}_{\mathcal{A}} \circ \mathbf{c}_{\mathcal{A}})^T\boldsymbol{\gamma}\}; \tag{30}$$

где " $\min^{+}$" *означает, что минимум берется только из положительных значений.*

Следствие 1. *На каждом шаге алгоритма абсолютная корреляция текущего вектора остатков на любой активный признак при линеаризации одинакова и больше абсолютной корреляции на любой неактивный признак, т.е справедливо*

$$\begin{cases} s_i c_i = s_j c_j, & \forall i, j \in \mathcal{A}; \\ s_i c_i > s_d c_d, & \forall i \in \mathcal{A}, \forall d \in \mathcal{A}^c. \end{cases}$$

Следствие 1 представляет собой аналог основного свойства метода наименьших углов в линейных моделях: на каждом шаге вектор остатков лежит на биссекторе добавленных признаков.

Все доказательства приведенных утверждений приводятся в приложении.

Вычислительные эксперименты

Сравним предложенный алгоритм с описанным в [2, 3] итеративным алгоритмом Forward Stagewise. На каждом шаге алгоритм выбирает признак $\mathbf{x}_{j^*}$, имеющий наибольшую корреляцию c_{j^*} с текущим вектором остатков $\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu})$ и делает небольшое смещение γ текущего приближения в направлении выбранного признака $\mathbf{x}_{j^*}$,

$$j^* = \arg\max|c_j| \quad \text{и} \quad \boldsymbol{\mu} \to \boldsymbol{\mu} + \gamma\,\mathrm{sign}(c_{j^*})\mathbf{x}_{j^*}.$$

Чем меньше абсолютная величина смещения γ, тем точнее получается оценка параметров $\boldsymbol{\beta}$. Но с уменьшением смещения увеличивается количество шагов и, тем самым, возрастает время выполнения алгоритма.

Модельные данные.

Сгенерируем $m = 50$ объектов с пятью независимыми, нормально распределенными признаками $\mathbf{x}_1, \cdots, \mathbf{x}_5$, т.е $\mathbf{x}_i = (x_{i1}, \cdots, x_{i5}) \sim \mathcal{N}_5(\mathbf{0}, \mathbf{I})$. Примем модель

$$\mathbf{y} = \frac{1}{1 + \exp(-(\beta_0 + \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \mathbf{x}_3\beta_3))} + \varepsilon,$$

в качестве параметров $\boldsymbol{\beta}$ возьмем, например, вектор $(\beta_0, \beta_1, \beta_2, \beta_3)^T = (1, -2, 6, 3)^T$. В нашей модели признаки $\mathbf{x}_4$ и $\mathbf{x}_5$ являются шумовыми. Результатом работы алгоритма является последовательность весов признаков, выбираемых на каждом шаге. В данном случае алгоритм сделает шесть шагов.

В таблице (1) представлены результаты работы алгоритма. Первый столбец — номера признаков, первая строка — номер шага, а соответствующая ячейка таблицы — вес признака. Признаку с номером 0 соответствует константный признак. На рис. 2 показано сравнение оценок коэффициетов, полученных с помощью LALR и Forward Stagewise.

По полученным результатам можно сделать вывод, что последовательность выбираемых признаков и их весов согласуется с исходной моделью.

Данные «SAHD».

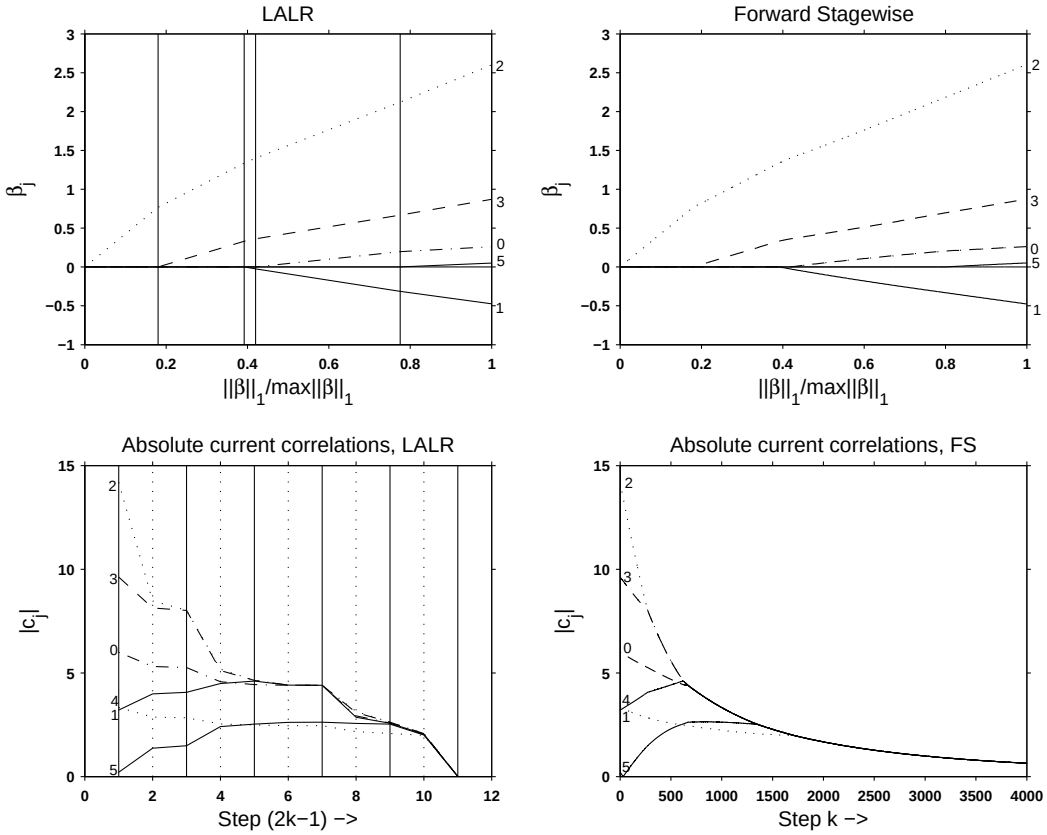

Рис. 2. Сравнение оценок коэффициентов для LALR и Forward Stagewise для модельных данных. Номера кривых соответствуют номерам признаков. Сплошные вертикальные линии обозначают шаги, а штриховые вертикальные — дополнительную итерацию для каждого шага.

Таблица 1. Результаты работы LALR

№	1	2	3	4	5	6
0	0	0	0	0.1969	0.2606	9.2868
1	0	0	-0.0250	-0.3142	-0.4733	-21.4689
2	0.7769	1.3359	1.4005	2.1215	2.5999	91.6048
3	0	0.3313	0.3615	0.6677	0.8713	36.5161
4	0	0	0	0	0	-8.2624
5	0	0	0	0	0.0513	1.6560

Проанализирована работа алгоритма на реальных данных «South African Heart Disease», см. [2]. Данные были впервые рассмотрены в [18]. Целью исследований являлось изучение факторов риска ишемических заболеваний сердца в районах с высокой заболеваемостью. Данные SAHD представляют собой сведения о физическом состоянии 462-х пациентов мужского пола белой расы возраста от 15 до 64 лет. Описание данных состоит из 9 признаков: $\mathbf{x}_1$ — sbp (sistolic blood pressure), $\mathbf{x}_2$ — tobacco, $\mathbf{x}_3$ — ldl (low-density lipoprotein), $\mathbf{x}_4$ — adiposity, $\mathbf{x}_5$ — famhist (family history), $\mathbf{x}_6$ — typea, $\mathbf{x}_7$ — obesity, $\mathbf{x}_8$ — alcohol, $\mathbf{x}_9$ — age; а также вектора меток класса chd: наличие — «1», или отсутствие — «0» инфаркта миокарда (MI) за время обследования. Перед использованием данные были стандартизованы согласно (1). На рис. 3 представлены результаты работы алгоритма.

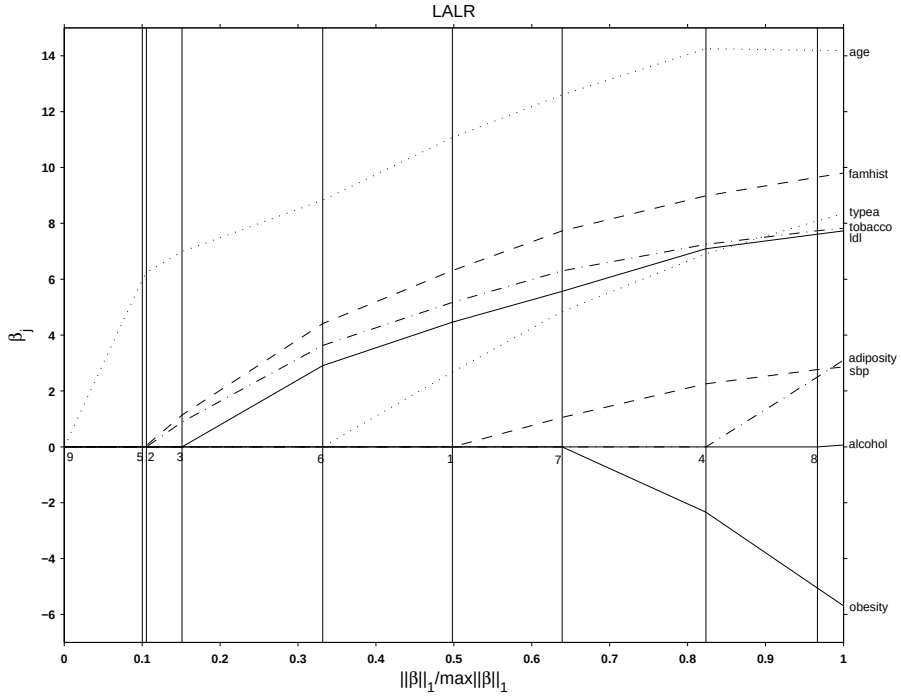

Рис. 3. Оценка коэффициентов алгоритма LALR для данных «South African Heart Disease». Вертикальные линии соответствуют шагам алгоритма.

Заключение

В данной работе предложен и исследован новый алгоритм LALR, решающий задачу отбора признаков в модели логистической регрессии. Разработан, исследован и математически обоснован алгоритм LALR, представляющий собой линеаризованный аналог алгоритма LARS для модели логистической регрессии. Проведена серия численных экспериментов на модельных и реальных данных «SAHD», результаты которых позволяют говорить об эффективности использования предложенного алгоритма.

Литература

[1] N. Draper and H. Smith. *Appied Regression Analysis*. Wiley, New York, 1966.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The ELemmaents of Statistical Learning: Data mining, Inference and Prediction*. Springer, New York, 2001.

[3] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

[4] R. R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[5] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal probLemmas. *Technometrics*, 12(1):55–67, 1970.

[6] A. N. Tikhonov. Regularization of incorrectly posed probLemmas. *SMD*, 4(3):1624–1627, 1963.

[7] A. Bjorkstrom. Ridge regression and inverse probLemmas. Technical report, Stockholm University, 2001.

[8] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[10] L. Lawson and R. J. Hanson. *Solving Least Squares ProbLemmas*. Englewood Cliffs: Prentice Hall, 1974.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[12] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[13] Ye Jianming. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, Mar 1998.

[14] D. Madigan and G. Ridgeway. Discussion of least angle regression. *Annals of Statistics*, 32(2):465–469, 2004.

[15] D. B. Rubin. Iteratively reweighted least squares. *Encyclopedia of statistical sciences*, 4:272–275, 1983.

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

[17] А. Ф. Измаилов. *Численные методы оптимизации*. Физматлит, Москва, 2005.

[18] J. Rousseauw, J. du Plessis, A. Benade, P. Jordan, J. Kotze, P. Jooste, and J. Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64:430–436, 1983.

Приложение к статье "Выбор признаков в задачах логистической регрессии"

Доказательство.[Леммы 1] Используя (20) и (23) разложим $\ell(\boldsymbol{\mu}_{\mathcal{A}_+})$ до первого члена,

$$\begin{aligned}\ell(\boldsymbol{\mu}_{\mathcal{A}_+}) = \ell(\boldsymbol{\mu}_{\mathcal{A}} + X_{\mathcal{A}}\boldsymbol{\gamma}) \approx \ell(\boldsymbol{\mu}_{\mathcal{A}}) + \sum_{j\in\mathcal{A}} \left(\frac{\partial \ell(\boldsymbol{\mu}_{\mathcal{A}_+})}{\partial \gamma_j}\right)_{\boldsymbol{\gamma}=\mathbf{0}} \gamma_j = \\ = \ell(\boldsymbol{\mu}_{\mathcal{A}}) + \sum_{j\in\mathcal{A}} s_j \mathbf{c}_j \gamma_j = \ell(\boldsymbol{\mu}_{\mathcal{A}}) + (\mathbf{s}_{\mathcal{A}} \circ \mathbf{c}_{\mathcal{A}})^T \boldsymbol{\gamma}.\end{aligned} \tag{31}$$

Далее, будем считать, что знак корреляции s_j не изменяется для любого $j \in \mathcal{A}$. Таким образом, используя выражение для корреляции (22), перепишем, раскрыв модуль, условия (27) в виде систем неравенств

$$(s_j \mathbf{x}_j - s_d \mathbf{x}_d)^T \left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})\right) \geqslant 0, \tag{32}$$

если $s_d \mathbf{x}_d^T \left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})\right) \geqslant 0$ и

$$(s_j \mathbf{x}_j + s_d \mathbf{x}_d)^T \left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})\right) \geqslant 0, \tag{33}$$

если $s_d \mathbf{x}_d^T \left(\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})\right) < 0$, для всех $j \in \mathcal{A}$ и $d \in \mathcal{A}^c$. Линеаризуем $\boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})$ при достаточно малых γ, используя (11)

$$\begin{aligned}\boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+}) = \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}} + X_{\mathcal{A}}\boldsymbol{\gamma}) \approx \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}}) + \sum_{j\in\mathcal{A}} \left(\frac{\partial \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}_+})}{\partial \gamma_j}\right)_{\boldsymbol{\gamma}=\mathbf{0}} \gamma_j = \\ = \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}}) + \sum_{j\in\mathcal{A}} W_{jj} s_j \mathbf{x}_j \gamma_j = \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}}) + W X_{\mathcal{A}} \boldsymbol{\gamma}.\end{aligned} \tag{34}$$

Таким образом, пользуясь (34), системы неравенств (32) и (33) перепишутся в следующем виде:

$$(s_j \mathbf{x}_j \pm s_d \mathbf{x}_d)^T W X_{\mathcal{A}} \boldsymbol{\gamma} \leqslant (s_j \mathbf{x}_j \pm s_d \mathbf{x}_d)^T (\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})), \tag{35}$$

для всех $j \in \mathcal{A}$ и $d \in \mathcal{A}^c$. Двойной знак «$\pm$» используется для компактной записи двух неравенств с «$+$» и «$-$». Как было введено ранее в (7), обозначим матрицы M_{d-} и M_{d+}:

$$M_{d\pm} = (\ \cdots \quad s_j \mathbf{x}_j \pm s_d \mathbf{x}_d \quad \cdots\ )_{j\in\mathcal{A}}, \tag{36}$$

для всех $d \in \mathcal{A}^c$. Тогда система (35) принимает следующий вид:

$$\left\{ \begin{array}{c} \vdots \\ M_{d\pm}^T W X_{\mathcal{A}} \boldsymbol{\gamma} \leqslant M_{d\pm}^T (\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})), \\ \vdots \end{array} \right. \tag{37}$$

для всех $d \in \mathcal{A}^c$. Обозначим матрицы A_{d-}, A_{d+} и векторы $\mathbf{b}_{d-}$, $\mathbf{b}_{d+}$

$$A_{d\pm} = M_{d\pm}^T W X_{\mathcal{A}},$$

$$\mathbf{b}_{d\pm} = M_{d\pm}^T (\mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}_{\mathcal{A}})),$$

тогда, используя (31) и (37), общая задача (26), (27) принимает линеаризованный вид (28). Что и требовалось доказать. ■

Доказательство.[Леммы 2] Докажем это от противного. Предположим, что векторы $(\mathbf{a}_1+$ $+\mathbf{a}_{k+1}), \ldots, (\mathbf{a}_k+\mathbf{a}_{k+1}), \mathbf{a}_{k+1}$ линейно зависимы. Тогда существуют $\gamma_1, \ldots, \gamma_{k+1}$ одновременно ненулевые, для которых справедливо

$$\gamma_1(\mathbf{a}_1 + \mathbf{a}_{k+1}) + \ldots + \gamma_k(\mathbf{a}_k + \mathbf{a}_{k+1}) + \gamma_{k+1}\mathbf{a}_{k+1} = 0.$$

Из этого следует, что

$$\gamma_1\mathbf{a}_1 + \ldots + \gamma_k\mathbf{a}_k + (\gamma_1 + \ldots + \gamma_k + \gamma_{k+1})\mathbf{a}_{k+1} = 0,$$

что противоречит линейной независимости векторов $\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}$. Поэтому векторы $(\mathbf{a}_1 + \mathbf{a}_{k+1}), \ldots, (\mathbf{a}_k + \mathbf{a}_{k+1}), \mathbf{a}_{k+1}$ линейно независимы. ■

Доказательство.[Леммы 3] Из линейной независимости векторов $\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}$ сразу следует, что матрица A имеет полный ранг по столбцам. Согласно лемме 2, матрица $(C|\mathbf{a}_{k+1})$ также имеет полный ранг по столбцам. Здесь знак $<< \mid >>$ обозначает присоединение вектора $\mathbf{a}_{k+1}$ к матрице C. Пользуясь свойствами ранга произвольной матрицы, получаем

$$\operatorname{rank}(A) = \operatorname{rank}(A^TA) = k \tag{38}$$

и аналогично

$$\operatorname{rank}\left((C|\mathbf{a}_{k+1})\right) = \operatorname{rank}\left((C|\mathbf{a}_{k+1})^T(C|\mathbf{a}_{k+1})\right) = k+1.$$

Из последнего заключаем, что матрица $(C|\mathbf{a}_{k+1})^T(C|\mathbf{a}_{k+1})$ является квадратной и полного ранга, а значит вектор-столбцы

$$(C|\mathbf{a}_{k+1})^T(\mathbf{a}_1 - \mathbf{a}_{k+1}), \ldots (C|\mathbf{a}_{k+1})^T(\mathbf{a}_k - \mathbf{a}_{k+1}), (C|\mathbf{a}_{k+1})^T\mathbf{a}_{k+1}$$

линейно независимы. Тогда, согласно лемме 2, векторы

$$(C|\mathbf{a}_{k+1})^T\mathbf{a}_1, \ldots, (C|\mathbf{a}_{k+1})^T\boldsymbol{a}_k$$

также линейно независимы. А из этого следует, что

$$\operatorname{rank}\left((C|\mathbf{a}_{k+1})^TA\right) = k.$$

Из последнего получаем, что матрица $A^T(C|\mathbf{a}_{k+1})$ имеет ровно k линейно независимых столбцов. Покажем, что при выполнении условия афинной независимости столбцов матрицы A^TA_+, вектор $A^T\mathbf{a}_{k+1}$ раскладывается в линейную комбинацию столбцов матрицы A^TC с ненулевыми коэффициентами.

Согласно (38), квадратная матрица A^TA имеет полный ранг, поэтому столбцы этой матрицы образуют базис в пространстве $\mathbb{R}^k$. А значит, любой ненулевой вектор этого пространства раскладывается по базису с ненулевыми коэффициентами единственным образом. Поэтому, для вектора $A^T\mathbf{a}_{k+1} \in \mathbb{R}^k$ существует и единственный ненулевой вектор $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)^T$, такой что

$$A^TA\boldsymbol{\lambda} = A^T\mathbf{a}_{k+1}. \tag{39}$$

Теперь, пусть выполнено условие афинной независимости столбцов матрицы A^TA_+. Это означает, что дополнительно выполняется условие

$$\eta = \sum_{i=1}^{k} \lambda_i \neq 1. \tag{40}$$

Покажем, что в этом случае вектор $A^T\mathbf{a}_{k+1}$ раскладывается по системе столбцов матрицы A^TC с коэффициентами

$$\boldsymbol{\xi} = \frac{1}{1-\eta}\boldsymbol{\lambda}. \tag{41}$$

Итак, пусть для некоторого $\boldsymbol{\xi}$ имеет место разложение

$$A^TC\boldsymbol{\xi} = A^T\mathbf{a}_{k+1}. \tag{42}$$

Преобразуем

$$A^TC = A^T(A - (\underbrace{\mathbf{a}_{k+1} \ \dots \ \mathbf{a}_{k+1}}_{k})) = A^TA - A^T\mathbf{a}_{k+1}\mathbf{1}^T.$$

Подставляя последнее в (42) получим $A^TA\boldsymbol{\xi} = A^T\mathbf{a}_{k+1}(1 + \mathbf{1}^T\boldsymbol{\xi})$. Делаем замену (39), $A^TA\boldsymbol{\xi} = A^TA\boldsymbol{\lambda}(1 + \mathbf{1}^T\boldsymbol{\xi})$. Откуда полагаем

$$\boldsymbol{\xi} = \boldsymbol{\lambda}(\underbrace{1 + \mathbf{1}^T\boldsymbol{\xi}}_{\text{число } \alpha}).$$

или подставляя $\boldsymbol{\xi} = \boldsymbol{\lambda}\alpha$ в предыдущее равенство, получим

$$\boldsymbol{\lambda}\alpha = \boldsymbol{\lambda}(1 + \underbrace{\mathbf{1}^T\boldsymbol{\lambda}}_{\eta}\alpha),$$

откуда $\alpha = \frac{1}{1-\eta}$. По условию (40) знаменатель дроби не обращается в 0. Таким образом, существует разложение (42) вектора $A^T\mathbf{a}_{k+1}$ по системе столбцов матрицы A^TC с ненулевыми коэффициентами (41). Поэтому

$$\operatorname{rank}(A^TC) = \operatorname{rank}\left(A^T(C|\mathbf{a}_{k+1})\right) = k,$$

т. е матрица A^TC имеет полный ранг.

Теперь, пусть столбцы матрицы A^TA_+ афинно зависимы. В этом случае, для разложения (39) дополнительно выполняется условие $\eta = \sum_{i=1}^{k}\lambda_i = 1$. Подставляя его в (39), получим

$$A^TA\boldsymbol{\lambda} = A^T\mathbf{a}_{k+1} = A^T\mathbf{a}_{k+1}\sum_{i=1}^{k}\lambda_i = A^T\mathbf{a}_{k+1}\mathbf{1}^T\boldsymbol{\lambda},$$

или, перенеся в одну сторону, имеем

$$A^T(A - \mathbf{a}_{k+1}\mathbf{1}^T)\boldsymbol{\lambda} = A^TC\boldsymbol{\lambda} = 0.$$

А полученная однородная система имеет нетривиальное решение $\boldsymbol{\lambda}$ тогда и только тогда, когда

$$\det(A^TC) = 0.$$

Таким образом, в этом случае матрица A^TC имеет неполный ранг. Что и требовалось доказать. ■

Доказательство.[Теоремы 4] Для доказательства воспользуемся методом математической индукции по шагам алгоритма.

1. База индукции. Возьмем в качестве базы первый шаг алгоритма, когда выбран первый активный признак $\mathbf{x}_i$. Матрицы $A_{d\pm}$ и векторы $\mathbf{b}_{d\pm}$ в этом случае представляют собой

действительные числа. Если решение ЗЛП существует, то оно достигается на границе области допустимых значений, т. е для некоторого d выполняется равенство $A_d\gamma^* = \mathbf{b}_d$, что эквивалентно выполнению условия (29). Здесь и далее, под $A_d\gamma^* = \mathbf{b}_d$ подразумевается выполнение $A_{d-}\gamma^* = \mathbf{b}_{d-}$, либо $A_{d+}\gamma^* = \mathbf{b}_{d+}$.

Покажем, что условие (30) также выполняется. Допустим, что это не так. Пусть $\gamma^* \in \Upsilon$ есть решение ЗЛП, тогда существует $d \in \mathcal{I}\backslash\{i\}$, для которого $A_d\gamma^* = \mathbf{b}_d$. Предположим, что существует $\tilde{\gamma} \in \Upsilon$, для которого выполнено $s_i c_i \tilde{\gamma} < s_i c_i \gamma^*$. Т. к. $\tilde{\gamma} \in \Upsilon$, то существует $\tilde{d} \in \mathcal{I}\backslash\{i\}$, для которого $A_{\tilde{d}}\tilde{\gamma} = \mathbf{b}_{\tilde{d}}$. Рассматриваются только значения $s_i c_i \gamma > 0$, которые соответствуют положительному приращению логарифма правдоподобия. А так как абсолютная корреляция $s_i c_i > 0$, то по предположению получаем, что $0 < \tilde{\gamma} < \gamma^*$. Из того, что вектор $\mathbf{x}_i$ имеет наибольшую абсолютную корреляцию с вектором остатков, следует $\mathbf{b}_{j\pm} = s_i c_i \pm s_j c_j > 0$, для $j \in \{d, \tilde{d}\}$. А это означает, что $A_{\tilde{d}} > 0$. Поэтому, учитывая, что γ^* есть решение, справедливо неравенство

$$\mathbf{b}_{\tilde{d}} = A_{\tilde{d}}\tilde{\gamma} < A_{\tilde{d}}\gamma^* \leqslant \mathbf{b}_{\tilde{d}}.$$

Получаем противоречие. Значит действительно, для данного шага условие (30) выполняется. Тем самым доказана справедливость теоремы для первого шага.

2. Допустим теперь, что утверждение верно для k-го шага алгоритма. Пусть $\boldsymbol{\gamma}^k$ есть решение ЗЛП на k-м шаге, тогда существует $d \in \mathcal{A}^c$, для которого верно $A_d\boldsymbol{\gamma}^k = \mathbf{b}_d$, причем для любого $j \in \mathcal{A}^c\backslash\{d\}$ справедливо $A_{j\pm}\boldsymbol{\gamma}^k < \mathbf{b}_{j\pm}$. Эти два условия есть не что иное, как линеаризованный вид условий (27). А это означает, что при линеаризации справедливо

$$\Big|\frac{d}{d\alpha}\ell(\boldsymbol{\mu}_{\mathcal{A}_+} + \mathbf{x}_i\alpha)\Big|_{\alpha=0} = \Big|\frac{d}{d\alpha}\ell(\boldsymbol{\mu}_{\mathcal{A}_+} + \mathbf{x}_d\alpha)\Big|_{\alpha=0}, \tag{43}$$

для любого $i \in \mathcal{A}$, и

$$\Big|\frac{d}{d\alpha}\ell(\boldsymbol{\mu}_{\mathcal{A}_+} + \mathbf{x}_i\alpha)\Big|_{\alpha=0} > \Big|\frac{d}{d\alpha}\ell(\boldsymbol{\mu}_{\mathcal{A}_+} + \mathbf{x}_j\alpha)\Big|_{\alpha=0}, \tag{44}$$

для любых $i \in \mathcal{A}$ и $j \in \mathcal{A}^c\backslash\{d\}$, где

$$\boldsymbol{\mu}_{\mathcal{A}_+} = \boldsymbol{\mu}_{\mathcal{A}} + X_{\mathcal{A}}\boldsymbol{\gamma}^k. \tag{45}$$

Теперь можно доказать утверждение для следующего шага.

3. Рассмотрим $(k+1)$-й шаг. В активное множество $\mathcal{A}$ добавился индекс d.

3.1 Сначала покажем что множество Υ не пусто. Для этого покажем существование обратных матриц A_d^{-1} для $d \in \mathcal{A}^c$. Покажем на примере матрицы A_{d-}. Условие (43), полученное на предыдущей итерации, означает, что при линеаризации для абсолютной корреляции справедливо $\mathbf{s}_{\mathcal{A}}\mathbf{c}_{\mathcal{A}} = c \cdot \mathbf{1}_{\mathcal{A}}$, где константа c есть значение абсолютной корреляции для активных признаков, а $\mathbf{1}_{\mathcal{A}}$ есть единичный вектор, размерности $|\mathcal{A}|$. Аналогично, условие (44) означает, что для любого $d \in \mathcal{A}^c$ верно $s_d c_d < c$. Далее, согласно (12), $A_{d-} = X_{\mathcal{A}}^T W X_{\mathcal{A}} - X_d^T W X_{\mathcal{A}}$ для любого $d \in \mathcal{A}^c$. Т. к признаки независимы, то матрица $X_{\mathcal{A}}^T W X_{\mathcal{A}}$ имеет полный ранг. Поэтому существует и единственный $\boldsymbol{\lambda} \neq \mathbf{0}$, для которого выполнено

$$X_{\mathcal{A}}^T W X_{\mathcal{A}}\boldsymbol{\lambda} = \mathbf{1}_{\mathcal{A}}, \tag{46}$$

причем, т. к $X_d = (s_d\mathbf{x}_d \ldots s_d\mathbf{x}_d)$, то

$$X_d^T W X_{\mathcal{A}}\boldsymbol{\lambda} = \tau \cdot \mathbf{1}_{\mathcal{A}}. \tag{47}$$

Откуда заключаем два важных результата.

3.1.1 Если для некоторого $d \in \mathcal{A}^c$ справедливо $\tau \neq 1$, то из (46) и (47) сразу следует, что матрица

$$\begin{pmatrix} X_{\mathcal{A}}^T W X_{\mathcal{A}} & \mathbf{1}_{\mathcal{A}} \\ s_d \mathbf{x}_d^T W X_{\mathcal{A}} & 1 \end{pmatrix}^T$$

имеет полный ранг. А это, в свою очередь, эквивалентно тому, что векторы

$$\{\ldots, X_{\mathcal{A}}^T W s_j \mathbf{x}_j, \ldots\}_{j \in \mathcal{A}} \quad \text{и} \quad X_{\mathcal{A}}^T W s_d \mathbf{x}_d$$

являются афинно независимыми. Теперь, применяя лемму 3 для векторов

$$\{\ldots, W^{\frac{1}{2}} s_j \mathbf{x}_j, \ldots\}_{j \in \mathcal{A}} \quad \text{и} \quad W^{\frac{1}{2}} s_d \mathbf{x}_d$$

получим, что матрица A_{d-} имеет полный ранг, а значит, для нее существует обратная. Для матрицы A_{d+} аналогичное условие $\tau \neq -1$.

3.1.2 Если для некоторого $d \in \mathcal{A}^c$ справедливо $\tau = 1$, то по лемме 3 получаем, что обратной матрицы A_{d-}^{-1} не существует, но в тоже время по лемме 3 существует обратная матрица A_{d+}^{-1}. Аналогично и для $\tau = -1$.

Тем самым показано, что множество Υ не пусто.

3.2 Покажем теперь выполнимость условия (29). Если решение ЗЛП существует, то оно достигается на границе области допустимых значений. Таким образом, решением ЗЛП является решение некоторой подсистемы ограничений максимального ранга. Докажем от противного, что множество Υ содержит решение.

Пусть $\boldsymbol{\gamma}$ есть решение ЗЛП, причем $\boldsymbol{\gamma}$ является решением некоторой подсистемы, отличной от $A_d \boldsymbol{\gamma} \leqslant \mathbf{b}_d$, $d \in \mathcal{A}^c$. Тогда для некоторых различных подсистем с индексами $d, d' \in \mathcal{A}^c$, $d \neq d'$, существуют строки с индексами $i, i' \in \mathcal{A}$, $i \neq i'$, для которых выполнено

$$\begin{cases} (\mathbf{x}_i \pm \mathbf{x}_d)^T W X_{\mathcal{A}} \boldsymbol{\gamma} = (c \pm s_d c_d) \cdot \mathbf{1}_{\mathcal{A}}, \\ (\mathbf{x}_{i'} \pm \mathbf{x}_{d'})^T W X_{\mathcal{A}} \boldsymbol{\gamma} = (c \pm s_{d'} c_{d'}) \cdot \mathbf{1}_{\mathcal{A}}. \end{cases}$$

3.2.1 Если справедливо

$$(s_i \mathbf{x}_i \pm s_{d'} \mathbf{x}_{d'})^T W X_{\mathcal{A}} \boldsymbol{\gamma} > (c \pm s_{d'} c_{d'}) \cdot \mathbf{1}_{\mathcal{A}},$$

то ограничение не выполнено и это противоречит тому, что $\boldsymbol{\gamma}$ решение ЗЛП.

3.2.2 Если справедливо

$$(s_i \mathbf{x}_i \pm s_{d'} \mathbf{x}_{d'})^T W X_{\mathcal{A}} \boldsymbol{\gamma} < (c \pm s_{d'} c_{d'}) \cdot \mathbf{1}_{\mathcal{A}},$$

то найдем

$$\begin{aligned} (s_{i'} \mathbf{x}_{i'} \pm s_d \mathbf{x}_d)^T W X_{\mathcal{A}} \boldsymbol{\gamma} &= (\mp s_{d'} \mathbf{x}_{d'}^T W X_{\mathcal{A}} \boldsymbol{\gamma} + (c \pm s_{d'} c_{d'}) \cdot \mathbf{1}_{\mathcal{A}}) - (\mp \mathbf{x}_i^T W X_{\mathcal{A}} \boldsymbol{\gamma} - (c \pm s_d c_d) \cdot \mathbf{1}_{\mathcal{A}}) \\ &= \pm (s_i \mathbf{x}_i - s_{d'} \mathbf{x}_{d'})^T W X_{\mathcal{A}} \boldsymbol{\gamma} + (c \pm s_{d'} c_{d'}) \cdot \mathbf{1}_{\mathcal{A}} + (c \pm s_d c_d) \cdot \mathbf{1}_{\mathcal{A}} \\ &> (c \pm s_d c_d) \cdot \mathbf{1}_{\mathcal{A}}. \end{aligned}$$

Ограничение не выполнено и это противоречит тому, что $\boldsymbol{\gamma}$ решение ЗЛП.

3.2.3 Последний случай, если справедливо

$$(s_i \mathbf{x}_i \pm s_{d'} \mathbf{x}_{d'})^T W X_{\mathcal{A}} \boldsymbol{\gamma} = (c \pm s_{d'} c_{d'}) \cdot \mathbf{1}_{\mathcal{A}},$$

то это означает, что для подсистем $d, d' \in \mathcal{A}^c, d \neq d'$ множество активных индексов, в которых справедливы ограничения-равенства, совпадают. Поэтому если рассмотреть некоторые k строк для подсистемы d в которых достигается равенство, то равенство будет достигаться в соответствующих строках для подсистемы d'. Причем разность любых двух строк для d будет равна разности соответствующих строк для d'. Отсюда можно сделать вывод, что ранг рассматриваемой матрицы равенств не больше $k+1$. Поэтому требуется, чтобы равенства достигались в каждой строке матриц A_d и $A_{d'}$, т. е на двух подсистемах сразу. В этом случае в активное множество придется добавлять сразу два индекса $\{d, d'\}$.

Отсюда следует выполнение (29).

3.3 Покажем выполнимость условия (30). Докажем это от противного.

Пусть $\boldsymbol{\gamma}_z \in \Upsilon$ есть решение ЗЛП. Индекс z соответствует номеру выбранного признака. Предположим, что условие (30) не выполняется. Тогда существует $\boldsymbol{\gamma}_d \in \Upsilon$, для которого верно $0 < (\mathbf{s}_\mathcal{A} \circ \mathbf{c}_\mathcal{A})^T \boldsymbol{\gamma}_d < (\mathbf{s}_\mathcal{A} \circ \mathbf{c}_\mathcal{A})^T \boldsymbol{\gamma}_z$. При линеаризации из последнего следует

$$0 < (\mathbf{s}_\mathcal{A} \circ \mathbf{c}_\mathcal{A})^T (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) = c \cdot \mathbf{1}_\mathcal{A}^T (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d), \tag{48}$$

Из того, что $\boldsymbol{\gamma}_d \in \Upsilon$, получаем, что $A_d \boldsymbol{\gamma}_d = \mathbf{b}_d$. Будем использовать двойной знак, чтобы учесть возможные случаи. Итак, в матричном виде при линеаризации системы запишутся в виде $(X_\mathcal{A} \pm X_d)^T W X_\mathcal{A} \boldsymbol{\gamma}_d = (c \pm s_d c_d) \cdot \mathbf{1}_\mathcal{A}$. По предположению $\boldsymbol{\gamma}_z$ есть решение, поэтому справедливо $(X_\mathcal{A} \pm X_d)^T W X_\mathcal{A} \boldsymbol{\gamma}_z < (c \pm s_d c_d) \cdot \mathbf{1}_\mathcal{A}$. Отнимем от второго первое, получим

$$(X_\mathcal{A} \pm X_d)^T W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) < 0. \tag{49}$$

Нетрудно показать, что для выбранных $\boldsymbol{\gamma}_z$ и $\boldsymbol{\gamma}_d$ справедливо $X_\mathcal{A} W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) = \alpha \cdot \mathbf{1}_\mathcal{A}$. Покажем, что $\alpha > 0$. Т. к матрица $X_\mathcal{A} W X_\mathcal{A}$ положительно определена, то по определению

$$0 < (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d)^T X_\mathcal{A} W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) = \alpha \cdot \mathbf{1}_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d).$$

И, пользуясь (48), получаем $\alpha > 0$.

Согласно (46), $\boldsymbol{\lambda} = \frac{\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d}{\alpha}$. Следовательно, из (47) получаем $X_d W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) = \alpha \cdot \tau \cdot \mathbf{1}_\mathcal{A}$, причем мы рассматриваем только те ограничения, для которых справедливо: $\tau < 1$ для матрицы A_{d-}, и $\tau > -1$ для матрицы A_{d+}, т. к в противном случае $\boldsymbol{\gamma}_z$ и $\boldsymbol{\gamma}_d$ не являются решениями. В итоге, получаем

$$\begin{cases} (X_\mathcal{A} - X_d)^T W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) = \alpha \cdot (1 - \tau) \cdot \mathbf{1}_\mathcal{A} > 0, & \tau < 1; \\ (X_\mathcal{A} + X_d)^T W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) = \alpha \cdot (1 + \tau) \cdot \mathbf{1}_\mathcal{A} > 0, & \tau > -1. \end{cases}$$

или просто

$$(X_\mathcal{A} \pm X_d)^T W X_\mathcal{A} (\boldsymbol{\gamma}_z - \boldsymbol{\gamma}_d) > 0, \tag{50}$$

что противоречит (49). Значит действительно, для данного шага условие (30) выполняется.

Таким образом, утверждение справедливо и для $(k+1)$-го шага, а значит, согласно методу математической индукции, утверждение справедливо для любого шага. Что и требовалось доказать. ■

Оценка параметров смеси распределений

К. В. Павлов
kirill.pavlov@phystech.edu
Московский физико-технический институт

В работе рассматриваются способы построения смеси моделей и экспертов. Предлагается EM-алгоритм для совместного нахождения параметров моделей и их весов в смеси, а так же для нахождения параметров смеси обобщенных линейных моделей.

Ключевые слова: *смеси моделей, обобщенно-линейные модели, смеси экспертов.*

Введение

При решении задачи анализа данных строится модель — отображение известных характеристик объекта в неизвестные. Часто оказывается, что качество алгоритма можно улучшить с помощью комбинирования нескольких моделей [3, p. 653–676]. Например, можно обучить l моделей и в качестве ответа выводить усредненный ответ по всем моделям. Подобные комбинации моделей называются комитетами. Один из наиболее важных случаев комитета является бустинг. Алгоритмы в комитет добавляются последовательно и их параметры зависят от уже созданного на момент добавления комитета. Другим важным частным случаем комитета является смесь экспертов. В этом случае ответы алгоритмов взвешиваются в зависимости от области пространства, в которой находится объект. Рассмотрим способы построения композиций.

Общий подход к оценке параметров моделей

В случае, когда одной модели для описания данных не хватает, используют смеси моделей. Предполагается, что исходная зависимость $p(\boldsymbol{y} \mid \boldsymbol{x})$ выражается как композиция моделей $p(y \mid \boldsymbol{x}, \boldsymbol{w}_k)$ формулой:

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{k=1}^{l} p(\boldsymbol{w}_k \mid \boldsymbol{x}) p(y \mid \boldsymbol{x}, \boldsymbol{w}_k) = \sum_{k=1}^{l} \pi_k p(y \mid \boldsymbol{x}, \boldsymbol{w}_k), \tag{1}$$

где $\pi_k = p(\boldsymbol{w}_k \mid \boldsymbol{x})$ — вероятность принадлежности к модели k. На π_k накладываются условия нормировки: вероятность каждой модели неотрицательна и сумма вероятностей равна единице.

$$\sum_{k=1}^{l} \pi_k = 1, \quad \pi_k \geqslant 0 \quad \forall k. \tag{2}$$

Далее предполагается, что объекты в выборке независимы и плотность совместного распределения преобразуется в произведение плотностей распределения каждого объекта.

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{k=1}^{l} \pi_k \prod_{i=1}^{n} p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}_k) = \prod_{i=1}^{n} \sum_{k=1}^{l} \pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}_k). \tag{3}$$

В формуле 3 произведена смена порядка суммирования перемножения. Используя принцип максимума правдоподобия, будет максимизировать $p(\boldsymbol{y} \mid \boldsymbol{x})$. Проще это делать, введя функцию правдоподобия $Q(\boldsymbol{w_1}, \ldots, \boldsymbol{w_l}, \boldsymbol{\pi})$ как логарифм плотности вероятности данных.