Машинное обучение. Метрические методы

Алексей Колесов

Белорусский государственный университет

8 ноября 2017 г.

Метод ближайших соседей

- определение ответа для объекта по ответам близких объектов
- гипотеза компактности и непрерывности
- lazy learning

Содержание

Метод ближайших соседей (k-NN)

- пусть на X определена метрика ρ
- для $X = \mathbb{R}^d$ естественна $\rho(x, x') = \sqrt{\sum_{i=1}^{d} (x_i - x'_i)^2}$
- пусть $\pi_1(x), \ldots, \pi_m(x)$ — объекты S, упорядоченные по $\rho(x, x_i)$

Алгоритм 1 k-NN

Вход: $S = ((x_1, y_1), \ldots, (x_m, y_m))$; $y_i \in [k]$

1: **return** $h(x) \in \operatorname{argmax}_{y \in Y} |\{y_{\pi_i(x)} = y : i \leqslant k\}|$

Для $k = 1$ получаем $h_S(x) = y_{\pi_1(x)}$

k-NN для регрессии

- для задач регрессии естественней: $h_S(x) = \frac{1}{k}\sum_{i=1}^k y_{pi_i(x)}$
- также часто применяют:

$$h_S(x) = \sum_{i=1}^{k} \frac{\rho\left(x, x_{\pi_i(x)}\right)}{\sum_{j=1}^k \rho\left(x, x_{\pi_j(x)}\right)} y_{\pi_i(x)}$$

Содержание

Анализ

- асимптотически k-NN приближается оптимальному решающему правилу (при «хороших» распределениях)
- мы хотим анализ для конечного m
- покажем, что можно достичь оценки $2L_D(h^*) + \epsilon$, где ϵ зависит от m

Предположения и обозначения

- задача бинарной классификации с $0 - 1$ функцией потерь
- $X = [0, 1]^d$, ρ-евклидова
- D_x — маргинальное распределение x, $\eta(x) = \mathbb{P}[y = 1|x]$
- напомним, что $h^*(x) = 1_{\eta(x)>1/2}$
- предположим, что $|\eta(x) - \eta(x')| \leqslant c||x - x'||$

Оценка для 1-NN

Оценка для 1-NN

Пусть $S = (x_1, y_1), \ldots, (x_m, y_m)$, h_S — 1-NN гипотеза, h^* — оптимальное решающее правило. Тогда:

$$\mathop{\mathbb{E}}_{S \sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + c \mathop{\mathbb{E}}_{S \sim D^m, x \sim D}[||x - x_{\pi_1(x)}||]$$

Доказательство

Хотим:

$$\mathop{\mathbb{E}}_{S \sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + c \mathop{\mathbb{E}}_{S \sim D^m, x \sim D}[||x - x_{\pi_1(x)}||]$$

- $L_D(h_S) = \mathbb{E}_{(x,y) \sim D}[1_{h_S(x) \neq y}]$
- $L_D(h_S)$ — вероятность получить S, затем ещё (x, y), такой что класс $\pi_1(x)$ не равен y
- $L_D(h_S)$ — вероятность получить $S_x = (x_1, \dots, x_m)$, затем $x \sim D_x$, затем найти $\pi_1(x)$ и выбрать $y \sim \eta(x)$ и $y_{\pi_1(x)} \sim \eta(\pi_1(x))$, так чтоб метки не совпали

Доказательство

Хотим:

$$\mathop{\mathbb{E}}_{S\sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + c \mathop{\mathbb{E}}_{S\sim D^m, x\sim D}[||x - x_{\pi_1(x)}||]$$

Распишем:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S}[L_D(h_S)] &= \mathop{\mathbb{E}}_{S_x\sim D_x^m, x\sim D_x, y\sim\eta(x), y'\sim\eta(\pi_1(x))}[1_{y\neq y'}] && (1)\\
&= \mathop{\mathbb{E}}_{S_x\sim D_x^m, x\sim D_x}\left[\mathop{\mathbb{P}}_{y\sim\eta(x), y'\sim\eta(\pi_1(x))}[y \neq y']\right] && (2)
\end{aligned}$$

Доказательство

Хотим:

$$\mathop{\mathbb{E}}_{S\sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + c \mathop{\mathbb{E}}_{S\sim D^m, x\sim D}[||x - x_{\pi_1(x)}||]$$

Распишем:

$$\begin{aligned} \mathop{\mathbb{P}}_{y\sim\eta(x), y'\sim\eta(x')}[y \neq y'] &= \eta(x')(1-\eta(x)) + (1-\eta(x'))\eta(x) && (3) \\ &= 2\eta(x)(1-\eta(x)) + (\eta(x) - \eta(x'))(2\eta(x) - 1) && (4) \\ &\leqslant 2\eta(x)(1-\eta(x)) + c||x - x'|| && (5) \end{aligned}$$

Доказательство

Хотим:

$$\mathop{\mathbb{E}}_{S\sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + c \mathop{\mathbb{E}}_{S\sim D^m, x\sim D}[||x - x_{\pi_1(x)}||]$$

Получили:

$$\mathop{\mathbb{E}}_{S}[L_D(h_S)] \leqslant \mathop{\mathbb{E}}_{x}[2\eta(x)(1-\eta(x))] + c \mathop{\mathbb{E}}_{S,x}[||x - x_{\pi_1(x)}||]$$

Заметим:

$$L_D(h^*) = \mathop{\mathbb{E}}_{x}[\min\{\eta(x), 1-\eta(x)\}] \geqslant \mathop{\mathbb{E}}_{x}[\eta(x)(1-\eta(x))]$$

Лемма

Безымянная лемма

Пусть $C_1, \ldots, C_r$ множество подмножеств X. Пусть S — множество из m объектов из X, выбранных независимо с одинаковым распределением. Тогда:

$$\mathop{\mathbb{E}}_{S \sim D^m}\left[\sum_{i: C_i \cap S = \varnothing} \mathbb{P}[C_i]\right] \leqslant \frac{r}{me}$$

Доказательство

Хотим:

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[\sum_{i:C_i\cap S=\varnothing}\mathbb{P}[C_i]\right]\leqslant\frac{r}{me}$$

- $\mathop{\mathbb{E}}_{S\sim D^m}\left[\sum_{i:C_i\cap S=\varnothing}\mathbb{P}[C_i]\right]=\sum_{i=1}^{r}\mathbb{P}[C_i]\,\mathbb{P}[C_i\cap S=\varnothing]$
- $\mathbb{P}[C_i\cap S=\varnothing]=(1-P(C_i))^m\leqslant e^{-P[C_i]m}$
- $\mathop{\mathbb{E}}_{S\sim D^m}\left[\sum_{i:C_i\cap S=\varnothing}\right]\leqslant\sum_{i=1}^{r}\mathbb{P}[C_i]e^{-\mathbb{P}[C_i]m}\leqslant r\max_i\mathbb{P}[C_i]e^{-\mathbb{P}[C_i]m}$
- $\max_\alpha \alpha e^{-m\alpha}\leqslant\frac{1}{me}$

Оценка для 1-NN

Оценка для 1-NN

Пусть $S = (x_1, y_1), \ldots, (x_m, y_m)$, h_S — 1-NN гипотеза, h^* — оптимальное решающее правило. Тогда:

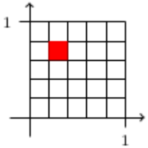
$$\mathop{\mathbb{E}}_{S \sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + 4c\sqrt{d}m^{-\frac{1}{d+1}}$$

Доказательство

- возьмём целое число T и положим $\epsilon = 1/T$, и положим сетку из $r = T^d$ кубиков со стороной ϵ $(C_1, \dots, C_r)$
- для любых x, x' из одного кубика: $||x - x'|| \leqslant \sqrt{d}\epsilon$
- для любых x, x': $||x - x'|| \leqslant \sqrt{d}$

$$\mathop{\mathbb{E}}_{x,S}[||x - x_{\pi_1(x)}||] \leqslant \mathop{\mathbb{E}}_{S}\left[\mathbb{P}\left[\bigcup_{i: C_i \cup S = \varnothing}\right]\sqrt{d} + \mathbb{P}\left[\bigcup_{i: C_i \cup S \neq \varnothing}\right]\epsilon\sqrt{d}\right]$$

$$\mathop{\mathbb{E}}_{x,S}[||x - x_{\pi_1(x)}||] \leqslant \sqrt{d}\left(\frac{r}{me} + \epsilon\right)$$



Доказательство

Имеем:

$$\mathop{\mathbb{E}}_{x,S}[||x - x_{\pi_1(x)}||] \leqslant \sqrt{d}\left(\frac{r}{me} + \epsilon\right)$$

Подставив $r = (1/\epsilon)^d$ и $\epsilon = 2m^{-1/(d+1)}$, получим:

$$\mathop{\mathbb{E}}_{S \sim D^m}[L_D(h_S)] \leqslant 2L_D(h^*) + 4c\sqrt{d}m^{-\frac{1}{d+1}}$$

Проклятие размерности

- $4c\sqrt{d}m^{-\frac{1}{d+1}}$ растёт с c и d
- чтоб получить оценку $\leqslant \epsilon$, необходимо выбрать $m \geqslant (4c\sqrt{d}/\epsilon)^{d+1}$
- размер выборки должен расти **экспоненциально** с размерностью пространства

NoFLT

NoFLT для липшицевых распределений

Для любого $c > 1$ и решающего правила L существует распределение $[0, 1]^d \times 0, 1$, такое что $\eta(x)$ является c-липшицевой, ошибка оптимального решающего правила равна нулю, но для размера выборки $m \leqslant (c + 1)^d/2$ ошибка L не меньше $1/4$

Доказательство: если выбрать решётку с шагом $1/c$ (из $(c + 1)^d$ точек), то любая функция распределения будет c-липшицевой $\Rightarrow$ NoFLT.

Применение

- наивный алгоритм работает за $\mathcal{O}(dm)$
- алгоритмы из вычислительной геометрии требуют $o(d^{\mathcal{O}(1)} \log(m))$, но занимают $m^{\mathcal{O}(d)}$ памяти
- чаще всего применяют приближённые алгоритмы (local-sensitive hashing)

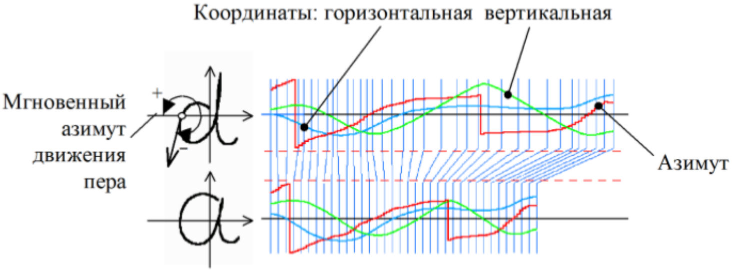
Содержание

Предположения

- гипотеза компактности: близкие объекты, как правило, лежат в одном классе
- гипотеза непрерывности: близким объектам соответствуют близкие ответы
- близкие объекты, у которых $\rho(x, x')$ мало

Примеры расстояний

- для сайтов в интернете: минимальное количество переходов по ссылкам
- для текстов: редакторское расстояние Левенштейна
- для сигналов: энергия сжатия



Обобщённый метрический классификатор

Обобщённый метрический классификатор:

$$h_S(x) = \underset{y \in Y}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{m} 1_{[y_{\pi_i(x)}=y]} w(i,x)}_{\Gamma_y(x)}$$

- $w(i,x)$ – степень важности i-го соседа
- $\Gamma_y(x)$ — оценка близости объекта x к классу y

1-NN

$w(i, x) = 1_{i=1}$

Преимущества:

- простота реализации
- интерпретируемость решения (вывод на основе прецедентов)

Недостатки:

- неустойчивость к шуму
- отсутствие настраиваемых параметров
- низкое качество классификации
- необходимо хранить всю выборку

k-NN

$w(i, x) = 1_{i \leqslant k}$

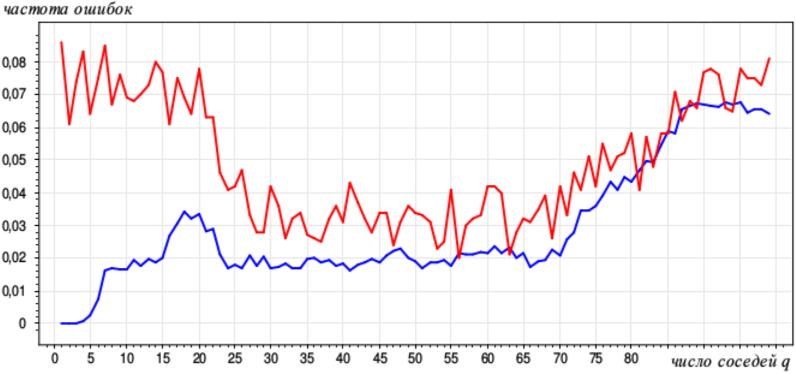
Преимущества:

- менее чувствителен к шуму
- возможность настраивать k

Недостатки:

- неоднозначность классификации $\Gamma_y(x) = \Gamma_{y'}(x)$

Пример


частота ошибок
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
число соседей q

Пример


частота ошибок на обучении и контроле (исключая и не исключая себя)
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400
число соседей k

Метод k взвшенных соседей

$w(i, x) = 1_{i \leqslant k} w_i$

Эвристики:

- $w_i = \frac{k+1-i}{k}$
- $w_i = q^i$ для $0 < q < 1$

Вопросы:

- как выбрать, какие веса использовать?
- может лучше, чтоб w_i зависел от $\rho(x, x')$, а не i?

Метод парзеновского окна

$$w(i,x) = K\left(\frac{\rho(x, x_{\pi_i(x)})}{h}\right)$$

$K(r)$ — ядро, не возрастает и положительно на $[0,1]$

Метод парзеновского окна фиксированной ширины:

$$h_{S,w,K}(x) = \operatorname*{argmax}_{y\in Y} \sum_{i=1}^{m} 1_{[y_{\pi_i(x)}=y]} K\left(\frac{\rho(x, x_{\pi_i(x)})}{w}\right)$$

Метод парзеновского окна переменной ширины:

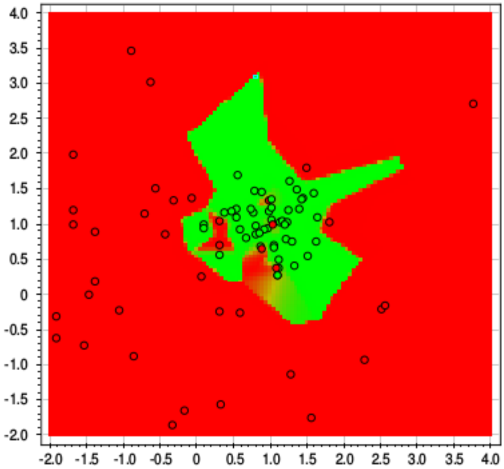
$$h_{S,k,K}(x) = \operatorname*{argmax}_{y\in Y} \sum_{i=1}^{m} 1_{[y_{\pi_i(x)}=y]} K\left(\frac{\rho(x, x_{\pi_i(x)})}{\rho(x, x_{\pi_{k+1}(x)})}\right)$$

Надо выбирать:

- ширину окна
- ядро

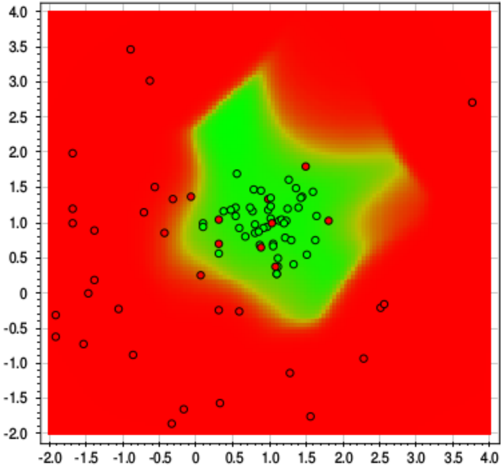
Пример

0.05


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
-2.0
-2.0 -1.5 -1.0 -0.5 0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

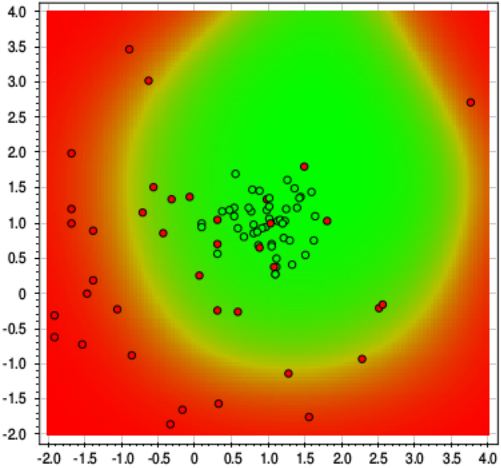
Пример

0.3


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
-2.0
-2.0 -1.5 -1.0 -0.5 0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

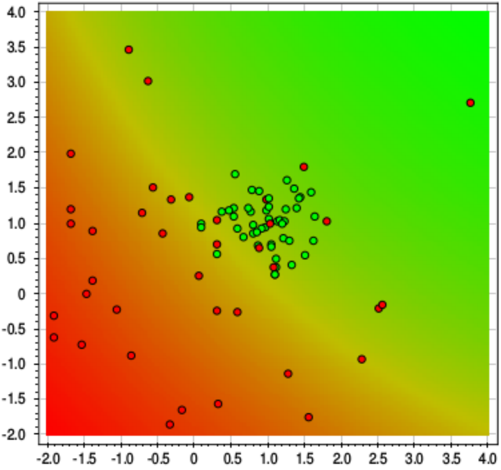
Пример

1.0


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
-2.0
-2.0 -1.5 -1.0 -0.5 0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

Пример

5.0


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
-2.0
-2.0 -1.5 -1.0 -0.5 0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0

Метод потенциальных функций

$$w(i, x) = \gamma_i K\left(\frac{\rho(x, x_{\pi_i(x)})}{h_i}\right)$$

Классификатор:

$$h_S(x) = \operatorname*{argmax}_{y \in Y} \sum_{i=1}^{m} 1_{[y_{\pi_i(x)} = y]} \gamma_i K\left(\frac{\rho(x, x_{\pi_i(x)})}{h_i}\right)$$

Аналогия:

- γ_i — величина заряда
- h_i — радиус действия
- y_i — знак заряда

Метод потенциальных функций — линейный классификатор

$$w(i, x) = \gamma_i K\left(\frac{\rho(x, x_{\pi_i(x)})}{h_i}\right)$$

Классификатор:

$$h_S(x) = \operatorname*{argmax}_{y \in Y} \sum_{i=1}^{m} 1_{[y_{\pi_i(x)} = y]} \gamma_i K\left(\frac{\rho(x, x_{\pi_i(x)})}{h_i}\right) = \operatorname{sign} \sum_{i=1}^{m} \gamma_i y_i K_i$$

Аналогия:

- $y_i K_i$ — признаковое описание
- γ_i — линейные коэффициенты

Содержание

Итоги

- разобрали метод ближайших соседей
- рассмотрели обобщения метода ближайших соседей
- доказали оценки для метода ближайших соседней ($k = 1$)

Литература

- Shai Shalev-Shwartz and Shai Ben-David — Understanding Machine Learning: From theory to algorithms (глава 19)
- К.Воронцов — метрические методы