Прикладная статистика 10. Логистическая регрессия.

8 ноября 2013 г.

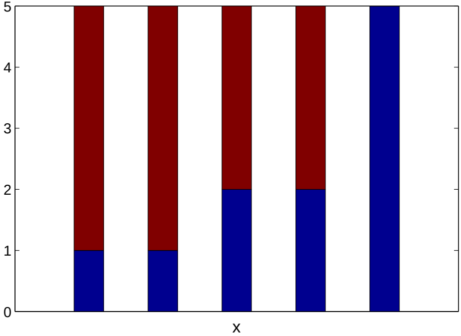
Постановка

Задача: оценить влияние одного или нескольких признаков на наступление какого-либо события и оценить его вероятность.

$$(x_i, y_i)\,,\ x_i \in \mathbb{R}^k,\ y_i \in \{0, 1\}\,,\ i = 1, \ldots, n;$$
$$\pi\,(x) = P\,(y = 1\,|x\,)\ \text{—?}$$

Пример 1

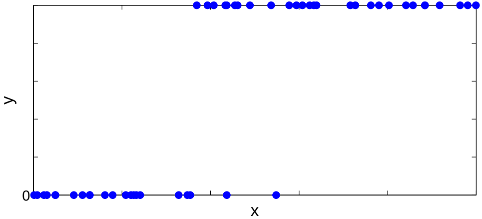
Разработка пестицидов: x_i — доза пестицида, y_i — смерть вредителя. Повторяемый эксперимент с фиксированными уровнями фактора:



$$\hat{\pi}(x) = \frac{\sum\limits_{i=1}^{n} y_i \left[x = x_i\right]}{\sum\limits_{i=1}^{n} \left[x = x_i\right]}.$$

Пример 2

Эконометрика, построение кривой спроса: x_i — цена товара, y_i — согласие купить товар.
Неповторяемый эксперимент со случайными уровнями фактора:



Можно построить непараметрическую оценку при помощи ядерного сглаживания:

$$\hat{\pi}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}.$$

Параметризация

Линейная регрессия:

$$\pi(x) = \theta_0 + \theta_1 x + \varepsilon.$$

- Оценка вероятности может выходить за $[0,1]$.
- В линейной регрессии $y = \mathbb{E}(y|x) + \varepsilon$, и МНК-оценка θ хороша, когда $\varepsilon \sim N(0,\sigma)$. Здесь же, если $y = \pi(x) + \varepsilon$, то $\varepsilon = 1 - \pi(x)$ или $\varepsilon = \pi(x)$, и МНК-оценка будет плохой.

Нужно такое нелинейное преобразование

$$g(\pi(x)) = \theta_0 + \theta_1 x + \varepsilon,$$

чтобы:

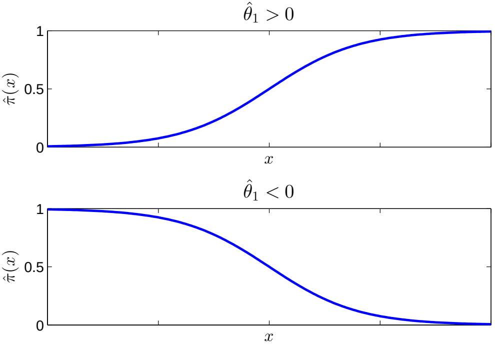
- $\hat{\pi}(x) = g^{-1}\left(\hat{\theta}_0 + \hat{\theta}_1 x\right)$ принимала значения из $[0,1]$;
- изменения на краях диапазона значений x приводили к меньшим изменениям $\pi(x)$ (x — годовой доход, y — покупка автомобиля, $\pi(10000000 + 200000) - \pi(10000000) <$ $\pi(500000 + 200000) - \pi(500000)$).

Параметризация

Logit:

$$g\left(\pi\left(x\right)\right) = \ln\frac{\pi\left(x\right)}{1-\pi\left(x\right)} = \theta_0 + \theta_1 x + \varepsilon,$$

$$\hat{\pi}\left(x\right) = \frac{e^{\hat{\theta}_0+\hat{\theta}_1 x}}{1+e^{\hat{\theta}_0+\hat{\theta}_1 x}}.$$



Относительный риск

Пусть $y \sim Ber(p)$, тогда **риск (odds)** события $y = 1$:

$$ODDS = \frac{p}{1-p}.$$

Если $y_1 \sim Ber(p_1)$, $y_2 \sim Ber(p_2)$, то **относительный риск (odds ratio)** события $y_1 = 1$ по сравнению с событием $y_2 = 1$:

$$OR = \frac{p_1/\left(1-p_1\right)}{p_2/\left(1-p_2\right)}.$$

Серд. заболевания \ Возраст	≥ 55	≤ 55
есть	21	22
нет	6	51

$$OR = \frac{21/6}{22/51} \approx 8.1.$$

Роль коэффициентов регрессии

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x.$$

Пусть $x = [\text{возраст} \geq 55]$, $y = [\text{есть сердечные заболевания}]$. По $\hat{\theta}_1$ легко оценить относительный риск получения заболевания пожилыми людьми:

$$\widehat{OR} = e^{\hat{\theta}_1}.$$

Пусть $x = \text{возраст}$, $y = [\text{есть сердечные заболевания}]$. $e^{\hat{\theta}_1}$ имеет смысл мультипликативного прироста риска получения заболевания при увеличении возраста на 1 год.

Настройка параметров

$\pi(x)$ оценивает $P(y=1\,|x)$,
$1-\pi(x)$ оценивает $P(y=0\,|x) \;\Rightarrow$
Вероятность $(x_i,1)$ — $\pi(x_i)$, $(x_i,0)$ — $1-\pi(x_i)$.

$$L(\theta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i},$$

$$LL(\theta)=-\sum_{i=1}^{n}\left(y_i\ln\pi(x_i)+(1-y_i)\ln\left(1-\pi(x_i)\right)\right),$$

$$\hat{\theta}=\operatorname*{argmax}_{\theta} LL(\theta).$$

$\hat{\theta}$ существует и единственна, находится методом Ньютона-Рафсона, является состоятельной и асимптотически эффективной оценкой θ, а также асимптотически нормальна.

Проблемы МП-оценки

$\hat{\theta}$ может не существовать или не быть конечной, если:

- наблюдения $y = 0$ и $y = 1$ линейно разделимы в пространстве признаков X;
- матрица X вырождена.

Итерационный процесс может не сойтись, если число признаков k слишком велико относительно числа наблюдений n.

Дисперсия оценок

Пусть $I(\theta) \in \mathbb{R}^{(k+1)\times(k+1)}$ — матрица вторых производных $LL(\theta)$:

$$\frac{\partial^2 LL}{\partial \theta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi(x_i)(1-\pi(x_i)),$$

$$\frac{\partial^2 LL}{\partial \theta_j \theta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi(x_i)(1-\pi(x_i)).$$

Другая форма записи:

$$I(\theta) = X^T V X,$$
$$V = \operatorname{diag}\left(\pi(x_1)(1-\pi(x_1)), \dots, \pi(x_n)(1-\pi(x_n))\right).$$

Из теории оценок максимума правдоподобия: $\mathbb{D}\hat{\theta} = I^{-1}\left(\hat{\theta}\right)$.

Доверительные интервалы

Для отдельного коэффициента θ_j:

$$\hat{\theta}_j \pm z_{1-\alpha/2}\sqrt{\left(I^{-1}\left(\hat{\theta}\right)\right)_{jj}}.$$

Для $g(x_0)$ — логита нового наблюдения x_0:

$$x_0\hat{\theta} \pm z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\theta}\right) x_0}.$$

Для вероятности $y = 1$ при $x = x_0$:

$$\left[\frac{e^{x_0\hat{\theta} - z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\theta}\right)x_0}}}{1 + e^{x_0\hat{\theta} - z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\theta}\right)x_0}}}, \frac{e^{x_0\hat{\theta} + z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\theta}\right)x_0}}}{1 + e^{x_0\hat{\theta} + z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\theta}\right)x_0}}}\right].$$

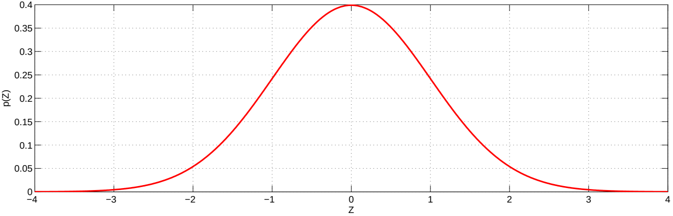
Критерий Вальда

нулевая гипотеза: $H_0: \theta_j = 0;$

альтернатива: $H_1: \theta_j <\neq> 0;$

статистика: $T = \dfrac{\hat{\theta}_j}{\sqrt{\left(I^{-1}(\hat{\theta})\right)_{jj}}};$

$T \sim N(0,1)$ при H_0;



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - ncdf(t,0,1), & H_1: \theta_j > 0, \\ ncdf(t,0,1), & H_1: \theta_j < 0, \\ 2\left(1 - ncdf(|t|,0,1)\right), & H_1: \theta_j \neq 0. \end{cases}$$

Критерий, основанный на правдоподобии

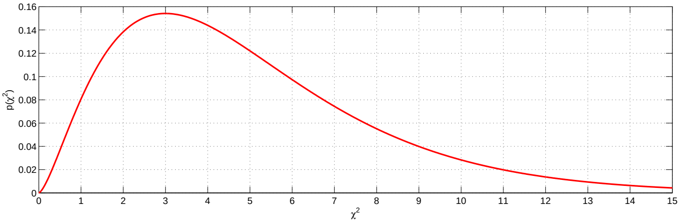
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2} \right); \quad \underset{(k+1)\times 1}{\theta^T} = \left(\underset{(k+1-k_1)\times 1}{\theta_1^T}, \underset{k_1\times 1}{\theta_2^T} \right)^T;$$

нулевая гипотеза: $H_0\colon \theta_2 = 0;$

альтернатива: $H_1\colon H_0$ неверна;

статистика: $G = 2\left(LL_{ur} - LL_r\right);$

$G \sim \chi^2_{k_1}$ при H_0;



достигаемый уровень значимости:

$$p(g) = 1 - chi2cdf(g, k_1).$$

Значимость признаков

Поскольку распределение $\hat{\theta}$ нормально только асимптотически, точность критериев невысока.

Поэтому значимость признаков рекомендуется проверять на уровне $\alpha = 0.25$.

Пошаговая логистическая регрессия

- **Шаг 0**. Настраивается модель с одной только константой, а также все модели с одной переменной. Рассчитывается статистика Вальда каждой модели и достигаемый уровень значимости. Выбирается модель с наименьшим достигаемым уровнем значимости. Соответствующая переменная X_{e1} включается в модель, если этот достигаемый уровень значимости меньше порогового значения $p_E = 0.15$.
- **Шаг 1**. Рассчитывается статистика Вальда и достигаемый уровень значимости для всех моделей, содержащих две переменные, одна из которых X_{e1}. Аналогично принимается решение о включении X_{e2}.
- **Шаг 2**. Если была добавлена переменная X_{e2}, возможно, X_{e1} уже не нужна. В общем случае просчитываются все возможные варианты исключения одной переменной, рассматривается вариант с наибольшим достигаемым уровнем значимости, соответствующая переменная исключается, если он превосходит пороговое значение $p_R = 0.2$.
- ...

Мультиколлинеарность

Признаки мультиколлинеарности:

- правдоподобие модели высоко, но оценки многих коэффициентов близки к своим стандартным отклонениям;
- коэффициенты сильно меняются при включении и исключении других признаков.

Порог классификации

Как по $\pi(x)$ оценить y?

$$y = [\pi(x) \geq p_0].$$

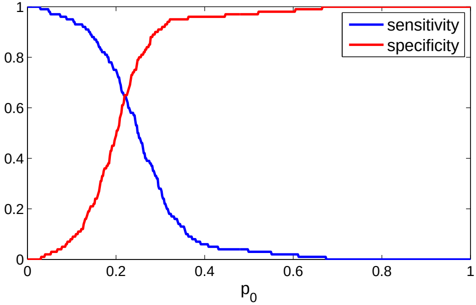
Чаще всего берут $p_0 = 0.5$, но можно выбирать по другим критериям, например, для достижения заданных показателей чувствительности или специфичности.

Пример

Эффективность терапии для наркозависимых, $p_0 = 0.5$:

$\hat{y}$ \ y	1	0
1	16	11
0	131	417

Чувствительность: $\frac{16}{16+131} \approx 10.9\%$.
Специфичность: $\frac{417}{11+417} \approx 97.4\%$.



Прикладная статистика
10. Логистическая регрессия.

Рябенко Евгений
riabenko.e@gmail.com