Прикладная статистика и анализ данных

Съезд III

Пропуски в данных

Неслучайные пропуски

Возможные причины:

- Отказ респондента дать ответ на вопрос:

  Сколько вы тратите в месяц на развлечения? ***<не скажу>***

  <не скажу> — важная информация!!!

- Признак не применим:

  Посещаете ли вы рестораны? ***Нет***

  Какой ваш средний чек? ***<NaN>***

В таких случаях:

1. Создается новый признак $x_{j'} = I\{x_j \text{ пропущено}\}$;
2. Пропуски заменяются на любую константу,

   не встречающуюся в данных.

Случайные пропуски

Возможные причины:

- Сломалось оборудование:

Время	8:00	9:00	10:00	11:00	12:00
Температура возд.	21.4	22.1	**NaN**	24.2	25.5

- Было лень:

⟵ ведро лени

Случайные пропуски

Варианты:

- Удалить строки, содержащие пропуски; **df.dropna**
- Заполнить пропуски:
  - по ближайшему соседу;
  - среднее/медиана; **df.fillna**
  - Интерполяция (временные ряды); **df.interpolate**
  - EM-алгоритм;
- Считать X^TX и X^TY только по полным парам

$$\frac{1}{n}\left(X^TX\right)_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, x_{ik} \approx \frac{1}{n_{jk}}\sum_{i=1}^{n} x_{ij}\, x_{ik}\, I\{x_{ij} \text{ и } x_{ik} \text{ не пропущены}\},$$

$$\frac{1}{n}\left(X^TY\right)_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, y_{i} \approx \frac{1}{n_j}\sum_{i=1}^{n} x_{ij}\, y_{i}\, I\{x_{ij} \text{ не пропущено}\}.$$

где n_{jk} — число полных пар (x_{ij}, x_{ik}); n_j — число заполненных x_{ij}.

Робастная регрессия

Метод Тейла

$$y = a + bx$$

$$\widehat{b} = \text{med}\left\{\frac{Y_j - Y_i}{x_j - x_i}, \quad 1 \leqslant i < j \leqslant n\right\}$$

$$\widehat{a} = \text{med}\left\{Y_i - \widehat{b}x_i, \quad i = 1, ..., n\right\}$$

Робастная оптимизация

$$\sum_{i=1}^{n} R\left(Y_i - x_i^T\theta\right) \to \min_{\theta \in \mathbb{R}^d}$$

- $R(x) = x^2$ — L_2-норма, получаем обычный МНК;
- $R(x) = |x|$ — L_1-норма;
- $R(x) = \frac{x^2}{2} I\{|x| < c\} + c(|x| - c/2) I\{|x| > c\}$ — Хубер;

Методы на основе ближайшего соседа

Метод ближайших соседей (kNN)

Пусть $\mathscr{X}$ — метрическое пространство.

$x_1, ..., x_n \in \mathscr{X}$ — обучающая выборка.

$Y_1, ..., Y_n$ — соответствующая целевая переменная.

Предположение:
свойства объекта меняются не сильно в его окрестности.

Тогда давайте смотреть на свойства k ближайших соседей.

Примеры.
Пусть $x \in \mathscr{X}$ — исследуемый объект.
$x_{(1)}, ..., x_{(k)}$ — k его соседей в порядке удаления от x.

1. *Классификация.*
   Предсказание — наиболее часто встречаемый класс.

2. *Регрессия.*
   Предсказание — усреднение отклика по соседям.

Взвешенный метод ближайших соседей

Пусть $x \in \mathscr{X}$ — исследуемый объект.

$x_{(1)}, ..., x_{(k)}$ — k его соседей в порядке удаления от x.

$Y_1, ..., Y_k$ — соответствующий отклик.

$w_1, ..., w_k$ — вклад k-го соседа, определяемый пользователем.

Способы определения веса:

- $w_j = 1 - j/k$ — зависящий от номера соседа
- $w_j = \|x - x_{(j)}\|^{-1}$ — зависящий от расстояния до соседа

$$\widehat{y}(x) = \arg\max_y \sum_{j=1}^{k} w_j I\{Y_j = k\} \text{ — Классификация}$$

$$\widehat{y}(x) = \frac{\sum_{j=1}^{k} w_j Y_j}{\sum_{j=1}^{k} w_j} \text{ — Регрессия}$$

Свойства

1. k — гиперпараметр модели;
2. Не редко на практике показывает хорошие результаты.
3. **Дорогое применение:**
   для каждого x результат вычисляется за $O(n \ln n)$.

Способы решения проблемы:

- Хранение данных в виде дерева: K-D Tree, Ball Tree
- Приближенный поиск: Locality-sensitive hashing —

  вероятностный метод понижения размерности многомерных данных.
  Подбирает хеш-функций так, чтобы похожие объекты
  с высокой степенью вероятности попадали в одну корзину.

https://github.com/facebookresearch/faiss — реализации методов

Непараметрическая регрессия

Метод локального усреднения

$$\widehat{y}(x) = \sum_{i=1}^{n} w_i(x) Y_i \Big/ \sum_{i=1}^{n} w_i(x),$$

где $w_i(x) \geqslant 0$ убывает при удалении x от X_i.

Варианты:

1. $w_i(x) = I\{|x - X_i| \leqslant c\}$ — усреднение по окрестности x;
2. $w_i(x) = (c - |x - X_i|)^k \, I\{|x - X_i| \leqslant c\}$

   — взвешенное усреднение по окрестности x;
3. Усреднение по k ближайшим соседям;
4. Ядерная оценка (см. далее).

Ядерная оценка Надарая-Ватсона

$$w_i(x) = \frac{1}{h} q\left(\frac{x - X_i}{h}\right),$$

где q — ядро (симметричная плотность),

$h > 0$ — ширина ядра.

Вероятностная интерпретация

Пусть $X_1, ..., X_n$ случайны и $Y_i = f(X_i, \varepsilon_i)$ — отклик. Тогда

$\frac{1}{n}\sum\limits_{i=1}^{n} w_i(x) = \frac{1}{nh}\sum\limits_{i=1}^{n} q\left(\frac{x-X_i}{h}\right)$ — ядерная оценка плотности X;

$\frac{1}{n}\sum\limits_{i=1}^{n} w_i(x) Y_i = \frac{1}{nh}\sum\limits_{i=1}^{n} Y_i \cdot q\left(\frac{x-X_i}{h}\right)$ — "ядерное мат. ожид." $\mathrm{E}Y$;

$\widehat{y}(x) = \sum\limits_{i=1}^{n} w_i(x) Y_i \Big/ \sum\limits_{i=1}^{n} w_i(x)$ — "ядерное УМО" $\mathrm{E}(Y|X)$.

Ядерная оценка: теорема о сходимости

Пусть

1. $\int\limits_{\mathbb{R}} |q(y)|\,dy < \infty$;
2. $yq(y) \to 0$ при $|y| \to \infty$;
3. $\mathsf{E}Y^2 < \infty$;
4. $h_n \to 0, nh_n \to \infty$ при $n \to \infty$.

Тогда $\widehat{y}(x) \xrightarrow{\mathsf{P}} y(x)$ в точках непрерывности функции $f(x)$, плотности $p_X(x)$ и условной дисперсии $\sigma^2(x) = \mathsf{D}(Y|X = x)$, если при этом $p(x) > 0$.

Наилучшая скорость сходимость квадратичного риска достигается при $h \sim n^{-1/5}$

Ядерная оценка: выбор ширины ядра

Функционал вида leave one out:

$$F(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{y}_{-i}(x_i)\right)^2,$$

где $\widehat{y}_{-i}(x)$ — ядерная оценка, построенная по выборке, из которой было исключено i-е наблюдение.

Утверждение

$$F(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{y}(x_i)\right)^2 \Bigg/ \left(1 - \frac{q(0)}{\sum_{k=1}^{n} q\left(\frac{x_i - x_k}{h}\right)}\right)$$

Выбор h: $F(h) \to \min\limits_{h}$

Ядерная оценка: доверительная лента

Предположения:

$\mu(x)$ — ожидаемый отклик;

$Y_i = \mu(x_i) + \varepsilon_i$ — наблюдаемый отклик, $\mathrm{E}\varepsilon_i = 0, \mathrm{D}\varepsilon_i = \sigma^2$;

q — гауссовское ядро.

Доверительная лента уровня доверия α:

$$\Big(\widehat{y}(x) - z_{3h}\delta(x),\ \ \widehat{y}(x) + z_{3h}\delta(x)\Big)$$

$$\delta(x) = \widehat{\sigma}\sqrt{\sum_{i=1}^{n} w^2(x_i)}\Big/\sum_{i=1}^{n} w(x_i), \qquad p = \frac{1+\alpha^{1/3h}}{2},$$

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(Y_{i+1} - Y_i)^2$$

Пояснение: $\mathrm{D}(Y_{i+1} - Y_i) = \mathrm{D}(\varepsilon_{i+1} - \varepsilon_i) = \mathrm{D}\varepsilon_{i+1} + \mathrm{D}\varepsilon_i = 2\sigma^2$

Ширина окна

$h = const$ не ок в случае "где-то пусто, а где-то густо".

Тогда $h(x) = ||x - X_{(k)}||$, где $X_{(k)}$ — k-й ближайший сосед для x.

Выбросы

Оценка Надарая-Ватсона крайне неустойчива к выборсам.

См. алгоритм LOWESS.

Локальная линейная регрессионная модель

Модель $f(x) = x^T\theta(x)$

Для каждого x применяется взвешенный МНК:

$$\sum_{i=1}^{n} w_i(x)\left(Y_i - X_i^T\theta(x)\right)^2 \longrightarrow \min_{\theta(x)}$$

где $w_i(x) = q_{h(x)}(x - X_i)$.

Взвешенный МНК:

$$\widehat{\theta}(x) = \left(X^T W(x) X\right)^{-1} X^T W(x) Y$$

$$W(x) = diag\left(w_1^2(x), ..., w_n^2(x)\right)$$

Проверка линейности логита в логистической регрессии

Сглаженные диаграммы рассеяния

Пусть $(x_1, Y_1), \ldots, (x_n, Y_n)$ — обучающая выборка, где $Y_i \in \{0, 1\}$.

Выберем признак j и построим ядерную регрессию $y \sim x_j$:

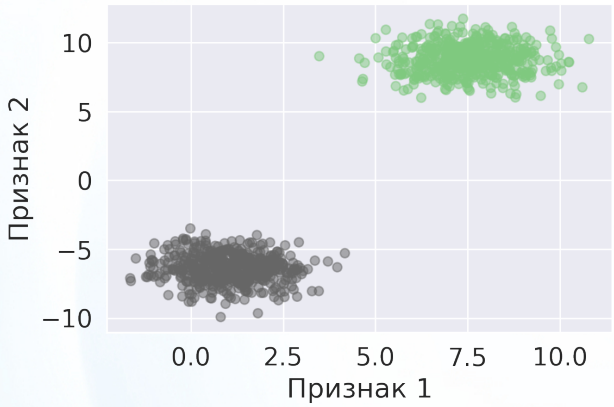
$$\widehat{y}(x_j) = \sum_{i=1}^{n} q\left(\frac{x_j - x_{ij}}{h}\right) Y_i \Bigg/ \sum_{i=1}^{n} q\left(\frac{x_j - x_{ij}}{h}\right),$$

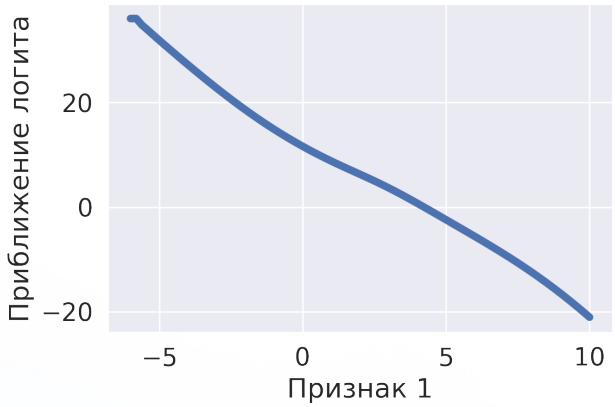
Эта регрессия — приближение вер-ти класса 1 в зависимости от x_j.

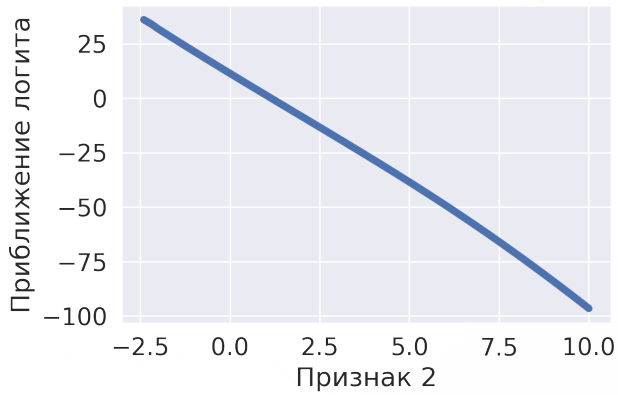
Отсюда делаем приближение логита:

$$\mathit{logit}(x_j) = \log \frac{\widehat{y}(x_j)}{1 - \widehat{y}(x_j)}.$$

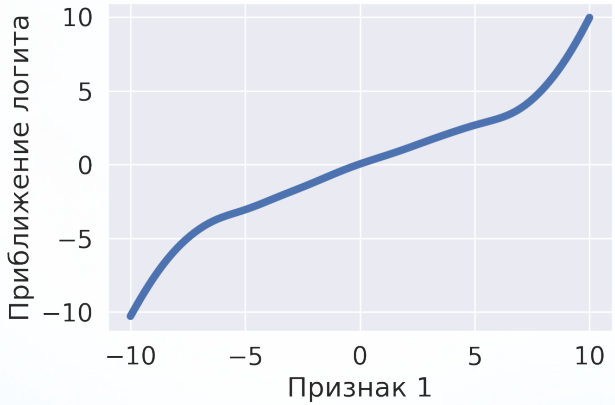
Проверка: график $\mathit{logit}(x_j)$ похож на прямую.

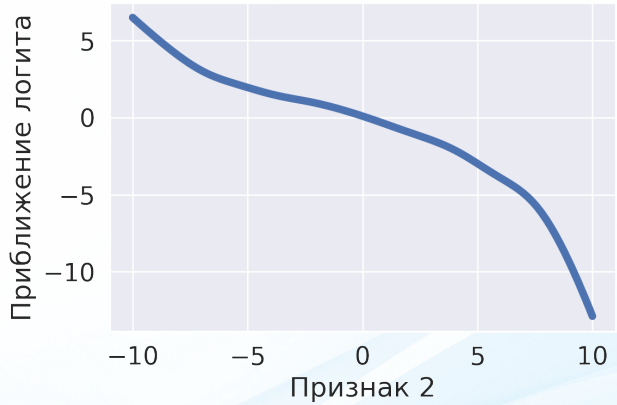

Признак 2
Признак 1

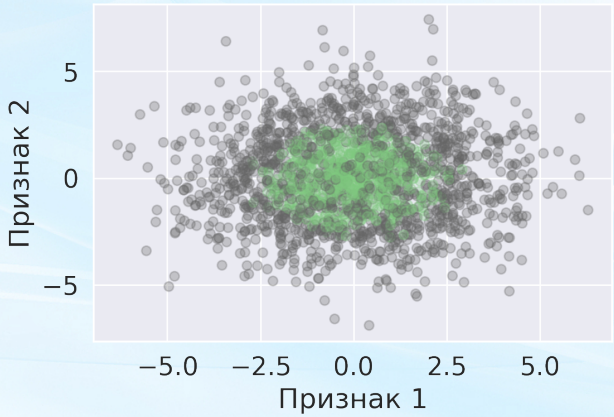

Приближение логита
Признак 1


Приближение логита
Признак 2

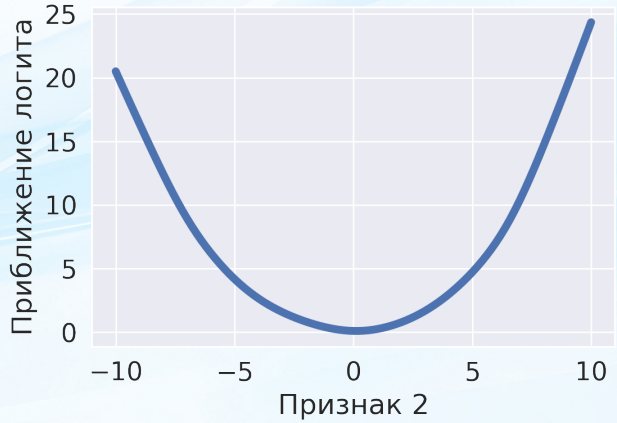

Признак 2
Признак 1

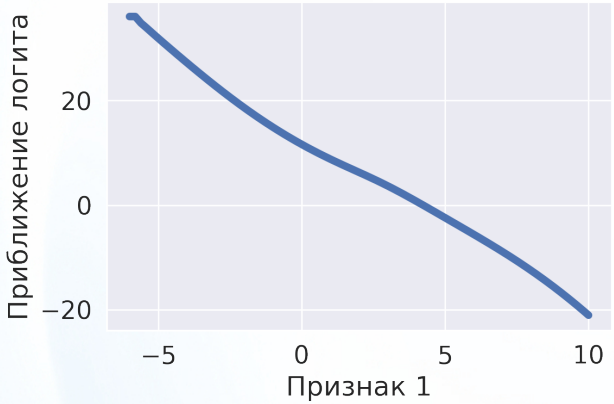

Приближение логита
Признак 1

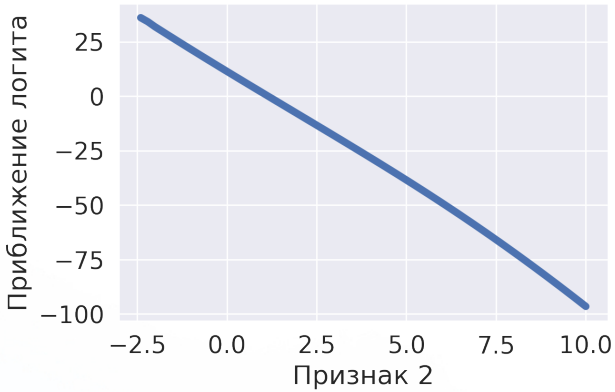

Приближение логита
Признак 2


Признак 2
Признак 1


Приближение логита
Признак 1

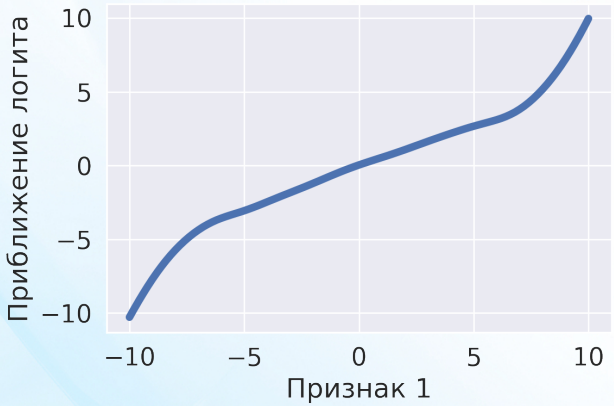

Приближение логита
Признак 2

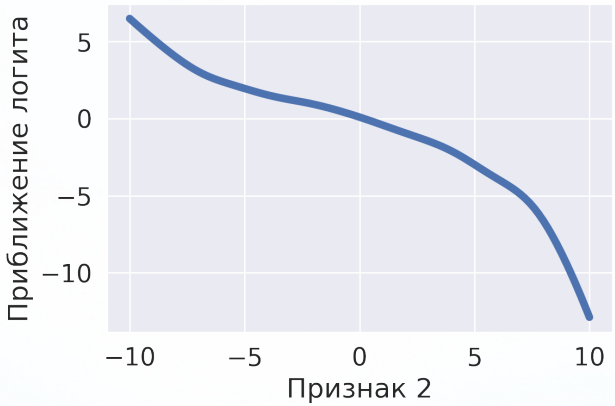




Логит линеен,

все хорошо

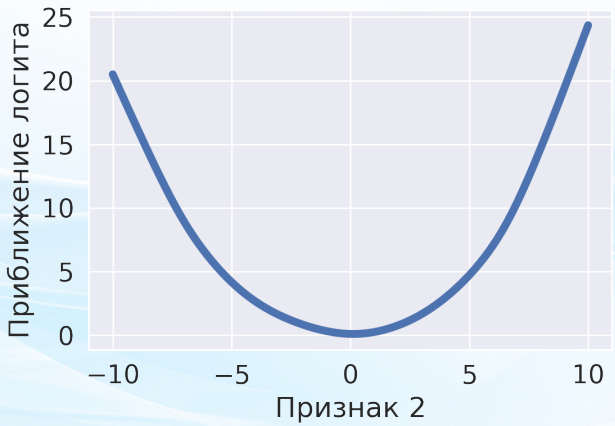




Классы линейно

разделимы, но зависи-

мость нелинейна



Классы не являются

линейно разделимыми


ВСЁ!