

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

Митяшов А. А.

Московский физико-технический институт

Научный руководитель В. В. Стрижов

Задача

В работе рассматривается построение модели логистической регрессии для банковского и менеджмент-скоринга.

Предлагается построить модель, которая отбирает признаки, учитывает экспертное мнение и повышает интерпретируемость полученных результатов.

Предложения по решению задачи

- 1 Для учета экспертных оценок предлагается ввести *предметно-экспертные* ограничения.
- 2 Для отбора признаков предлагается использовать штрафную функцию *elastic-net*.

Здесь и далее коэффициенты признаков обозначаются как β

Ограничение	Пример использования
$\beta_j \geq 0$	Экспертно предполагается, что с увеличением признака (количественного) увеличиваются риски.
$\beta_{j+1} \geq \beta_j$	Аналогично предыдущему ограничению, если количественный признак градуировать.
$\ \beta_A\ \leq t$, β_A – это подвектор β .	Позволяет снизить эффект переобучения, ограничив веса определенных признаков.
$C\beta = \mathbf{b}$ $C \in \mathbb{R}^{n \times p}$ $\mathbf{b} \in \mathbb{R}^n$	Позволяет ввести линейные ограничения на коэффициенты признаков. Если заменить знак $=$ в ограничении на \leq , то является обобщением предыдущих ограничений.

Имеется матрица "объект-признак" $X \in \mathbb{R}^{n \times p}$, а также вектор ответов \mathbf{y} . Стандартизируем данные, т.е.

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

Затем решается задача классификации с помощью логистической регрессии.

$$y_i = \pi(\mathbf{x}_i) + \varepsilon, \text{ где } \pi(\mathbf{x}_i) = \frac{1}{1 + e^{-\beta_0 - \mathbf{x}_i^T \boldsymbol{\beta}}}.$$

Здесь $\boldsymbol{\beta}$ -вектор коэффициентов признаков, определяемый по обучающей выборке с помощью принципа максимума правдоподобия.

Постановка задачи

Для отбора признаков предлагается использовать штраф elastic-net:

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

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

Здесь функция лог-правдоподобия:

$$L = \frac{1}{n} \sum_{i=1}^n y_i (\beta_0 + x_i^T \beta) - \log(1 + e^{\beta_0 + x_i^T \beta}).$$

Решается задача поиска условного минимума с ограничениями типа

$$\beta_j \geq 0, \text{ или } \beta_{j+1} \geq \beta_j, \text{ или}$$

$$\|\beta_{\mathcal{A}}\| \leq t, \text{ где } \beta_{\mathcal{A}} \text{ — подвектор } \beta.$$

В базовом алгоритме, не учитывающем предметно-экспертные ограничения, для упрощения задачи предлагается использовать квадратичную аппроксимацию функции правдоподобия, получаемую по формуле Тейлора:

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

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

Для того, чтобы найти выражение для β_j , используем необходимое и достаточное условие глобального экстремума для выпуклой функции.

Теорема

Пусть решается задача

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

без предметно-экспертных ограничений, тогда оценка оптимальных параметров:

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

Здесь S – оператор сжатия, $S(a, b) = \text{sign}(a)(|a| - b)_+$, а

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

Модификация алгоритма

Для модификации алгоритма воспользуемся методом внешних штрафных функций.

Рассмотрим, например, ограничение типа

$$\varphi(\beta) \leq 0,$$

где $\varphi(\beta)$ – выпуклая функция.

Предлагается вместо исходной задачи решать следующие задачи:

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

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

Функция $\delta_k(\beta)$ удовлетворяет свойствам внешней штрафной функции, если $r_{k+1} > r_k$.

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

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

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

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.

Теорема (Митяшов)

Пусть решается задача

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

$$\varphi(\beta) \leq 0, \varphi(\beta) - \text{выпуклая функция,}$$

тогда вместо данной задачи можно решать следующие:

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

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

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

Рассмотрим, например, ограничение типа $\beta_{j+1} \geq \beta_j$, тогда, используя данную теорему, можно получить выражения для β_l :

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

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

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

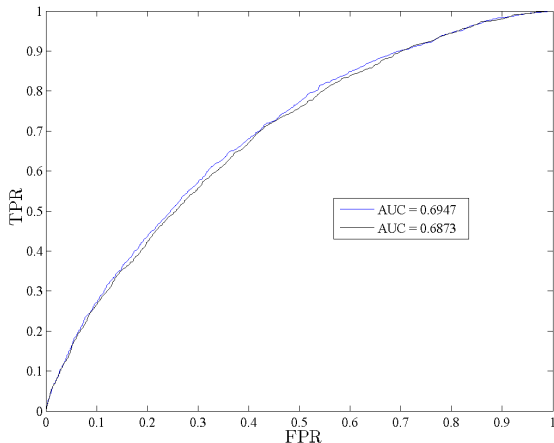
Аналогичным образом можно выписать выражения, которые получаются для остальных ограничений.

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.

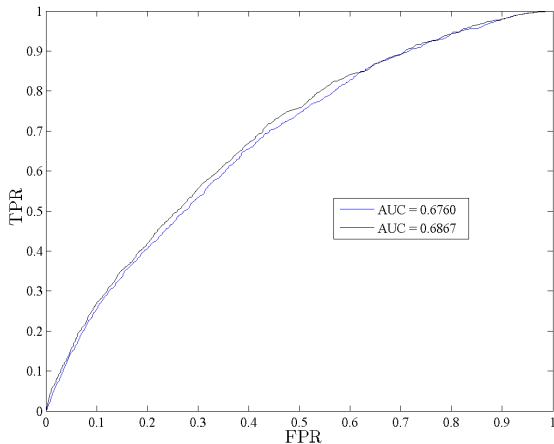
Были проверены гипотезы, представленные ранее.

Составлялись соответствующие обучающие выборки и на них строились модели с предметно-экспертными ограничениями и без них.



Здесь показано, что на исходной выборке AUC не увеличивается, так как мы ищем условный минимум.

Некорректные выборки



Здесь показан случай, когда обучающая выборка составлена некорректно, и введение дополнительных ограничений улучшает качество.

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.