	Принят	Нет
Приехал на машине	20	88
Пешком	35	17

Вероятн., что кандид., пришедш. пешком будет принят на должн.

$\frac{35}{35+17} = \frac{35}{52} \approx 0{,}67$

Когда AUC = 0.5, тогда классифик. не может различ. положит. и отриц. баллы класса => либо предсказыв. случайн., либо существ. 1 постоянн. для всех точек

На плоск. даны точки в двумерн. простр. $X = [(-1,1), (1,-1), (1,1), (0,0)]$ с соотв. метками классов $y = [1, 1, 1, -1]$

С помощью leave-one-out кросс-валидации найдите оптим. число соседей $k \in [1, 3]$ в методе ближ. соседей

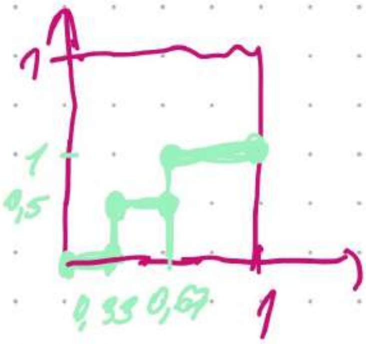
В кач-ве меры близости использ. евклидово расстоян., а метрика кач-ва accuracy

Первые 3 точки относ. к классу 1, остальн. -1

x_i	y_i
0.7	+1
0.6	−1
0.3	−1
0.45	+1
0.92	−1

Отсортируем по убыван.

0.92	−1
0.7	+1
0.6	−1
0.45	+1
0.3	−1

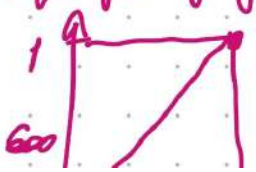


$0{,}33 \cdot 0{,}5 + 1 \cdot 0{,}33 = 0{,}165 + 0{,}33 = 0{,}495$
$= 0{,}5$

$$\frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot 1 = \frac{1}{6} + \frac{2}{3} = \frac{5}{6}$$

600+ и 400− , всего 1000

предсказыв. одну и ту же вероятн. попасть в + класс, то есть:



$$\frac{1 \cdot 1}{2} = 0{,}5$$

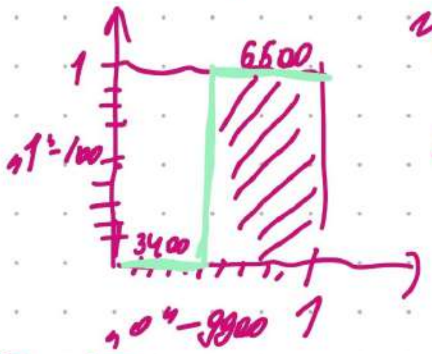
y

$\left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}3400$

$\left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}100$

$\left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}6500$

Строим квадрат 1 на 1



горизонталь делим на столько частей, сколько нулей, у нас их = 9900

вертикаль делим на кол-во единиц

идем сверху вниз по отсорт. табл

сначала идём на 3400 шагов вправо по 0, потом на все шаги вверх, потом снова вправо

$$1 \cdot \frac{6500}{9900} = \frac{65}{99} \approx 0,66$$

x_i	y_i
0.7	+1
0.6	−1
0.3	−1

Отсортируем по убыван.

0.92 −1

0.7 +1

Алгоритм бинарн. классифик. выдаёт значен. b_i - принадл. отр. $[0,1]$, всего 10 000 наблюд., если ранжир. по возраст., то $y_i = 1$ занимают 6501 по 6600. Площадь под ROC кривой - ?

$b(x)$ - прогнозы ; y - ответы

$b(x)$

y

0
⋮
0] 6500

1
⋮
1] 6501-6600, а в оставш. местах 0
100

нам неважно, что стоит в $b(x)$, важны метки

0
0
0] 3400

Когда строим ROC кривую, то сортир. по убыванию

y

0 1

Левая вершина $X \le 5$

x	y
3	10
2	5
5	15

$$\bar{y} = \frac{10+5+15}{3} = 10$$

$$H(Q_l) = \frac{1}{2} \cdot (y-\bar{y})^2 \sum_1^3$$

$$H(Q_l) = \frac{1}{2}\left[(10-10)^2 + (5-10)^2 + (15-10)^2\right] = \frac{1}{2}(25+25) = 25$$

Правая вершина $X > 5$

x	y
6	14

$$\bar{y} = 14$$

$$H(Q_r) = 0$$

$$IG = \frac{62}{3} - \frac{3}{4} \cdot 25 - 0 = \frac{62}{3} - \frac{75}{4} = \frac{23}{12} = 1{,}917 \approx 1{,}9$$

Выборка

X	y
3	10
2	5
5	15
6	14

Разбиения по усл. $X > \theta$, в данном случ. по $X > 5$

$|R| = 4$ - кол-во объекта

Формула несмещ. дисперсии $S_0^2 = \frac{\sum_{i=1}^{n}(X-\bar{X})^2}{n-1}$

$\bar{X}$ - в нашем случае это y, $\bar{y}$ - среднее

$$y_{средн} = \frac{\overset{15}{10+5}\overset{30}{+}15+14}{4} = \frac{44}{4} = 11$$

$$H(R) = \frac{1}{3} \cdot \sum_{1}^{4} (y-\bar{y})^2$$

$$H(R) = \frac{1}{3}\left[(10-11)^2+(5-11)^2+(15-11)^2+(14-11)^2\right] = \frac{1}{3}(\overset{52}{1+36+16}+9) = \frac{62}{3}$$

Левая вершина $X \le 5$

X	y
3	10

$\bar{y} = \frac{10+5+15}{...} = 10$

Критерий Джини $H(R_l) = 0{,}2857(1-0{,}2857) + 0{,}7143(1-0{,}7143) =$

$= 0{,}2857 \cdot 0{,}7143 + 0{,}7143 \cdot 0{,}2857 = 0{,}40815102$

В правой вершине R_r 20 1 и 10 0

Доля $p_{1,r} = \frac{20}{30} = \frac{2}{3}$

Доля $p_{0,r} = \frac{1}{3}$

Критерий Джини $H(R_r) = \frac{2}{3} \cdot \left(1 - \frac{2}{3}\right) + \frac{1}{3}\left(1 - \frac{1}{3}\right) = \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9} + \frac{2}{9} =$

$= \frac{4}{9} = 0{,}4444444$

Подставляем: $0{,}48 - \frac{70}{100} \cdot 0{,}40815102 - \frac{30}{100} \cdot 0{,}4444444 =$

$= 0{,}48 - 0{,}28570571 - 0{,}1333333 = 0{,}06$

Выборка

X	Y
3	10
2	5

Разбиения по усл. $X > \Theta$, в данном случ. по $X > 5$

$|R| = 4$ - всего объектов

• в правой остальные

$|A|$ - кол-во объектов в верш. A

$H(R) = \sum_{k=1}^{2} p_k(1-p_k)$ - значение крит. Джини

В исходн. верш. всего 100 объект.

Доля 1 $p_1 = \frac{40}{100} = 0{,}4$

Доля 0 $p_0 = \frac{60}{100} = 0{,}6$

Критерий Джини $H(R) = 0{,}4(1-0{,}4) + 0{,}6(1-0{,}6) = 0{,}48$

В левой вершине R_l 20 1 и 50 0

Доля 1 $p_{1,l} = \frac{20}{70} = \frac{2}{7} = 0{,}2857$

Доля 0 $p_{0,l} = \frac{50}{70} = 0{,}7143$

11:03 Пн 10 июня

101
110
111

$a = (1,1,1) \quad b = (1,2,0)$

$a - b = (0,-1,1)$

$|a-b| = \sqrt{1+1} = \sqrt{2}$

$K(a,b) = e^{(-2)}$

$\cos = \dfrac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|} = \dfrac{K(a,b)}{\sqrt{K(a,a)\cdot K(b,b)}} = \dfrac{e^{(-2)}}{\sqrt{1}} = 0{,}14$ - верно!

$k(a,a) = \exp(-0) = 1$

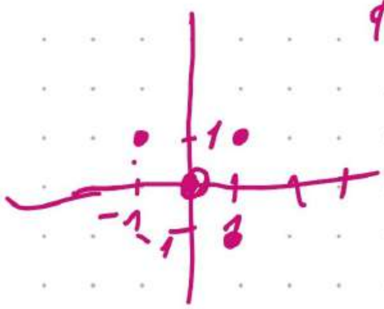
$k(b,b) = \exp(0) = 1$

повторить

$Q = H(R) - \dfrac{|R_l|}{|R|} H(R_l) - \dfrac{|R_r|}{|R|} H(R_r)$

В вершине 40 объект. 1 и 60 объект. 0 После разбиения:

- в левой верш. 20 1 и 50 0



Первые 3 точки относ. к классу 1, остальн. −1

Всего 3 варианта: $k=1$ or $k=2$ or $k=3$

accur. $=0$

потому что k ближ. сосед с меткой −1

Смотрим по тому, насколько много соседов нужн. класса, при $k=3$ лучше всех

Объект описыв. двумерн. вектором $x=(x_1, x_2)$

Дан вектор $w=(2,3)$ и число $w_0=7$

Найти ширину полосы между $\langle w, x\rangle = w_0+1$ и $\langle w, x\rangle = w_0-1$, где $\langle w, x\rangle$ — скалярн. произв. вектора w и x.

$2x_1+3x_2=7$ — центр. прямая

$L = \dfrac{2}{\|w\| \text{ (без } w_0)} = \dfrac{2}{\sqrt{13}}$ — метод опорных векторов