Pavlov K.V., Strijov V.V. Multilevel model selection in the bank credit scoring applications // Proceedings of Mathematical Methods of Pattern Recognition. МАКС~Пресс, 2011 : 158-161.

Выбор многоуровневых моделей в задачах банковскового кредитного скоринга*

Павлов К. В., Стрижов В. В.
kirill.pavlov@phystech.edu, strijov@ccas.ru
Московский физико-технический институт, Вычислительный центр Дородницина РАН, Москва, Россия.

Решается задача классификации с использованием логистической регрессии. Предлагается новый подход, заключающийся в совместной кластеризации объектов и выборе признаков моделей. Результатом подхода является многоуровневая модель — набор моделей оптимальной сложности. Для построения моделей предлагается использовать EM-алгоритм. На E-шаге происходит отнесение объектов к моделям. На M-шаге происходит выбор наиболее вероятных параметров модели. Алгоритм тестировался на данных кредитным займам наличными.

Введение

Данная работа посвящена проблеме выбора и настройки моделей логистической регрессии в задачах классификации. Авторы предлагают новый подход, заключающийся в совместной кластеризации объектов и выборе признаков многоуровневых моделей. Его результатом является набор моделей оптимальной сложности.

Известные подходы к выбору моделей заключаются в использовании шаговой регрессии с критерием Маллоуза [6], итеративного перевзвешивающего метода наименьших квадратов [2], порождения нелинейных регрессионных моделей [5].

Для построения моделей предлагается использовать EM-алгоритм [2]. На E-шаге происходит отнесение объектов к моделям на основе оценки правдоподобия многоуровневой модели. На M-шаге происходит выбор наиболее вероятных параметров модели по объектам, которые к ней отнесли.

Преимуществом данного подхода является его способность описывать принципиально многомодельные выборки и сегментировать объекты в соответствии с используемыми моделями. Алгоритм тестировался на модельных и реальных данных по кредитному займу наличными. Эксперименты показали преимущество использования многоуровневых моделей по сравнению с использованием одной модели.

Постановка задачи

Рассмотрим задачу восстановления регрессии

$$\mathsf{E}(y \,|\, \boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{w}), \tag{1}$$

в которой измеряемые данные представляют собой пары значений зависимой переменной y и независимой переменной $\boldsymbol{x}$. Зависимость f является функцией регрессии от независимой переменной $\boldsymbol{x}$, называемой регрессором, и вектора параметров $\boldsymbol{w}$. Задачей регрессионного анализа является нахождение функции f и параметров $\boldsymbol{w}$.

Определение 1. *Регрессионная выборка* $D = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^m$ *— множество* m *пар, состоящих из векторов* $\boldsymbol{x}_i = (x_1, \ldots, x_n)^\intercal$ *и соответствующих этим векторам значений* y^i.

Далее предполагается, что переменные определены на подмножестве действительных чисел: $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n,\ y \in \mathcal{Y} \subseteq \mathbb{R}$. Индексы элементов i и компонент вектора независимой переменной j являются элементами конечных множеств $i \in \mathcal{I} = \{1, \ldots, m\},\ j \in \mathcal{J} = \{1, \ldots, n\}$.

Определение 2. *Матрица плана* X *— матрица, строки которой есть компоненты независимой переменной* $\boldsymbol{x}$, $X = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m)^\intercal$.

Регрессионную выборку, определенную в (1) будем обозначать $D = (X, \boldsymbol{y})$. Выборка может быть как функцией дискретного аргумента, так и отношением, при этом одному значению переменной $\boldsymbol{x}$ может соответствовать несколько значений переменной y. Для нахождения функции регрессии используется понятие регрессионной модели.

Определение 3. *Регрессионная модель — параметрическое семейство функций, отображающих декартово произведение областей определения объектов* $\mathcal{X}$ *и параметров* $\mathcal{W}$ *в область значений* $\mathcal{Y}$ *зависимой переменной*

$$f \colon \mathcal{X} \times \mathcal{W} \to \mathcal{Y}.$$

Определение 4. *Многоуровневой регрессионной моделью называется набор регрессионных моделей* f_k, $k = 1, \ldots, l$ *такой, что при разбиении множества индексов объектов* $\mathcal{I} = \sqcup \mathcal{I}_k$ *для всех объектов с индексами из* $\mathcal{I}_k$ *используется модель* f_k.

Ниже будут исследованы обобщенные линейные модели.

Обобщенные линейные модели

Впервые обобщенные линейные модели были введены Джоном Нельдером и Робертом Веддербурном в 1972г. [4]. В основе обобщенных линейных моделей лежат следующие предположения. Во-первых, считается, что зависимая переменная $\boldsymbol{y}$ имеет экспонентную плотность распределения [1] с вектором параметров $\boldsymbol{\theta}$,

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = h(\boldsymbol{y}) g(\boldsymbol{\theta}) \exp\big(\boldsymbol{T}(\boldsymbol{y})^\intercal \boldsymbol{\eta}(\boldsymbol{\theta})\big), \tag{2}$$

Работа выполнена при поддержке РФФИ: 10-07-00422

где h, g, $\boldsymbol{T}$ и $\boldsymbol{\eta}$ — известные функции. Оказывается [1], что в случае экспонентного распределения и только в нем, $\boldsymbol{T}(\boldsymbol{y})$ является достаточной статистикой. Для удобства дальнейшей записи перепишем функцию плотности в следующем виде

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \exp\big(\boldsymbol{T}(\boldsymbol{y})^{\intercal}\boldsymbol{\eta}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + c(\boldsymbol{y})\big). \quad (3)$$

Второе предположение заключается в том, что предиктор $\boldsymbol{\eta}$ линеен по координатам независимой переменной $\boldsymbol{x}$.

$$\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta}) = X\boldsymbol{w}. \quad (4)$$

Предполагается так же, что математическое ожидание $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ зависимой переменной $\boldsymbol{y}$ есть монотонная функция вектора $\boldsymbol{\eta}$ [3]. При этом регрессионная модель имеет вид

$$\mathsf{E}(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \boldsymbol{\mu} = f(\boldsymbol{\eta}) = f(X\boldsymbol{w}). \quad (5)$$

Функция f называется функцией активации. В силу её монотонности существует обратная функция f^{-1}, которая называется функцией связи [2].

В частном случае экспонентного распределения $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. При этом говорят [3], что распределение имеет каноническую форму. Функция плотности при этом

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \exp\big(\boldsymbol{T}(\boldsymbol{y})^{\intercal}\boldsymbol{\theta} - b(\boldsymbol{\theta}) + c(\boldsymbol{y})\big). \quad (6)$$

Для случая канонической формы можно выписать выражения математического ожидания и дисперсии достаточной статистики зависимой величины

$$\mathsf{E}\big(\boldsymbol{T}(\boldsymbol{y})\big) = \boldsymbol{\mu} = \nabla b(\boldsymbol{\theta}); \quad \mathsf{D}\big(\boldsymbol{T}(\boldsymbol{y})\big) = \nabla\nabla^{\intercal} b(\boldsymbol{\theta}). \quad (7)$$

В двухклассовой задаче классификации переменная y принимает два значения, $y \in \{0, 1\}$. Предположим, что зависимая величина принадлежит распределению Бернулли. Ниже рассмотрим этот случай.

Распределение Бернулли

Пусть случайная величина имеет распределение Бернулли с параметром p, $y \sim B(p)$, тогда

$$y = \begin{cases} 1, & p\,; \\ 0, & 1-p. \end{cases} \quad (8)$$

Покажем, что распределение Бернулли есть частный случай экспонентного распределения (6). Функция плотности $p(y\,|\,p)$ имеет вид

$$p(y\,|\,p) = p^y(1-p)^{1-y}. \quad (9)$$

Логарифмируя плотность получим функцию правдоподобия

$$l(y\,|\,p) = y\log p + (1-y)\log(1-p). \quad (10)$$

Сгруппируем члены

$$l(y\,|\,p) = y\log\frac{p}{1-p} + \log(1-p). \quad (11)$$

Полученное выражение имеет форму экспонентного семейства (6) для случая $\boldsymbol{T}(\boldsymbol{y}) = y$

$$\log p(y\,|\,p) = y\theta - b(\theta) + c(y) \quad (12)$$

со следующим соответствием: из вида первого слагаемового получим, что каноническй параметр соответствует логистической функции p:

$$\theta = \log\frac{p}{1-p}. \quad (13)$$

Решая данное уравнение получим, что

$$p = \frac{e^\theta}{1+e^\theta} = \sigma(\theta), \quad 1-p = \frac{1}{1+e^\theta} = \sigma(-\theta). \quad (14)$$

Во втором слагаемом

$$\log(1-p) = \log\Big(\frac{1}{1+e^\theta}\Big) = -\log(1+e^\theta),$$

откуда определяется функция $b(\theta)$:

$$b(\theta) = \log(1+e^\theta). \quad (15)$$

В случае распределения Бернулли $c(y) = 0$.

Проверим значения математического ожидания и дисперсии. Дифференцируя $b(\theta)$ получим

$$\mathsf{E}(y) = b^{'}(\theta) = \frac{e^\theta}{1+e^\theta} = p. \quad (16)$$

Вторая производная даст дисперсию

$$\mathsf{D}(y) = b^{''}(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = p(1-p). \quad (17)$$

Для подбора параметров модели воспользуемся итеративным перевзвешивающим методом наименьших квадратов (IRLS).

Оценка правдоподобия модели

Рассмотрим распределение бернуллиевского случайного вектора $\boldsymbol{y}$ с независимыми компонентами $y_i \sim B(p_i)$. В рамках обобщенных линейных моделей натуральный параметр θ представляется как

$$\theta = \sum_{j=1}^{n} x_j w_j = \boldsymbol{x}^{\intercal}\boldsymbol{w}. \quad (18)$$

Функция плотности вектора $\boldsymbol{y}$ имеет вид

$$p(\boldsymbol{y}\,|\,\boldsymbol{w}) = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}. \quad (19)$$

Определим функцию штрафа как минус логарифм правдоподобия

$$E(\boldsymbol{w}) = -\ln p(\boldsymbol{y} \mid \boldsymbol{w}) = \\ = -\sum_{i=1}^{m} y_i \ln p_i + (1-y_i)\ln(1-p_i). \tag{20}$$

Используя тождество

$$\frac{d\sigma(\theta)}{d\theta} = \sigma(1-\sigma) \tag{21}$$

и то, что $p = \sigma(\boldsymbol{x}^\intercal \boldsymbol{w})$, вычислим градиент функции штрафа.

$$\nabla E(\boldsymbol{w}) = -\sum_{i=1}^{m} \big(y_i(1-\sigma_i) - (1-y_i)\sigma_i\big)\boldsymbol{x}_i = \\ = \sum_{i=1}^{m} (\sigma_i - y_i)\boldsymbol{x}_i = X^\intercal(\boldsymbol{\sigma} - \boldsymbol{y}), \tag{22}$$

где $\sigma_i = \sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w})$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)^\intercal$.

Для подбора параметров $\boldsymbol{w}$ модели воспользуемся методом Ньютона-Рафсона, который на каждой итерации вычисляет квадратичную аппроксимацию функции, используя её градиент и гессиан. Формула обновления весов

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} - H^{-1}(\boldsymbol{w}^{\text{old}})\nabla E(\boldsymbol{w}^{\text{old}}). \tag{23}$$

Гессиан функции штрафа

$$H(\boldsymbol{w}) = \\ = \nabla\nabla^\intercal E(\boldsymbol{w}) = \sum_{i=1}^{m} \sigma_i(1-\sigma_i)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal = X^T \Sigma X, \tag{24}$$

где введено обозначение Σ — диагональная матрица, $\Sigma_{ii} = \sigma_i(1-\sigma_i)$. Используя (17) заметим, что $\Sigma_{ii} = \mathsf{D}y_i$, а так как компоненты вектора $\boldsymbol{y}$ по предположению независимы, то Σ является корреляционной матрицей.

Из свойств сигмоидной функции $\Sigma_{ii} > 0$, ковариационная матрица положительно определена, а значит и гессиан положительно определён (он является матрицей Грама в пространстве весов) из чего следует, что функция $E(\boldsymbol{w})$ выпукла и имеет единственный минимум.

Формула Ньютона–Рафсона для обновления весов для модели логистической регрессии

$$\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} - (X^T\Sigma X)^{-1}X^\intercal(\boldsymbol{\sigma} - \boldsymbol{y}) = \\ = (X^T\Sigma X)^{-1}X^T\Sigma(X\boldsymbol{w}^{\text{old}} - \Sigma^{-1}(\boldsymbol{\sigma} - \boldsymbol{y})) = \\ = (X^T\Sigma X)^{-1}X^T\Sigma\boldsymbol{z}; \tag{25}$$

$$\boldsymbol{z} = X\boldsymbol{w}^{\text{old}} - \Sigma^{-1}(\boldsymbol{\sigma} - \boldsymbol{y}). \tag{26}$$

Процедура выбора модели

Для подбора многоуровневых моделей при решении задачи классификации для объекта нужно выбрать соответствующую ему модель. Это можно сделать на основе правдоподобия этой модели. Вероятность того, что объект $(\boldsymbol{x}_i, y_i)$ был порожден моделью f_k

$$p(f_k \mid \boldsymbol{x}_i, y_i) = \frac{p(f_k, \boldsymbol{x}_i, y_i)}{p(\boldsymbol{x}_i, y_i)} = \frac{p(y_i \mid f_k, \boldsymbol{x}_i)p(f_k, \boldsymbol{x}_i)}{p(\boldsymbol{x}_i, y_i)}. \tag{27}$$

Априорная вероятность объекта $p(\boldsymbol{x}_i, y_i)$ одинакова для всех моделей. Величина $p(f_k, \boldsymbol{x}_i)$ называется априорной вероятностью модели. Предположим, что заранее нет никаких предпочтений в выборе моделей и априорные вероятности их равны. Если мы предполагаем, что объект относится к наиболее вероятной модели, то принцип максимума правдоподобия модели можно представить в виде задачи оптимизации

$$k^* = \arg\max_{k \in \{1..l\}} p(y_i \mid f_k, \boldsymbol{x}_i). \tag{28}$$

Класс объекта неизвестен, в этом случае будем рассматривать наихудший вариант: объект имеет класс, доставляющий минимум $p(y_i \mid f_k, \boldsymbol{x}_i)$:

$$k^* = \arg\max_k \min_{y_i} p(y_i \mid f_k, \boldsymbol{x}_i). \tag{29}$$

Вероятности принадлежности объектов к классам выражаются через логистическую функцию (14), перепишем решающее правило для выбора модели

$$k^* = \arg\max_k \min\{\sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w}_k), \sigma(-\boldsymbol{x}_i^\intercal \boldsymbol{w}_k)\}. \tag{30}$$

Преобразуем выражение

$$k^* = \arg\max_k \sigma(-|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|) = \\ = \arg\min_k \sigma(|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|) = \arg\min_k |\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|. \tag{31}$$

Заметим, что $\frac{|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|}{|\boldsymbol{w}_k|}$ есть расстояние от $\boldsymbol{x}_i$ до гиперплоскости с нормальным вектором $\boldsymbol{w}_k$. Объекты относятся к той модели, расстояние до разделяющей гиперплоскости которой минимально с точностью до модуля нормального вектора. Ввиду того, что минимизация $|\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|$ эквивалентна максимизации $\sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w}_k)\big(1 - \sigma(\boldsymbol{x}_i^\intercal \boldsymbol{w}_k)\big) = \mathsf{D}(y_i \mid w_k)$, решающее правило (31) можно интерпретировать так: объекты относятся к модели, дисперсия относительно которой максимальна.

Перейдем к способу построения многоуровневых моделей.

Алгоритм выбора модели

Для построения моделей используется EM-алгоритм со следующими шагами:

M-step. На M-шаге настраиваются параметры моделей с помощью логистической регрессии и метода Ньютона–Рафсона (IRLS).

Алгоритм 1. EM-алгоритм для l моделей.

Вход: $X = \{\boldsymbol{x}_i^\intercal\}_{i=1}^m$ — матрица плана;
$\boldsymbol{y} = \{y_i\}_{i=1}^m$ — метки классов;
l — число моделей;
Выход: Набор моделей $(\text{model}_k)_{k=1}^l$;

```
EmIrls(X, Y, l):
Инициализировать модели объектов случайно;
повторять
    M-step:
    для k = 1..l
        Оценить параметры k-ой модели;
        X^k — объекты, отнесенные к k-ой модели;
        y^k — классы объектов k-й модели;
        w_k = IRLS(X^k, y^k);
    E-step:
    для всех i = 1..m
        model(x_i) = arg min_k |x_i^T w_k|;
пока модели не стабилизируются.
```

E-step. На E-шаге происходит отнесение объекта к моделям на основании их правдоподобия. Решающее правило имеет вид

$$k^* = \arg\min_k |\boldsymbol{x}_i^\intercal \boldsymbol{w}_k|. \tag{32}$$

Численный эксперимент

Алгоритм тестировался на модельных и реальных данных. Модельные данные представляли собой два кластера в каждом из которых объекты разных классов распределены нормально и линейно разделимы, однако сама выборка линейно разделимой не является. Алгоритм выявил наличие двух моделей и безошибочно классифицировал объекты гиперплоскостями, рис. 1.

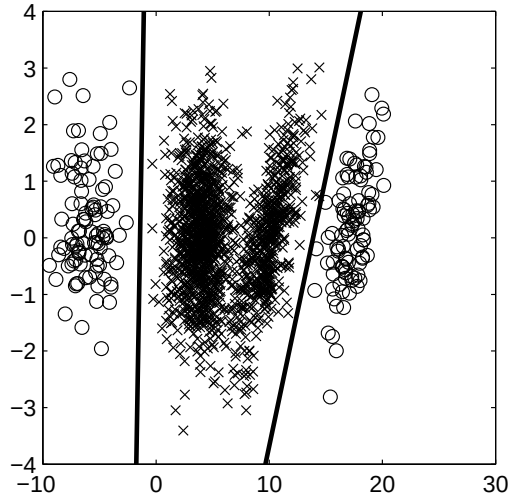

Рис. 1. Классификация модельной выборки.

Реальные данные представляли собой кредитные истории займа наличными. Выборка содержала данные о шести тысячах клиентах, каждый из которых описывался пятидесятью признаками. В качестве функции качества использовалась площадь под ROC-кривой. Метод сравнивался с логистической регрессией с градиентным методом настройки весов и итеративным перевзвешивающим методом наименьших квадратов, рис. 2.

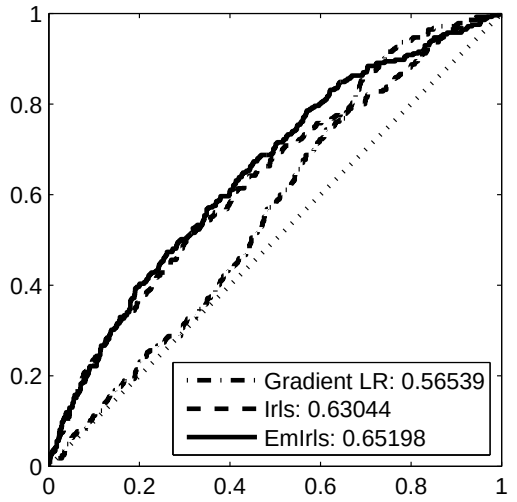

Рис. 2. ROC кривые и значения площади под кривой для различных моделей.

Заключение

В работе был предложен алгоритм выбора многоуровневых моделей; его работа проиллюстрирована на реальных и синтетических данных. Так же в работе показано преимущество использования многоуровневых моделей по сравнению с использованием одной модели на примере классификации линейно неразделимой выборки.

Литература

[1] *Ивченко Г., Медведев Ю.* Введение в математическую статистику. — Издательство ЛКИ, 2010. — P. 600.

[2] *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.

[3] *Lee Y., Nelder J. A., Pawitan Y.* Generalized Linear Models with Random Effects. — Taylor and Francis Group, LLC, 2006. — P. 396.

[4] *Nelder J., Wedderburn R.* Generalized linear models // Journal of the Royal Statistical Society. — 1972. — Pp. 370–384. — Series A (General) (Blackwell Publishing).

[5] *Strijov V., Weber G. W.* Nonlinear regression model generation using hyperparameter optimization // Computers and Mathematics with Applications. — 2010. — Vol. 60, no. 4. — Pp. 981–988.

[6] *Tibshirani R. J.* Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society. Series B (Methodological). — 1996. — Vol. 58, no. 1. — Pp. 267–288.