Правила остановки последовательного выбора признаков и статистические методы для моделей с бинарным откликом

Научный руководитель: Булинский Александр Вадимович
Мехмат МГУ

Сперва рассмотрим результат из статье [1] Jan Mielniczuk, Pawel Teisseyre.

Будем рассматривать p-тое количество признаков $X_1, \ldots, X_p$ и целевая классовая величина(отклик) - Y. Пусть X_S - подмножество $X_1, \ldots, X_p$, $S \subset \{1, \ldots, p\}$. Определим $p\left(x_j\right) := P\left(X_j = x_j\right), x_j \in \mathcal{X}_j$, где $\mathcal{X}_j$ - область значений X_j и $\left|\mathcal{X}_j\right|$ - мощность множества(конечно). Область значений классовой величины - $\mathcal{Y}$(конечно). Совместная вероятность - $p\left(x_i, x_j\right) = P\left(X_i = x_i, X_j = x_j\right)$. $\hat{p}\left(x_j\right)$ - обозначение оценки $p\left(x_j\right)$.

Задача - зная значения признаков $X_1, \ldots, X_p$, предположить значение Y. Например, задача можем сводиться к вопросу о $P(Y = y|X_1 = x_1, \cdots, X_p = x_p)$

По известному набору наблюдений необходимо найти зависимость отклика от значений признаков.

Однако, наш отклик Y может не зависить от существенной части признаков, либо же иметь пренебрежительно слабую зависимость. Рассмотрение всех признаков может быть существенно затруднено, например, в силу высоких вычисленных требований в виде экспоненцильного роста сложности.

Цель - выбрать наиболее подходящие признаки.

$$\arg \max_{S:|S|=k} I\left(X_S, Y\right)$$

Однако, перебор по всем признаков является крайне затратным, так как сложность растет экспоненциально от числа признаков

Следовательно, можно перейти к последовательному выбору признаков:

$$\arg \max_{j \in S^c} \left[I\left(X_{S \cup \{j\}}, Y\right) - I\left(X_S, Y\right) \right] = \arg \max_{j \in S^c} I\left(X_j, Y \mid X_S\right)$$

Разложим условную совместную информацию через информации многостороннего взаимодействия:

$$
\begin{aligned}
& I\left(X_{S\cup\{j\}}, Y\right) - I\left(X_S, Y\right) \\
& = I\left(X_j, Y \mid X_S\right) = \sum_{k=0}^{|S|} \sum_{\{i_1,\ldots,i_k\}\subseteq S} II\left(X_{i_1}, \ldots, X_{i_k}, X_j, Y\right) \\
& \quad = I\left(X_j, Y\right) + \sum_{i\in S} II\left(X_i, X_j, Y\right) + \sum_{i_1,i_2\in S: i_1<i_2} II\left(X_{i_1}, X_{i_2}, X_j, Y\right) \\
& + \ldots + II\left(X_{i_1}, \ldots, X_{i_{|S|}}, X_j, Y\right)
\end{aligned} \tag{1}
$$

Для упрощения вычислений можно взять второй порядок приближения условной совместной информации:

$$\begin{aligned} J\left(X_j, S\right) &= I\left(X_j, Y\right) + \sum_{i \in S} II\left(X_i, X_j, Y\right) \\ &= I\left(X_j, Y\right) + \sum_{i \in S} \left[I\left(Y, X_j \mid X_i\right) - I\left(Y, X_j\right)\right] \\ &= I\left(X_j, Y\right)\left(1 - |S|\right) + \sum_{i \in S} I\left(Y, X_j \mid X_i\right) \end{aligned} \tag{2}$$

На практике приближение выше второго использовать затруднительно. Для r-того порядка приближения необходимо оценивать $(r+1)$-мерные вероятности. Пусть n - число наблюдений, для любого i - $|\mathcal{X}_i| = b$, следовательно, всего b^{r+1} - возможных комбинаций. n/b^{r+1} - в среднем наблюдений на комбинацию.

Например, если $n = 1000$, $b = 5$, $r = 2$, то $n/b^{r+1} = 8$. Если $r = 3$, то $n/b^{r+1} = 1.6$. То есть не удастся приближенно оценить вероятности.

Авторы строят последовательный выбор признаков. Необходима точка остановки, на которой выбор предполагаемых значимых признаков будет окончен.

Предположим выбрали S множество индексов признаков. Тогда выбор признаков останавливается при $I\left(Y, X_j \mid X_S\right) = 0$ для любого $j \in S^c$. Пользуемся приближением второго порядка.

Пусть S_k - множество индексов, выбранных на k-том шаге, где $S_0 = \emptyset$. На k-том шаге определим $S_{k+1} = S_k \cup \{j_k\}$ такое, что $j_k = \arg\max_{j \in S_k^c} J\left(X_j, S_k\right)$. Таким образом момент остановки в приближении:

$t := \arg\min_{1 \leq k \leq p}(J\left(X_j, S_k\right) = 0), \forall j \in S_k^c.$

Так как мы не знаем распределений, нужны оценки.

$$\begin{aligned}\hat{J}\left(X_j, S_k\right) &= \hat{I}\left(X_j, Y\right) + \sum_{i \in S_k} \hat{I}\left(X_i, X_j, Y\right) \\ &= \hat{I}\left(X_j, Y\right) + \sum_{i \in S_k} \left[\hat{I}\left(Y, X_j \mid X_i\right) - \hat{I}\left(Y, X_j\right)\right] \\ &= \hat{I}\left(X_j, Y\right)\left(1 - |S_k|\right) + \sum_{i \in S_k} \hat{I}\left(Y, X_j \mid X_i\right) \end{aligned} \tag{3}$$

Заметим, что при $J\left(X_j, S_k\right) = 0$, может выполняться $\hat{J}\left(X_j, S_k\right) > 0$.

Цель - построить приближение распределения $2n\hat{J}\left(X_j, S_k\right)$ при гипотезе, что $X_j \perp Y \mid X_{S_k}$; и ввести правило остановки $\hat{t}$, основанное на кватилях распределения.

Теорема 1

Пусть X, Y и Z - случайные величины, принимающие значения в $|\mathcal{X}|$, $|\mathcal{Y}|$ и $|\mathcal{Z}|$, соответственно. Предположим, что Y и Z независимы при условии X. Тогда $2n\hat{I}(Y, Z \mid X) \approx \sum_{i=1}^{|\mathcal{X}|} W_i$, где W_i имеет χ^2 асимптотическое распределение со $(|\mathcal{Y}| - 1)(|\mathcal{Z}| - 1)$ степенями свободы и $\approx$ - означает, что обе части отличаются только на величину, стремящуюся к нулю по вероятности.

□ Для упращения записи $\hat{p}_{ijk} = \hat{p}\left(X = x_i, Y = y_j, Z = z_k\right), \hat{p}_{ij} = \hat{p}\left(X = x_i, Y = y_j\right)$ и т.д.. Запишем, используя определение условной совместной информации:
$2n\hat{I}(Y, Z \mid X) = 2n \sum_{i,j,k} \hat{p}_{ijk} \log \frac{\hat{p}_{ijk}\hat{p}_i}{\hat{p}_{ij}\hat{p}_{ik}}$

$$= 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \hat{p}_{ijk}\hat{p}_i \log\left(1 + \frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}}\right) \tag{4}$$

Используя $\log(1 + x) = x - x^2/2 + O\left(x^3\right)$ для малых x, получим:

$$\log\left(1 + \frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}}\right) = \frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}} - \frac{1}{2}\frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\left(\hat{p}_{ij}\hat{p}_{ik}\right)^2} + O\left(\frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^3}{\left(\hat{p}_{ij}\hat{p}_{ik}\right)^3}\right) \tag{5}$$

Вставим выражение выше в (4), видим, что член содержащий последний член в (5) ограничен некоторой $C > 0$ как

$$
\begin{aligned}
& C \times 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^3}{\left(\hat{p}_{ij}\hat{p}_{ik}\right)^3} \\
& \leq C \times 2n \sum_i \frac{1}{\hat{p}_i} \times \frac{\max_{j,k} \left|\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right|}{\min_{j,k} \left(\hat{p}_{ij}\hat{p}_{ik}\right)^2} \sum_{j,k} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}} \\
& = C \times 2 \sum_i \frac{1}{\hat{p}_i^2} \times \frac{\max_{j,k} \left|\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right|}{\min_{j,k} \left(\hat{p}_{ij}\hat{p}_{ik}\right)^2} \\
& \quad \times n\hat{p}_i \sum_{j,k} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}}
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
&= C \times 2 \sum_i \frac{1}{\hat{p}_i^2} \times \frac{\max_{j,k} \left| \hat{p}_{ijk} \hat{p}_i - \hat{p}_{ij} \hat{p}_{ik} \right|}{\min_{j,k} \left(\hat{p}_{ij} \hat{p}_{ik} \right)^2} \\
&\times n \hat{p}_i \sum_{j,k} \frac{\left(\hat{p}_{ijk} \hat{p}_i - \hat{p}_{ij} \hat{p}_{ik} \right)^2}{\hat{p}_{ij} \hat{p}_{ik}}
\end{aligned} \tag{6}
$$

В силу условной независимости и сходимости $\hat{p}_{ij} \to p_{ij} > 0, \hat{p}_{ik} \to p_{ik} > 0$ имеем

$$
\left| \hat{p}_{ijk} \hat{p}_i - \hat{p}_{ij} \hat{p}_{ik} \right| \to \left| p_{ijk} p_i - p_{ij} p_{ik} \right| = 0 \tag{7}
$$

при $n \to \infty$.

В силу (7) и (10) ниже, последний член в (6) - сумма двух членов таких, что первый член сходится к нулю, и второй член имеет распределение хи-квадрат. Следовательно, из теоремы Слуцкого следует, что оценка в (6) сходится по вероятности к 0 , когда $n \to \infty$. Таким образом $2n\hat{\hat{I}}(Y, Z \mid X)$ примерно равен

$$
\begin{aligned}
& 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \left[\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik} + \hat{p}_{ij}\hat{p}_{ik}\right] \\
& \times \left[\frac{\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}}{\hat{p}_{ij}\hat{p}_{ik}} - \frac{1}{2}\frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\left(\hat{p}_{ij}\hat{p}_{ik}\right)^2}\right] \\
& = 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik} + \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}}\right. \\
& \left. - \frac{1}{2}\frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}} + \frac{1}{2}\frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^3}{\left(\hat{p}_{ij}\hat{p}_{ik}\right)^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= 2n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik} + \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}} \right. \\
&\left. - \frac{1}{2} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}} + \frac{1}{2} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^3}{\left(\hat{p}_{ij}\hat{p}_{ik}\right)^2} \right) \\
&\approx n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \left(2(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}) + \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}} \right) \\
&= n \sum_i \frac{1}{\hat{p}_i} \sum_{j,k} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}}
\end{aligned} \tag{8}
$$

где приближение в (8) получено аналогично как в (6), и последнее равенство следует из то, что $\sum_{j,k}(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}) = 0$. Последнее полученное выражение равно

$$n\sum_i \hat{p}_i \sum_{j,k} \frac{\left(\hat{p}_{ijk}/\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2\right)^2}{\hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2}$$

$$= \sum_i n_i \sum_{j,k} \frac{\left(\hat{p}_{ijk}/\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2\right)^2}{\hat{p}_{ij}\hat{p}_{ik}/\hat{p}_i^2} =: \sum_{i=1}^{|\mathcal{X}|} W_i \tag{9}$$

где $n_i = n\hat{p}_i$.

Заметим, что W_i есть хи-квадрат статистика для тестирования независимости Y и Z при $X = x_i$, которая при предположении независимости Y и Z при X имеет асомптотическое χ^2 распределение со $(|\mathcal{Y}| - 1)(|\mathcal{Z}| - 1)$ степенями свободы

$$W_i = n_i \sum_{j,k} \frac{\left(\hat{p}_{ijk}\hat{p}_i - \hat{p}_{ij}\hat{p}_{ik}\right)^2}{\hat{p}_{ij}\hat{p}_{ik}} \approx \chi^2_{(|\mathcal{Y}|-1)(|\mathcal{Z}|-1)} \tag{10}$$

в силу (Теорема 6.9, Shao Mathematical Statistics). Автором было показано, что число степеней свободы предельного χ^2 распределения равно $p - s - 1$, где $p = |\mathcal{Y}||\mathcal{Z}|$ и $s = (|\mathcal{Y}| - 1) + (|\mathcal{Z}| - 1)$, при предположении условной независимости $(|\mathcal{Y}| - 1) + (|\mathcal{Z}| - 1)$ условных вероятностей Y и Z при $X = x_i$ нужно определить, чтобы определить условное распределение (Y, Z) при $X = x_i$. Таким образом общее число степеней свободы равно $|\mathcal{Y}||\mathcal{Z}| - (|\mathcal{Y}| - 1) - (|\mathcal{Z}| - 1) - 1 = (|\mathcal{Y}| - 1)(|\mathcal{Z}| - 1)$. $\square$

Далее перейдём к теме потенциального применения правил остановки. Далее материал из статьи [2].

$X_1^j, \cdots, X_n^j$ - случайные величины признаков j-того наблюдения. Y^j - соотвественный бинарный отклик (-1 или 1). Пусть

$$\xi = \left(\xi^1, \cdots, \xi^N\right)$$

где

$$\xi^j = \left(X^j, Y^j\right), \quad j = 1, \cdots, N$$

Будем считать $\xi^1, \cdots, \xi^N$ - н.о.р. случайные вектора. Введём (X, Y) независимый с ξ и имеющий распределение ξ^1. Все случайные вектора рассматриваются на $(\Omega, \mathcal{F}, \mathrm{P}), \mathrm{E}$ - интегрирование по P.

Пусть $X := \{0, 1, 2\}^n$ - пространство возможных значений объясняющих переменных. Функция $f : X \to \{-1, 1\}$ называется теоритической функцией предсказания. Определим сбалансированную и нормированную ошибку предсказания для f как

$$\mathrm{Err}(f) := \mathrm{E}|Y - f(X)|\psi(Y)$$

где $\psi : \{-1, 1\} \to \mathbf{R}_+$ - штрафная функция. Следовательно,

$$\begin{aligned}\mathrm{Err}(f) =& 2\psi(-1)\mathrm{P}(f(X) = 1, Y = -1) \\ &+ 2\psi(1)\mathrm{P}(f(X) = -1, Y = 1)\end{aligned} \tag{1}$$

Очевидно $\mathrm{Err}(f)$ зависит от распределения (X, Y), но опустим обозначение. Ссылаясь на идеи предыдущих работ авторы устанавливают

$$\psi(y) = \frac{1}{4\mathrm{P}(Y = y)}, \quad y \in \{-1, 1\}$$

где случаи $\mathrm{P}(Y = -1) = 0$ и $\mathrm{P}(Y = 1) = 0$ исключаются. Тогда

$$\mathrm{Err}(f) = \frac{1}{2}\mathrm{P}(f(X) = 1 \mid Y = -1) + \frac{1}{2}\mathrm{P}(f(X) = -1 \mid Y = 1) \tag{2}$$

Если $\mathrm{P}(Y = -1) = \mathrm{P}(Y = 1) = 1/2$, то называем сбалансированным: $\mathrm{Err}(f) = \mathrm{E}|Y - f(X)|/2$.

Следовательно, $\mathrm{Err}(f)$ равно ошибке классификации $\mathrm{P}(Y \neq f(X))$. В общем,

$$\mathrm{Err}(f) = \frac{1}{2}\mathrm{E}\,|Y^* - f\,(X^*)|$$

где (X^*, Y^*) имеет распределение

$$\mathrm{P}\,(X^* = x, Y^* = y) = \frac{1}{2}\mathrm{P}(X = x \mid Y = y)$$

$$(x, y) \in X \times \{-1, 1\}$$

Можно заметить, что оптимальная теоритическая функция предсказания минимизирующая сбалансированную ошибку предсказания:

$$f^*(x) = \begin{cases} 1, p(x) > \mathrm{P}(Y = 1) \\ -1, \text{ иначе} \end{cases}$$

где

$$p(x) = \mathrm{P}(Y = 1 \mid X = x), x \in X \tag{4}$$

Тогда каждый $x \in \mathcal{X}$ классифицируется как вероятный, если $f^*(x) = 1$, иначе при $f^*(x) = -1$.

$p(x)$ и $\mathrm{P}(Y = 1)$ неизвестны. Необходимо найти приближение f^*, используя

$$f_{PA} = f_{PA}(x, \xi(S))$$

со значениями в $\{-1, 1\}$, которая зависит от $x \in \mathcal{X}$ и наблюдений

$$\xi(S) = \left\{\xi^j, j \in S\right\}$$

где

$$S \subset \{1, \cdots, N\} \tag{5}$$

Авторы вводят оценки $p(x)$ и $\mathrm{P}(Y = 1)$ как

$$\hat{p}(x, \xi(S)) = \frac{\sum_{j \in S} I\left\{Y^j = 1, X^j = x\right\}}{\sum_{j \in S} I\left\{X^j = x\right\}}, x \in \mathcal{X} \tag{6}$$

и

$$\hat{\mathrm{P}}_S(Y = 1) = \frac{1}{\#S} \sum_{j \in S} I\left\{Y^j = 1\right\} \tag{7}$$

где $\#D$ - мощность множества D.

Устанавливается аналогично для f_{PA}:

$$\begin{aligned}&\operatorname{Err}\left(f_{PA}(\cdot, \xi(S))\right)\\&= \frac{1}{2} \sum_{y \in \{-1,1\}} \mathrm{P}\left(f_{PA}(X, \xi(S)) \neq y \mid Y = y\right)\end{aligned} \tag{9}$$

Распределение (X, Y) неизвестно, строится оценка $\widehat{Err}\left(f_{PA}(\cdot, \xi(S))\right)$

Строится оценка для f_{PA} с помощью K-свёрточной кросс-валидации $(K > 1)$:

$$\hat{\mathrm{Err}}_K\left(f_{PA}(\cdot, \xi), \xi\right) = \frac{1}{2} \sum_{y \in \{-1,1\}} \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{(k)} I\left\{f_{PA}\left(X^j, \xi\left(\overline{S_k}\right)\right) \neq y, Y^j = y\right\}}{\sum_{(k)} I\left\{Y^j = y\right\}}$$

где сумма $\sum_{(k)}$ берётся по всем j принадлежащим

$$S_k = \{(k-1)[N/K] + 1, \cdots \\ k[N/K] I\{k < K\} + N I\{k = K\}\} \tag{11}$$

$$C_{k_1,\cdots,k_r}(x) = \left\{u = (u_1, \cdots, u_n) \in X : u_{k_i} = x_{k_i}, i = 1, \cdots, r\right\}.$$

Теорема 2

Предположим существует подмножество $U \subset X$ и подмножество $\{k_1, \cdots, k_r\} \subset \{1, \cdots, n\}$ такие, что выполнено:

1. *Для любого $x \in \mathcal{X}$ и каждого конечномерного вектора v с компонентами в $X \times \{-1, 1\}$, функции $f_{PA}(\cdot, v)$ and f постоянны на $C_{k_1,\cdots,k_r}(x)$.*
2. *Для любого $x \in U$ и каждого $W_N \subset \{1, \cdots, N\}$ с $\#W_N \to \infty$, верно $f_{PA}(x, \xi(W_N)) \to f(x)$ п.н. при $N \to \infty$.*
3. $\mathrm{P}\left(Y = 1 \mid X_{k_1} = x_{k_1}, \ldots, X_{k_r} = x_{k_r}\right) = \mathrm{P}(Y = 1)$, *если* $x \in X \backslash U$.
4. f *постоянна на* $X \backslash U$.

Тогда $\hat{\mathrm{Err}}_{\mathrm{K}}(f_{PA}(\cdot, \xi), \xi) \to \mathrm{Err}(f)$ *п.н.,* $N \to \infty$.

Замечание. Если заменить условие 3 в Теореме 1 более усиленным
3') $\mathrm{P}(Y = 1 \mid X = x) = \mathrm{P}(Y = 1)$ для любого $x \in X \backslash U$, то можно брать $\{k_1, \cdots, k_r\} = \{1, \cdots, n\}$, чтобы убрать условие 1.
□. Доказательство основано на следующей лемме:

Лемма 1

Пусть $\left\{ \left(Z_j^{(m)}, Y_j^{(m)} \right), \quad 1 \leq j \leq m, m \in N \right\}$ *- набор по строчно независимых случайных величин распределенных как* (Z, Y)*, где* Z *принимает значения в конечном множестве* Z*, и* Y *- в* $\{-1, 1\}$*.*
Предположим $\{f_m(z), m \in N, \mathrm{z} \in Z\}$ *- набор случаных величин со значениями в* $\{-1, 1\}$*.*
Предположим существует $U \subset Z$ *такое, что верны следующие условия:*

1. $f_m(z) \to f(z)$ *п.н. для любого* $z \in U$*, при* $m \to \infty$*, где неслучайная величина* $f : Z \to \{-1, 1\}$*.*
2. $\mathrm{P}(Y = 1 \mid Z = z) = \mathrm{P}(Y = 1)$*, если* $z \in Z \backslash U$*.*
3. f *- постоянная на* $Z \backslash U$*.*

Тогда $m \to \infty$*,*

$$\frac{1}{2} \sum_{y \in \{-1,1\}} \frac{\sum_{j=1}^{m} I\left\{ f_m\left(Z_j^{(m)} \right) \neq y, Y_j^{(m)} = y \right\}}{\sum_{j=1}^{m} I\left\{ Y_j^{(m)} = y \right\}} \to \mathrm{Err}(f) \tag{12}$$

□. Авторы вводят $Q_m(y) = \sum_{j=1}^{m} I\left\{Y_j^{(m)} = y\right\}$ и определяют следующие события:

$$A_j^{(m)}(y) = \left\{f_m\left(Z_j^{(m)}\right) \neq y, Y_j^{(m)} = y\right\}$$
$$B_j^{(m)}(y) = \left\{f\left(Z_j^{(m)}\right) \neq y, Y_j^{(m)} = y\right\}$$

Тогда левая часть (12) равна

$$\sum_{y \in \{-1,1\}} \frac{1}{2Q_m(y)} \sum_{j=1}^{m} I\left\{A_j^{(m)}(y)\right\}$$

Для $y \in \{-1, 1\}$ верно

$$\begin{aligned} &\frac{1}{Q_m(y)} \sum_{j=1}^{m} I\left\{A_j^{(m)}(y)\right\} \\ =&\frac{1}{m\mathrm{P}\{Y = y\}} \sum_{j=1}^{m} I\left\{A_j^{(m)}(y)\right\} \\ &+ \frac{1}{m} \sum_{j=1}^{m} I\left\{A_j^{(m)}(y)\right\} \left(\frac{m}{Q_m(y)} - \frac{1}{\mathrm{P}(Y = y)}\right) \end{aligned} \tag{13}$$

Абсолютная величина второго члена в правой части (13) не превышает $|m/\left(Q_m(y)\right) - 1/\mathrm{P}(Y = y)|$ и стремится к 0 п.н. при $m \to \infty$. Это следует из усиленного закона больших чисел для массивов(УЗБЧМ). Заметим

$$
\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m} I\left\{A_j^{(m)}(y)\right\} \\
=&\frac{1}{m}\sum_{j=1}^{m} I\left\{B_j^{(m)}(y)\right\} \\
&+\frac{1}{m}\sum_{j=1}^{m} I\left\{Z_j^{(m)} \in U\right\}\left(I\left\{A_j^{(m)}(y)\right\} - I\left\{B_j^{(m)}(y)\right\}\right) \\
&+\frac{1}{m}\sum_{j=1}^{m} I\left\{Z_j^{(m)} \notin U\right\}\left(I\left\{A_j^{(m)}(y)\right\} - I\left\{B_j^{(m)}(y)\right\}\right)
\end{aligned}
\tag{14}
$$

Согласно УЗБЧМ первый член в правой части (14) стремится $P(f(Z) \neq y, Y = y)$ п.н. Предположим, второй член стремится к 0 п.н. Действительно, множество Z - конечно и функции f, f_m принимают два значения. Следовательно, из условия 1, для почти всех $\omega \in \Omega$, существует $N_1 = N_1(\omega)$ такой, что $f_m(z) = f(z)$ для всех $z \in U$ и $m > N_1$. Тогда второй член в правой части (14) равен 0 для всех $m > N_1$, что доказывает предположении. Таким образом остаётся оценить третий член.

Из условия 3, б.о.о. можем принять $f(z) = -1$ для $z \in Z \backslash U$. Тогда получается

$$
\begin{aligned}
V_m &:= \sum_{y \in \{-1,1\}} \sum_{j=1}^{m} \frac{I\left\{Z_j^{(m)} \notin U\right\}\left(I\left\{A_j^{(m)}(y)\right\} - I\left\{B_j^{(m)}(y)\right\}\right)}{m\mathrm{P}(Y = y)} \\
&= \frac{1}{m} \sum_{j=1}^{m} I\left\{Z_j^{(m)} \notin U\right\} I\left\{f_m\left(Z_j^{(m)}\right) = 1\right\} R_j^{(m)} \qquad (15) \\
&= \frac{1}{m} \sum_{z \in Z \backslash U} I\left\{f_m(z) = 1\right\} \sum_{j=1}^{m} I\left\{Z_j^{(m)} = z\right\} R_j^{(m)}
\end{aligned}
$$

где

$$
R_j^{(m)} = \frac{I\left\{Y_j^{(m)} = -1\right\}}{\mathrm{P}(Y = -1)} - \frac{I\left\{Y_j^{(m)} = 1\right\}}{\mathrm{P}(Y = 1)}
$$

УЗБЧН и условие 2 приводят к тому, что для $z \in Z \backslash U$ и $y \in \{-1, 1\}$:

$$\sum_{j=1}^{m} \frac{I\left\{Z_j^{(m)} = z\right\} I\left\{Y_j^{(m)} = y\right\}}{m\mathrm{P}(Y = y)}$$

$$\to \frac{\mathrm{P}(Z = z, Y = y)}{\mathrm{P}(Y = y)} = \mathrm{P}(Z = z)$$

почти наверное. Следовательно, для почти всех $\omega \in \Omega$, существует $N_2 = N_2(\omega)$ такое, что

$$\left| \frac{1}{m} \sum_{j=1}^{m} I\left\{Z_j^{(m)} = z\right\} R_j^{(m)} \right| < \varepsilon$$

для всех $z \in Z \backslash U$ и $m > N_2$. Используя последнюю оценку и (15), получается, что для $m > N_2$,

$$|V_m| \leq \sum_{z \in Z \backslash U} \varepsilon I\left\{f_m(z) = 1\right\} \leq \varepsilon \cdot \#Z$$

Тогда $V_m \to 0$ п.н. при $m \to \infty$. Собирая (12)-(16), получается необходимый результат. $\square$
Теперь к доказательству Теоремы 1. Зафиксируем $1 \leq k \leq K$ и определим

$$f_m(z) := f_{PA}\left(x, \xi\left(\overline{S_k}\right)\right)$$

где $z \in Z := \{0,1,2\}^r, \quad m := \#S_k, S_k$ - определен в (11), и x - любой элемент из $\mathcal{X}$ с $\left(x_{k_1}, \cdots, x_{k_r}\right) = z$.
Из условия 1 f_m корректно определён.
Применим лемму к набору

$$\left\{\left(Z_j^{(m)}, Y_j^{(m)}\right), 1 \leq j \leq m\right\} := \left\{\left(\left(X_{k_1}^j, \ldots, X_{k_r}^j\right), Y^j\right), j \in S_k\right\}$$

и $\{f_m(z), z \in Z\}$, получается, что п.н.

$$\frac{1}{2} \sum_{y \in \{-1,1\}} \frac{\sum_{(k)} I\left\{f_{PA}\left(X^j, \xi\left(\overline{S_k}\right)\right) \neq y, Y^j = y\right\}}{\sum_{(k)} I\{Y^j = y\}} \to \mathrm{Err}(f)$$

при $\#S_k \to \infty$.

Таким образом

$$\hat{\mathrm{Err}}_K\left(f_{PA}(\cdot,\xi),\xi\right) \to \frac{1}{K}\sum_{k=1}^{K}\mathrm{Err}(f) = \mathrm{Err}(f)$$

при $N \to \infty$. $\square$

Как упомянуто ранее, отклик может зависить лишь от ограниченного числа признаков. В статье вводится понятие набора значимых признаков - существует $l \in \mathbf{N}, l < n$, и вектор $\left(k_1^*, \cdots, k_l^*\right)$, где $1 \leq k_1^* < \cdots < k_l^* \leq n$ такой, что для каждого $x = (x_1, \cdots, x_n) \in \mathcal{X}$ верно:

$$p(x) = \mathrm{P}\left(Y = 1 \mid X_{k_1^*} = x_{k_1^*}, \cdots, X_{k_l^*} = x_{k_l^*}\right) \tag{17}$$

Набор $\left(k_1^*, \cdots, k_l^*\right)$ имеющий наименьший размер l называется наиболее значимым.

Для $x \in \mathcal{X}$ и $\{k_1, \cdots, k_r\} \subset \{1, \cdots, n\}$, вводится

$$f_{k_1,\cdots,k_r}(x) = \begin{cases} 1, \mathrm{P}\left(Y = 1 \mid X_{k_1} = x_{k_1}, \ldots, X_{k_r} = x_{k_r}\right) > \mathrm{P}(Y = 1) \\ -1, \text{ иначе.} \end{cases}$$

Аналогично вводится оценка $\widehat{f}_{k_1,\cdots,k_r}(x)$.

Теорема 3

Пусть $\left(k_1^*, \cdots, k_l^*\right)$ *- наиболее значимый набор признаков. Тогда для любого фиксированного* $\{k_1, \cdots, k_r\} \subset \{1, \cdots, n\}$ *верно*

1. $\mathrm{Err}\left(f_{k_1^*, \cdots, k_l^*}\right) \leq \mathrm{Err}\left(f_{k_1, \cdots, k_r}\right)$;
2. $\hat{\mathrm{E}}\mathrm{r}r_K\left(\hat{f}_{k_1, \cdots, k_r}\right)$ *- сильная состоятельная асимптотически несмещенная оценка* $\mathrm{Err}\left(f_{k_1, \cdots, k_r}\right)$ *при* $N \to \infty$;
3. *Для всех* $\varepsilon, \delta > 0$ *и для всех достаточно больших* N

$$\mathrm{P}\left(\widehat{Err}_K\left(\widehat{f}_{k_1^*, \cdots k_l^*}\right) < \widehat{Err}_K\left(\widehat{f}_{k_1, \cdots, k_r}\right) + \varepsilon\right) > 1 - \delta$$

□ 1) Из (17) следует, что $f_{k_1^*,\cdots,k_l^*}$ совпадает с функцией f^*, которая является минимумом весовой ошибки прогнозирования.

2. Проверим условия теоремы 1 для $f_{PA}(x,\xi(S)) := \widehat{f}_{k_1,\cdots,k_r}(x,\xi(S))$. Условие 1 следует из определения $f_{k_1,\cdots,k_r}(x,\xi(S))$. Далее

$$U := \left\{x \in X : \mathrm{P}\left(Y=1 \mid C_{k_1,\ldots,k_r}(x)\right) \neq \mathrm{P}(Y=1)\right\} \tag{19}$$

где $C_{k_1,\cdots,k_r}$ введено раньше. Предположим, что для каждого $x \in U$ и любого $W_N \subset \{1,\cdots,N\}$, что $\#W_N \to \infty$, верно:

$$\widehat{f}_{k_1,\cdots,k_r}\left(x,\xi\left(W_N\right)\right) \to f_{k_1,\cdots,k_r}(x) \text{ п.н. при } N \to \infty$$

Действительно, предположим, что для некоторого $\varepsilon > 0$,

$$\mathrm{P}\left(Y=1 \mid X_{k_1}=x_{k_1},\cdots,X_{k_r}=x_{k_r}\right) - \mathrm{P}(Y=1) > \varepsilon$$

Из УЗБЧН следует, что

$$\hat{\mathrm{P}}_{W_N}\left(Y=1 \mid X \in C_{k_1,\ldots,k_r}(x)\right) - \hat{\mathrm{P}}_{W_N}(Y=1)$$

сходится п.н. к

$$\mathrm{P}\left(Y=1 \mid X_{k_1}=x_{k_1},\cdots,X_{k_r}=x_{k_r}\right) - \mathrm{P}(Y=1)$$

$N \to \infty$. Тогда, для почти всех $\omega \in \Omega$ существует $N_0 = N_0(\omega)$ такой, что

$$\hat{\mathbf{P}}_{W_N}\left(Y=1 \mid X \in C_{k_1,\cdots,k_r}(x)\right) - \hat{\mathrm{P}}_{W_N}(Y=1) > \varepsilon/2$$

для всех $N > N_0$. Следовательно, для всех $N > N_0$ имеем $\hat{f}_{k_1,\cdots,k_r}(x,\xi(W_N)) = 1 = f_{k_1,\cdots,k_r}(x)$, что доказывает предположение. Условие 2 Теоремы 1 удовлетворено.
Условия 3 и 4 следуют из (19) и определения $\widehat{f}_{k_1,\cdots,k_r}(x,\xi(S))$ и $f_{k_1,\cdots,k_r}(x)$.
Следовательно, из Теоремы 1 следует, что $\hat{\mathrm{E}}\mathrm{r}r_K\left(\hat{f}_{k_1,\cdots,k_r}\right) \to \mathrm{Err}\left(f_{k_1,\cdots,k_r}\right)$ п.н. и в L_2 (в силу теоремы Лебега из ограниченности $\widehat{Err}_K\left(\widehat{f}_{k_1,\cdots,k_r}\right)$ величиной 1).

3. Следует из 1) и 2).

□

Стоит отметить, что частично аналогичные результаты вместе с другими дополнительными результатами, но уже для небинарного отклика получены с статьях [4], [5], [6].

Рассмотрим процесс стратификации из [3]

$(X^1, Y^1), (X^2, Y^2), \ldots$ - н.о.р как (X, Y). Общий набор наблюдений разбивается на два набора:

$$\zeta_{N_1}^1 := \left\{ \left(X^{j_1^1}, 1 \right), \ldots, \left(X^{j_1^{N_1}}, 1 \right) \right\}, \zeta_{N_{-1}}^{-1} := \left\{ \left(X^{j_{-1}^1}, -1 \right), \ldots, \left(X^{j_{-1}^{N_{-1}}}, -1 \right) \right\}$$

где j_1^k и $j_{-1}^k, k \in \mathbb{N}$, упорядочены. $N := N_1 + N_{-1}$ для $\zeta_N := \zeta_{N_1}^1 \cup \zeta_{N_{-1}}^{-1}$. В отличие от $\xi_N := \left\{ (X^1, Y^1), \ldots, (X^N, Y^N) \right\}$, имеющий закон $\mathrm{law}(X, Y)$, есть $\zeta_{N_1}^1$ и $\zeta_{N_{-1}}^{-1}$ с распределениями $X \mid Y = 1$ и $X \mid Y = -1$. Поэтому нельзя использовать оценки (например, of $\mathrm{P}(Y = 1)$ или $\mathrm{P}(X \in B, Y = 1)$, где $B \subset \mathbb{X}$) построенные на средних ζ_N. Пусть $N_1 = \max\{[aN], 1\}$ и $N_{-1} = N - N_1$

Предположим есть оценка $\widehat{\mathrm{P}}_N^y$ от $\mathrm{P}(Y = y)$:

$$\widehat{\mathrm{P}}_N^y \to \mathrm{P}(Y = y) \text{ п.н.}, N \to \infty, y \in \{-1, 1\} \tag{8}$$

Например, $\widehat{\mathrm{P}}_N^y$ использует $\left\{ \left(X^k, Y^k \right), 1 \leqslant k \leqslant \max \left\{ j_1^{N_1}, j_{-1}^{N_{-1}} \right\} \right\}$. Тогда частотные оценки$\mathrm{P}(Y = 1)$ или $\mathrm{P}(X \in B, Y = 1), B \subset \mathbb{X}$, сильно состоятельны, $\max \left\{ j_1^{N_1}, j_{-1}^{N_{-1}} \right\} \to \infty$ п.н. $N \to \infty$.

$\widehat{\mathrm{P}}_N := \left(\widehat{\mathrm{P}}_N^{-1}, \widehat{\mathrm{P}}_N^1 \right)$. $\mathrm{P}(Y = -1) = 0$ и $\mathrm{P}(Y = 1) = 0$ не рассматривается.

Пусть $f_{PA}\left(x, \zeta_N, \widehat{\mathrm{P}}_N\right)$ - функция определяющая алгоритм предсказания, т.е. функция в $\{-1,1\}$ зависящая от $x \in \mathbb{X}$, наблюдений ζ_N и $\widehat{\mathrm{P}}_N$. Аналогично определяется на поднаборах. $f_{PA}\left(x, \zeta_N(S), \widehat{\mathrm{P}}_N\right)$ for $\zeta_N(S) := \left\{\left(X^j, Y^j\right), j \in S\right\}, S \subset \left(\left\{j_1^1, \ldots, j_1^{N_1}\right\} \cup \left\{j_{-1}^1, \ldots\right.\right.$ Для каждой $f : \mathbb{X} \to \{-1,1\}$ находится оценка f_{PA}, которая используется на оценки $\mathrm{Err}(f)$. Для этого используется K-свёрточная кросс-валидация или по-другому поднаборны алгоритм усреднения. Для фиксированного $K \in \mathbb{N}$ и любого $y \in \{-1,1\}$, вводится разбиение $\left\{j_y^1, \ldots, j_y^{N_y}\right\}$ на K поднаборов $S_k^y\left(N_y, \omega\right), k = 1, \ldots, K$:

$$S_k^y\left(N_y, \omega\right) := \left\{j_y^i(\omega) : i \in \left\{(k-1)\left[\frac{N_y}{K}\right] + 1, \ldots, k\left[\frac{N_y}{K}\right] \mathbb{I}\{k < K\} + N_y \mathbb{I}\{k = K\}\right\}\right\} \tag{9}$$

Вводится $S_k(N,\omega) = S_k^1(N_1,\omega) \cup S_k^{-1}(N_{-1},\omega)$ и

$$\widehat{\mathrm{Err}}_K\left(f_{PA},\zeta_N,\widehat{\mathrm{P}}_N\right) := \frac{2}{K}\sum_{y\in\{-1,1\}}\sum_{k=1}^{K}\sum_{j\in S_k^y(N_y)}\frac{\widehat{\psi}\left(y,\zeta_N\left(\overline{S_k(N)}\right),\widehat{\mathrm{P}}_N\right)\mathbb{I}\left\{f_{PA}^j(N,k)\neq y\right\}\widehat{\mathrm{P}}_N^y}{\sharp S_k^y(N_y)} \tag{10}$$

где $f_{PA}^j(N,k) := f_{PA}\left(X^j,\zeta_N\left(\overline{S_k(N)}\right),\widehat{\mathrm{P}}_N\right)$ и $\widehat{\psi}\left(y,\zeta_N\left(\overline{S_k(N)}\right),\widehat{\mathrm{P}}_N\right)$ - оценка величины $\psi(y), y\in\{-1,1\}$, построенная на наблюдениях $\zeta_N\left(\overline{S_k(N)}\right)$ и $\widehat{\mathrm{P}}_N$; $\overline{S_k(N)} = \left\{j_1^1,\ldots,j_1^{N_1}\right\}\cup\left\{j_{-1}^1,\ldots,j_{-1}^{N_{-1}}\right\}\setminus S_k(N)$, $\sharp$ - мощность множества.

Предполагается, что для любого $k = 1, \ldots, N$,

$$\widehat{\psi}\left(y, \zeta_N\left(\overline{S_k(N)}\right), \widehat{\mathrm{P}}_N\right) \to \psi(y) \text{ п.н.}, N \to \infty, \quad y \in \{-1, 1\} \tag{11}$$

Определяется $L(x) = \psi(1)\mathrm{P}(X = x, Y = 1) - \psi(-1)\mathrm{P}(X = x, Y = -1), x \in \mathbb{X}$.

Теорема 4

ζ_N, ψ. $\widehat{\psi}, f : \mathbb{X} \to \{-1, 1\}$ и f_{PA} - определены как ранее. Пусть существует не пустой $U \subset \mathbb{X}$ такой, что для каждого $x \in U$ и $k = 1, \ldots, K$, верно:

$$f_{PA}\left(x, \zeta_N\left(\overline{S_k(N)}\right), \widehat{\mathrm{P}}_N\right) \to f(x) \text{ п.н., } N \to \infty \tag{12}$$

Тогда для каждого $a \in (0, 1)$ (с $N_1 = \max\{[aN], 1\}, N_{-1} = N - N_1$),

$$\widehat{\mathrm{Err}}_K\left(f_{PA}, \zeta_N, \widehat{\mathrm{P}}_N\right) \to \mathrm{Err}(f) \text{ a.s., } N \to \infty \tag{13}$$

равносильно

$$\sum_{k=1}^{K} \sum_{y \in \{-1,1\}} \sum_{x \in \mathbb{X}_y} y \mathbb{I}\left\{f_{PA}\left(x, \zeta_N\left(\overline{S_k(N)}\right), \widehat{\mathrm{P}}_N\right) = -y\right\} L(x) \to 0 \text{ a.s., } N \to \infty \tag{14}$$

где

$$\mathbb{X}_y = (\mathbb{X} \backslash U) \cap \{x \in \mathbb{X} : f(x) = y\}, \quad y \in \{-1, 1\} \tag{15}$$

Для доказательства сперва необходимо воспользоваться леммой

Пусть $(X, Y), \left(X^1, Y^1\right), \left(X^2, Y^2\right), \ldots$ н.о.р. на вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$. Для каждого $\omega \in \Omega$ рассматривается $Y^1(\omega), Y^2(\omega), \ldots$ и вводятся индексы $1 \leqslant j_{-1}^1(\omega) < j_{-1}^2(\omega) < \ldots$, для которых $Y^{j_{-1}^k(\omega)}(\omega) = -1, k \in \mathbb{N}$. Аналогично для всех $Y^i(\omega)$ со значениями 1 как $\left\{Y^{j_1^m(\omega)}(\omega)\right\}_{m \in \mathbb{N}}$, где $1 \leqslant j_1^1(\omega) < j_1^2(\omega) < \ldots$ Используя распределение Бернулли с вероятностью успеха p, вводится отрицательная биномиальная величина $U_{r,p}$ - число успехов до r неудач
где $r \in \mathbb{N}$ ($U_{r,p} \sim NB(r, p)$).

Тогда

$$\mathrm{P}\left(U_{r,p}=k\right)=\left(\begin{array}{c}k+r-1\\k\end{array}\right)p^{k}(1-p)^{r},\quad k=0,1,\ldots$$

$\left\{Y^{i}=1\right\}$ и $\left\{Y^{i}=-1\right\}$ рассматриваются как успехи и неудачи (с вероятностью $p=\mathrm{P}(Y=1)$ успеха), тогда j_{-1}^{r} также распределён как $U_{r,p}+r$. Следовательно,

$$\mathrm{P}\left(j_{-1}^{r}=m\right)=\begin{cases}\left(\begin{array}{c}m-1\\m-r\end{array}\right)p^{m-r}(1-p)^{r}, & m=r,r+1,\ldots\\0, & m=1,\ldots,r-1\end{cases}\tag{4}$$

j_1^r распределён как $U_{r,1-p}+r$, где $U_{r,1-p} \sim NB(r, 1-p)$. j_1^r распределен также как $G_p^1+\ldots+G_p^r$, где $G_p^1, \ldots, G_p^r$ - независимые величины с геометрическими распределениями и с параметром p ($\mathrm{P}\left(G_p^1=k\right)=p(1-p)^{k-1}$, $k=1,2,\ldots$). Тогда $\mathrm{E}j_1^r=\frac{r}{p}$ и $j_1^r<\infty$ п.н. для любого $r \in \mathbb{N}$ ($\mathrm{E}j_{-1}^r=\frac{r}{1-p}$ для $r \in \mathbb{N}$).

Определяются $Z^k := X^{j_1^k}$ для каждого $k \in \mathbb{N}$. $\mathcal{B}$ - семейство всех подмножеств $\mathbb{X}$.

Лемма 2

Для каждого $m \in \mathbb{N}$, случайные величины $Z^1, \ldots, Z^m$ - независимы и распределены как X при $Y = 1$ ($X \mid Y = 1$), т.е., для каждого $B \in \mathcal{B}$ и $k = 1, \ldots, m$,

$$\mathrm{P}\left(Z^k \in B\right) = \mathrm{P}(X \in B \mid Y = 1)$$

Теперь затронем тему логистической регрессии. Как упоминалось ранее:

$$f^*(x) = \begin{cases} 1, p^*(x) > 1/2 \\ -1, \text{ otherwise} \end{cases}$$

Случаи $p^*(x) \in \{0, 1\}$ исключаются. Для оценки $p^*(x)$ используется

$$q^*(x) = \lambda\left(p^*(x)\right) \tag{24}$$

где $\lambda(z) = \log(z/(1-z)), \quad z \in (0, 1)$, обратная логистическая функция. Логистическая функция - $\Lambda(t) = \left(1 + e^{-t}\right)^{-1}, t \in \mathbf{R}$.

$p^*(x) \in (0,1)$ для $x \in \mathcal{X}$, $q^*(x)$ сожет принимать любые действительные значения. Рассматривается класс G всех вещественно значных функций на тернарных значениях $x_1, \cdots, x_n$. $M \subset \mathcal{G}$ - называется моделью зависимости отклика и объясняющих переменных. Введем

$$\hat{\psi}(y, \xi(S)) = \frac{1}{4\hat{\mathrm{P}}_S(Y = y)}, y \in \{-1, 1\}$$

$\hat{\mathrm{P}}_S(Y = y)$ - оценка .

Введем функцию оценки регрессии:

$$L(h, \xi(S)) = \frac{1}{\#S} \sum_{j \in S} \varphi\left(-Y^j h\left(X^j\right)\right) \hat{\psi}\left(Y^j, \xi(S)\right) \tag{25}$$

$\varphi(t) = \log_2\left(1 + e^t\right)$ для $t \in \mathbf{R}$, и $h \in M$. В отличие от других работ в этой статье авторы рассматривают нормировку, т.е. наблюдения с весами от отношения случаев в подвыборке $\xi(S)$. $\arg\min_{h \in M} L(h, \xi(S))$ равен $\arg\max_{h \in M}$ для

$$\frac{1}{\#S} \sum_{j \in S} \left(\frac{I\left\{Y^j = 1\right\}}{2\hat{\mathrm{P}}_S(Y = 1)} \log \gamma_j + \frac{I\left\{Y^j = -1\right\}}{2\hat{\mathrm{P}}_S(Y = -1)} \log\left(1 - \gamma_j\right) \right)$$

где $\gamma_j = \Lambda\left(h\left(X^j\right)\right)$. Минимизация эквивалентна нормированной оценки максимума правдоподобия q^*.

Следующая теорема о строгой состоятельности оценки при корректной модели, т.е. $q^* \in M$. Введём $h(\cdot, \xi(S)) := \arg\min_{q \in M} L(q, \xi(S))$.

Теорема 5

Пусть $q^* \in M, h_0 \equiv 0$ *принадлежит* M *и*

$$\min_{(x,y) \in X \times \{-1,1\}} P(X = x \mid Y = y) > 0$$

Рассмотрим $W_N \subset \{1, \cdots, N\}$ *и множество*

$$h_N(\cdot) = h\left(\cdot, \xi\left(W_N\right)\right)$$

Тогда $h_N(x) \to q^*(x)$ *п.н. для всех* $x \in \mathcal{X}$*, когда* $\#W_N \to \infty$*. Более того,*

$$\mathrm{Err}_K\left(f_{PA}(\cdot, \xi), \xi\right) \to \mathrm{Err}\left(f^*\right) \text{ a.s., } N \to \infty$$

где $f_{PA}(\cdot, \xi) = 2I\{\Lambda(h(\cdot, \xi)) > 1/2\} - 1$.

□ Сперва показывается

$$h_N(x) < \varphi^{-1}\left(\frac{4}{\mathrm{P}(X = x \mid Y = -1)}\right) \text{ п.н.}$$

для всех $x \in \mathcal{X}$ и всех $N > N_1 = N_1(\omega)$. Положим $l_N = \#W_N$. По определению

$$\frac{\varphi\left(h_N(x)\right)}{4l_N} \sum_{j \in W_N} \frac{I\left\{X^j = x, Y^j = -1\right\}}{\hat{\mathrm{P}}_{W_N}(Y = -1)} \leq L\left(h_N, \xi\left(W_N\right)\right)$$

$$\leq L\left(0, \xi\left(W_N\right)\right) = \frac{1}{4l_N} \sum_{(j,y) \in W_N \times \{-1,1\}} \frac{I\left\{Y^j = y\right\}}{\hat{\mathrm{P}}_N(Y = y)}$$

Пользуясь УЗБЧН, получается сходимость п.н.

$$\max_{x \in X} \left| \frac{1}{l_N} \sum_{j \in W_N} \frac{I\left\{X^j = x, Y^j = -1\right\}}{\hat{\mathrm{P}}_{W_N}(Y = -1)} \right.$$
$$\left. -\mathrm{P}(X = x \mid Y = -1) \right| \to 0$$

Очевидно

$$\frac{1}{l_N} \sum_{(j,y) \in W_N \times \{-1,1\}} \frac{I\left\{Y^j = y\right\}}{\hat{\mathrm{P}}_{W_N}(Y = y)} = 2.$$

Эти выражения приводят к желаемой оценке $h_N(x)$. Аналогично

$$h_N(x) > -\varphi^{-1}\left(\frac{4}{\mathrm{P}(X = x \mid Y = 1)}\right)$$

для каждого $x \in \mathcal{X}$ и всех $N > N_2 = N_2(\omega)$. Следовательно, $h_N \in M_C := M \cap \{h : \|h\|_\infty \leq C\}$ для $N > \max(N_1, N_2)$, здесь $\|h\|_\infty = \max_{x \in X} |h(x)|$ и

$$C = \max_{(x,y) \in X \times \{-1,1\}} \varphi^{-1}\left(\frac{4}{\mathrm{P}(X = x \mid Y = y)}\right)$$

Если $h \in M_C$, тогда $|L(h, \xi(W_N)) - \mathrm{E}\varphi(-Yh(X))\psi(Y)|$ меньше

$$\frac{\varphi(\|h\|_\infty)}{2} \sum_{y \in \{-1,1\}} \left(\left| \frac{1}{\hat{\mathrm{P}}_{W_N}(Y=y)} - \frac{1}{\mathrm{P}(Y=y)} \right| \right.$$
$$\left. + \sum_{x \in X} \left| \sum_{j \in W_N} \frac{I\{X^j = x, Y^j = y\}}{l_N \mathrm{P}(Y=y)} - \frac{\mathrm{P}(X^j = x, Y^j = y)}{\mathrm{P}(Y=y)} \right| \right)$$

Из УЗБЧН следует

$$L(h, \xi(W_N)) - \mathrm{E}\varphi(-Yh(X))\psi(Y) \to 0 \text{ a.s.} \tag{26}$$

равномерно по $\{h : \|h\|_\infty \leq C\}$.
Также

$$2\mathrm{E}\varphi(-Yh(X))\psi(Y) = \mathrm{E}\varphi(-Y^* h(X^*))$$
$$= -\mathrm{E}\log_2 \left(\Lambda(h(X^*))^{I\{Y^*=1\}} (1 - \Lambda(h(X^*)))^{1-I\{Y^*=1\}} \right)$$

Из (условного) информационного неравенства,

$$\mathrm{E}\log_2\left(\Lambda\left(h\left(X^*\right)\right)^{I\left\{Y^*=1\right\}}\left(1-\Lambda\left(h\left(X^*\right)\right)\right)^{1-\left\{Y^*=1\right\}}\right)$$

достигает своего максимума по функциям h только на q^*, представленный в (24). В условиях теоремы, $q^* \in M$. Следовательно, по определению h_N и q^* получается

$$L\left(h_N, \xi\left(W_N\right)\right) \leq L\left(q^*, \xi\left(W_N\right)\right)$$
$$\mathrm{E}\varphi\left(-Y^* h\left(X^*\right)\right)\big|_{h=h_N} \geq \mathrm{E}\varphi\left(-Y^* q^*\left(X^*\right)\right)$$

Из (26) и УЗБЧН

$$L\left(h_N, \xi\left(W_N\right)\right) - \frac{1}{2}\mathrm{E}\varphi\left(-Y^* h\left(X^*\right)\right)\Big|_{h=h_N} \to 0 \text{ a.s.},$$
$$L\left(q^*, \xi\left(W_N\right)\right) \to \frac{1}{2}\mathrm{E}\varphi\left(-Y^* q^*\left(X^*\right)\right) \text{ a.s.}$$

Тогда

$$\mathrm{E}\varphi\left(-Y^* h\left(X^*\right)\right)\big|_{h=h_N} \to \mathrm{E}\varphi\left(-Y^* q^*\left(X^*\right)\right) \text{ a.s.}$$

Это возможно только когда $h_N(x) \to q^*(x)$ п.н. для каждого $x \in X$. Действительно, для почти всех $\omega \in \Omega$, можно всегда выбрать подпоследовательность $h_{N_k} = h_{N_k(\omega)}(\cdot, \omega)$ сходящаяся к некоторой функции $\mu = \mu(\cdot, \omega)$ и

$$\mathrm{E}\varphi\left(-Y^*\mu\left(X^*\right)\right) = \mathrm{E}\varphi\left(-Y^*q^*\left(X^*\right)\right)$$

Тогда из информационного неравенства $\mu(\cdot, \omega) = q(\cdot)$.
Для доказательства второй части Теоремы 3 нужно заметить, что $f_{PA}(x, \xi) = 2I\{\Lambda(h(x, \xi)) > 1/2\} - 1$ сходится п.н. к $f^*(x) = 2I\left\{\Lambda\left(q^*(x)\right) > 1/2\right\} - 1$ для почти всех $x \in U$, где

$$\begin{aligned} U &:= \{x \in X : p^*(x) \neq 1/2\} \\ &= \{x \in X : \mathrm{P}(Y = 1 \mid X = x) \neq \mathrm{P}(Y = 1)\} \end{aligned}$$

Остаётся только воспользоваться Теоремой 1 и замечанием к ней. □

У меня есть некоторые вопросы про нефиксированное число признаков.

S - множество номеров всех признаков, m - количество фиксированных признаков, n - число наблюдений, $\mathbb{S}_m$ - семейство множеств номеров релевантных признаков размера m. Q_m — все подмножества S размера m. Цель:

$$argmax_{L \in Q_m} I(X_L, Y)$$

У Алексея Кожевина при дополнительных ограничениях доказывается в статье 2020 года:

$$\widehat{S}_{n,k}(w) = argmax_{L \in Q_m} \widehat{I}_{n,k,L}(w)$$

$$P(\widehat{S}_{n,k} \subset S_m) \longrightarrow 1$$

$$n \to \infty$$

У меня есть вопрос. Пусть введём максимум по всем наборам:

$$\widehat{S}_{n,k}(w) = argmax_{L \subset S} \widehat{I}_{n,k,L}(w)$$

$$u_{n,k} = min_{M \in \widehat{S}_{n,k}(w)} |M|$$

$$\widehat{S}^{min}_{n,k}(w) = \{M \in \widehat{S}_{n,k}(w), |M| = u_{n,k}\}$$

Пусть $\mathbb{S}$ - семейтсво множеств релевантных признаков. Пусть $\mathbb{S}^{min}$ - семейство наименьших по размеру множеств релевантных признаков. Следующие предположения выглядят правдоподобно при некоторых ограничениях:

$$P(\widehat{S}_{n,k}(w) \subset \mathbb{S}) \longrightarrow 1$$

$$P(\widehat{S}^{min}_{n,k}(w) \subset \mathbb{S}^{min}) \longrightarrow 1$$

$$n \to \infty$$

Насколько вообще такие предположения могут быть интересными? Я полагаю, что слишком большой перебор признаков, если оценку двигать по всем возможным комбинациям признаков.

Перейду к аналогичному вопросу, но уже с последовательным выбором признаков.
Главная цель:

$$argmax_{L \in Q_m} I(X_L, Y)$$

Последовательный выбор признаков на (k+1)-том шагу:

$$j_{k+1} = argmax_{j \in S_k^c}(I(X_{S_k \cup \{j\}}, Y) - I(X_{S_k}, Y)) = argmax_{j \in S_k^c} I(X_j, Y | X_{S_k})$$

$$S_{k+1} = S_k \cup \{j_{k+1}\}$$

$$t^{stop} := argmin_{1 \le k \le p}(I(X_j, Y | S_k) = 0)$$

Оценочный stopping rule:

$$\widehat{t}_n^{stop} = argmin_{1 \le k \le p}(\widehat{I}(X_j, Y | S_k) = 0)$$

Насколько сильны и неправдоподобны следующие предположения? А именно:

1. $S_{t^{stop}} \subset \mathbb{S}$
2. $S_{t^{stop}} \subset \mathbb{S}^{min}$
3. $P(S_{\widehat{t}_n^{stop}} \subset \mathbb{S}) \longrightarrow 1, \quad n \to \infty$
4. $P(S_{\widehat{t}_n^{stop}} \subset \mathbb{S}^{min}) \longrightarrow 1, \quad n \to \infty$

Аналогично для функции ошибки в статье для бинарного отклика:

$$\widehat{Err}_K(f_{\hat{t}_n^{stop}}) \longrightarrow Err(f_M)$$

где $M \in \mathbb{S}^{min}$

Ссылки

1. Stopping rules for mutual information-based feature selection, Jan Mielniczuk, Pawe l Teisseyre
   https://www.sciencedirect.com/science/article/abs/pii/S0925231219307544
2. Statistical Methods of SNP Data Analysis and Applications, Alexander Bulinski, Oleg Butkovsky, Victor Sadovnichy, Alexey Shashkin, Pavel Yaskov, Alexander Balatskiy , Larisa Samokhodskaya, Vsevolod Tkachuk
   https://www.scirp.org/journal/paperinformation?paperid=16881
3. New version of the MDR method for stratified samples, Alexander Bulinski, Alexey Kozhevin
   https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=soic&paperid=1&option_lang=rus
4. Estimation of nonbinary random response, Bulinski A.V., Rakitko A.S.
   https://link.springer.com/article/10.1134/S1064562414020306
5. MDR method for nonbinary response variable, Bulinski A., Rakitko A.
   https://doi.org/10.1016/j.jmva.2014.11.008
6. Forward Selection of Relevant Factors by Means of MDR-EFE Method, Alexander Bulinski
   https://www.mdpi.com/2227-7390/12/6/831