Машинное обучение. Задание 1

Сергей Володин, 374 гр.

19 марта 2016 г.

Задача 1

Пусть $x \in \mathbb{R}^2$. Для этой точки упорядочим объекты выборки x_i по увеличению $\rho(x, x_i)$: $x^{(1)}, ..., x^{(6)}$. $+1$ — синий класс. $y_1 = +1$ Алгоритм классификации: $a(x, X^l) = \underset{y \in \{-1,+1\}}{\operatorname{argmax}} \sum\limits_{i=1}^{l} [y^{(i)} = y][i \leqslant k]$. Точка x на границе классов $\Leftrightarrow \sum\limits_{i=1}^{k} [y^{(i)} = -1] = \sum\limits_{i=1}^{k} [y^{(i)} = +1]$.

1. Пусть $k > 1$. Расмотрим последовательность $y^{(1)}, ..., y^{(k)}$. Поскольку $k \geqslant 2$, в ней должно быть не менее 2 элементов класса $+1$, что невозможно (их всего 1). Значит, границе не принадлежит ни одна точка, т.е. всё $\mathbb{R}^2$ классифицируется как -1.
2. Пусть $k = 1$. Точка лежит на границе $\Leftrightarrow \min\limits_{i \in \overline{2,6}} ||x - x_i|| = ||x - x_1||$. Получаем ломаную на плоскости *(дописать)*

Задача 2

	Q_E	Q_G	Q_H
Правило 1	x	89500	0.1709
Правило 2	x	109500	0.3219
best	x	Правило 2	Правило 2

1. Индекс Джини $Q_G(x) = \#\{(x_i, x_j) \colon i \neq j, x(x_i) = x(x_j), y(x_i) = y(x_j)\}$. Для первого правила $Q_G(x^1) = 200 \cdot 199 \cdot 2 + 100 \cdot 99 = 89500$, для второго $Q_G(x^2) = 100 \cdot 99 \cdot 2 + 300 \cdot 299 = 109500$
2. Энтропийный (для класса c и правила x и выборки длины l). $h(q) \stackrel{\text{def}}{=} -q \log_2 q - (1 - q) \log_2 (1 - q)$. $P = \#\{x_i = c\}$. $p = \#\{x_i \colon x(x_i) = 1, y_i = c\}$, $n = \#\{x_i \colon x(x_i) = 1, y_i \neq c\}$. $Q_H(x) = h(\frac{P}{l}) - \frac{p+n}{l} h(\frac{p}{p+n}) - \frac{l-p-n}{l} h(\frac{P-p}{l-p-n})$. В нашем случае $P = 200$, $l = 500$. Для первого правила (считаем, что оно предсказвыет первый класс) $p = 200$, $n = 200$. $Q_H(x^1) \approx 0.1709$, Для второго правила $p = 100$, $n = 0$, $Q_H(x^2) \approx 0.3219 \Rightarrow$ берем второе.
3. Что такое Q_E???

Задача 3

1. Выборка разделима при всех h: гиперплоскость $(x, w) - w_0$ при $w = (0, 1) - \frac{1}{2}$.
2. Картинка не соответствует условию. Какое правильное условие???

Задача 4а

Рассмотрим $K(x, y) - K(y, x) = (y + x, 2y + x) - (x + y, 2x + y) = (x + y, y - x) \neq 0$ в случае $x = 0$, $y \neq 0$. Получаем, что функция K не симметрична $\Rightarrow$ не ядро.

Задача 4а

$K(x, y) \stackrel{\text{def}}{=} \underbrace{\operatorname{ch}(x, y)}_{K_1(x,y)} + 3 \underbrace{\operatorname{sh}(x, y)}_{K_2(x,y)}$

1. . Докажем, что K_1, K_2 — ядра. Функции $\operatorname{ch} t$ и $\operatorname{sh} t$ разлагаются в сходящийся на $\mathbb{R}$ ряд с неотрицательными коэффициентами, (x, y) — ядро $\Rightarrow K_1 = \operatorname{ch}(x, y)$ и $K_2 = \operatorname{sh}(x, y)$ — ядра.
2. $K(x, y)$ — ядро как сумма K_1 и K_2 с положительными коэффициентами 1 и 3.

Задача 5

1. Нет. Склонность к переобучению уменьшается, т.к. увеличивается «усреднение» по объектам (меньше чувствительность к выбросам).

2. Нет. При увеличении количества элементов в листе наоборот получается «огрубление» модели.

3. Да. $C \to +\infty \Rightarrow$ вес регуляризатора $\to 0$. В предельном случае регуляризатор отсутствует, т.е. величина весов может быть сколь угодно большой, что как раз приводит к переобучению на мультиколлинеарной обучающей выборке.

Задача 6

Обозначим $n \stackrel{\text{def}}{=} |R_m|$. $p_k \stackrel{\text{def}}{=} \frac{n_k}{n}$, где $n_k \equiv \sum_{x_i \in R_m} [y_i = k]$ — количество объектов класса k в R_m. $k \in \overline{1,l}$ — всего l классов. По условию,

$$P\{a(x) = k\} = p_k$$

.

Частота ошибок — случайная величина $\xi = \frac{1}{n}\sum\limits_{i=1}^{n}[a(x_i) \neq y_i] = \frac{1}{n}\sum\limits_{i=1}^{n}\xi_i$, где $\xi_i = [a(x_i) \neq y_i]$ — также случайная величина. Эта величина принимает только значения 0 и 1, откуда находим $M\xi_i = 1 \cdot P\{\xi_i = 1\} + 0 \cdot P\{\xi_i = 0\} = P\{\xi_i = 1\} = P\{a(x_i) \neq y_i\} = 1 - P\{a(x_i) = y_i\} = 1 - p_{y_i}$

Найдем $M\xi = \frac{1}{n}\sum\limits_{i=1}^{n} M\xi_i = \frac{1}{n}\sum\limits_{i=1}^{n}(1 - p_{y_i}) = 1 - \sum\limits_{i=1}^{n}\frac{p_{y_i}}{n} = \boxed{=}$. Запишем $1 = \sum\limits_{k=1}^{l}[y_i = k]$, подставим это выражение в сумму: $\boxed{=} 1 - \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{l}[y_i = k]p_{y_i}$. Переставим суммы местами и воспользуемся тем, что при $y_i \neq k$ слагаемое равно 0: $\boxed{=} 1 - \frac{1}{n}\sum\limits_{k=1}^{l} p_k \sum\limits_{i=1}^{n}[y_i = k] = 1 - \frac{1}{n}\sum\limits_{k=1}^{l} p_k n_k = \boxed{1 - \sum\limits_{i=1}^{l} p_k^2}$.

Поскольку решающее правило как функция $a\colon \overline{1,n} \to \overline{1,l}$ определено неоднозначно, индекс Джини правила $a(x)$ — также случайная величина $\eta = \#\{(x_i, x_j) \in R_m^2 | y_i = y_j, a(x_i) = a(x_j)\} = \sum\limits_{i,j \in \overline{1,n}^2, y_i = y_j} \underbrace{[a(x_i) = a(x_j)]}_{\xi_{ij}}$. ξ_{ij} принимает значения только 0 и 1, откуда выразим $M\xi_{ij} = P\{\xi_{ij} = 1\} = P\{a(x_i) = a(x_j)\} = \sum\limits_{k=1}^{l} P\{a(x_i) = a(x_j) | a(x_i) = k\}P\{a(x_i) = k\}\boxed{=}$. Случайные величины $a(x_i)$ и $a(x_j)$ независимы при $i \neq j$, поэтому $\boxed{=}\sum\limits_{k=1}^{l} p_k^2$. При $i = j$ $M\xi_{ij} = 1$. Вернемся к $M\eta = \sum\limits_{i,j \in \overline{1,n}^2}[y_i = y_j] \cdot [i = j] \cdot 1 + \sum\limits_{i,j \in \overline{1,n}^2}\sum\limits_{k=1}^{l}[i \neq j][y_i = y_j]p_k^2 = n + \sum\limits_{k=1}^{l} p_k^2 \sum\limits_{i,j \in \overline{1,n}^2}[y_i = y_j][i \neq j]$. Рассмотрим последнюю сумму: $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}[i \neq j][y_i = y_j] = \sum\limits_{k=1}^{l} n_k(n_k - 1) = \sum\limits_{k=1}^{l} p_k^2 n^2 - n$ — количество пар различных объектов, принадлежащих одному классу. Получаем $M\eta = n + (\sum\limits_{k=1}^{l} p_k^2)(n^2 \sum\limits_{k=1}^{l} p_k^2 - n) \neq M\xi$.

??? проверил для $k = 2$, тоже не равно.