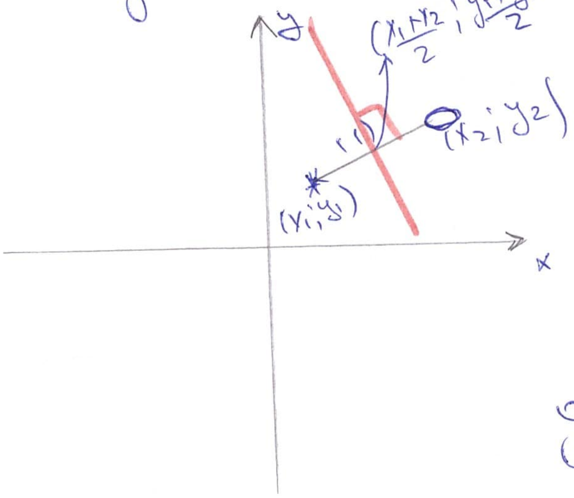
Задачи

Метрические методы, kNN

№1. 2 объекта в $\mathbb{R}^2$ => евклидово расстояние примет следующий вид. $\rho(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$

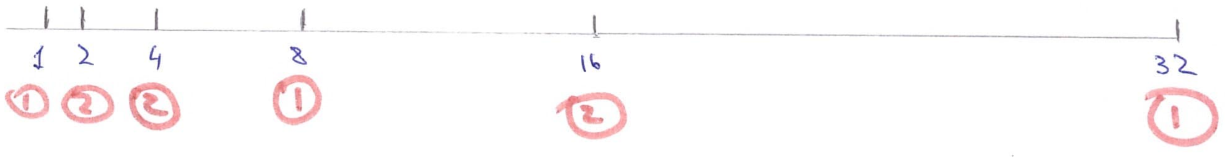
Тогда это $\mathbb{R}^2$ со стандартной метрикой.



Построим расстояние (1) между 2-умя объектами. Это прямая линия. Тогда, разделяющей поверхностью будет равноудалённый от двух ~~пр~~ объектов перпендикуляр, а это прямая линия. ч.т.д.

№2. Рассматриваем $\mathbb{R}$; $X=[1,2,4,8,16,32]$
$y=[1,2,2,1,2,1]$

Нарисуем картинку:



а) 1-NN. В случае 1-го ближайшего соседа без весов, согласно гипотезе компактности, x^* будет принадлежать тому же классу, что и его ближайший сосед, а в случае одинакового расстояния меньшему классу. Тогда, наш алгоритм примет вид: (рассматриваем граничные точки — точки между соседей)

$$a(x)=\begin{cases}1, & x\in(-\infty;1{,}5]\cup[6;12]\cup[24;+\infty)\\ 2, & x\in(1{,}5;6)\cup(12;24)\end{cases}$$

2-kNN

б) В отличии от пункта а), здесь нужно рассматривать точку смены дальнего соседа — эта точка равноудалённая от точек из X. Т.к., возьмём x^* левее 1, $x^* \in 1$, далее рассмотрим точку смены соседа 1 на соседа 4, эта точка $\frac{4+1}{2} = 2,5 \Rightarrow$ правее 2,5 новый ближ. соседа $\Rightarrow$ 2 и 4 ближ. соседи. Тогда, алгоритм примет вид:

$$a(x) = \begin{cases} 1, & x \in (-\infty; 2,5] \cup [5; +\infty) \\ 2, & x \in (2,5; 5). \end{cases}$$

2-NN с весами.

в) Т.к. в данном kNN мы учитываем веса $w = \frac{1}{\rho(x_1, x_2)}$, то каждый раз, когда мы определяем класс x^*, то нам нужно считать ρ до каждого из 2ух соседей, что знать вес какого из них нужнее. Для этого нужно отследить точки равноудалённые от двух соседей, а также нужно отслеживать точки смены соседей. Тогда $a(x)$ примет вид:

$$a(x) = \begin{cases} 1, & x \in (-\infty; 1,5] \cup [6; 12] \cup [24; +\infty) \\ 2, & x \in (1,5; 6) \cup (12; 24) \end{cases}$$

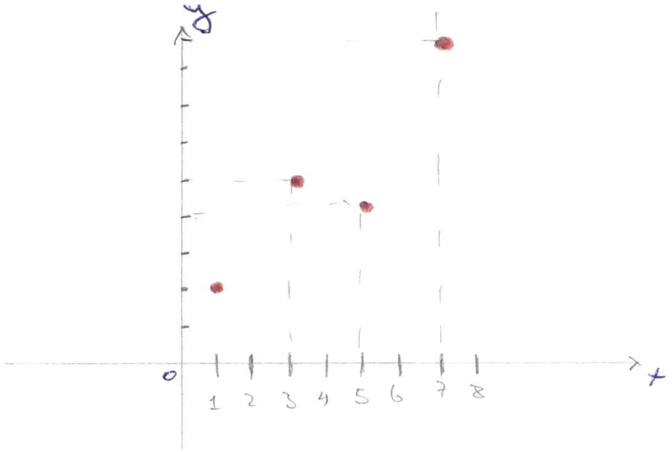
г) Данный пункт делается аналогично пункту б), только у нас 3 соседа и мы смотрим каких соседей больше. С учётом этого а алгоритм примет вид:

$$a(x) = \begin{cases} 1, & x \in [18; +\infty) \\ 2, & x \in (-\infty; 18) \end{cases}$$

N3. В алгоритме kNN присутствует всего один гиперпараметр — k. Если $k = 1$ — переобучение, а если $k = \ell$ — недообучение.

№4. $a(x) = \dfrac{\sum_{i=1}^{3} w_i y_i}{\sum_{i=1}^{3} w_i}$ → алгоритм 3-NN (взвешенный)

$w_i = \dfrac{k-i+1}{k}$



Нужно посчитать координату y для точки $x = 3.5$. Для начала найдём 3х ближайших соседей

1-й – 3
2-ой – 5
3-ий – 1

$\Rightarrow$ $w_1 = \dfrac{3-1+1}{3} = 1$

$w_2 = \dfrac{5-1+1}{3} = \dfrac{2}{3}$

$w_3 = \dfrac{1-1+1}{3} = \dfrac{1}{3}$

Тогда $a(3.5) = \dfrac{5 \cdot 1 + 4 \cdot \frac{2}{3} + 2 \cdot \frac{1}{3}}{2} = \dfrac{25}{6}$.

№5. $O(NB)$

Линейные методы

№1. 10 признаков ⇒ 10 весов (чисел), которые надо настроить. Нахождение весов в линейной регрессии ($a(x) = w_0 + \sum_{i=1}^{\ell} w_i x_i$, ℓ – кол-во признаков) сводится к следующей задаче:

1. Без регуляризации: $Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w_i, x_i \rangle - y_i)^2 \to \min_w$

2. С регуляризацией: $Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i) + C\|w\|^2 \to \min_w$,

где C – гиперпараметр.

Тогда, наша таблица примет вид:

Параметры:

	Без регуляр.	С регуляр.
Исходные признаки	10	10
Исходные + квадр.	20	20

Гиперпараметры

	Без регуляр.	С регул.
Исход. признаки	0	1
Исх. + квадр.	0	1

№2. Чтобы остаток было равно 1, $a(x)$ должен быть равен y, тогда на основании этого составим систему неравенств и решим её:

$$\begin{cases} w_0 + 0{,}2w_1 + 0{,}4w_2 + 0 \cdot w_3 > 0 \\ w_0 + 0{,}5w_1 + 0{,}9w_2 + 0 \cdot w_3 > 0 \\ w_0 + 0{,}3w_1 + 0{,}3w_2 + 0 \cdot w_3 > 0 \\ w_0 + 0{,}1w_1 + 0{,}8w_2 + w_3 < 0 \\ w_0 + 0{,}5w_1 + 0{,}7w_2 + w_3 < 0 \\ w_0 + 0{,}9w_1 + 0{,}9w_2 + w_3 < 0 \\ w_0 + 0{,}1w_1 + 0{,}3w_2 + w_3 < 0 \end{cases}$$

Данная система имеет бесконечное множество решений. => бесконечное число коэффициентов.

б) Возьмём след. коэфф.: $w_0 = 0$, $w_1 = 1$, $w_2 = 1$, $w_3 = -3$; $w_0 = 0$, $w_1 = 1$, $w_2 = 1$, $w_3 = -2$; $w_0 = 0$, $w_1 = 1$, $w_2 = 1$, $w_3 = -4$

№3. Запишем $MSE(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$

Для того, чтобы ~~на~~ $MSE = 0$ нужно, чтобы $a(x_i) = y_i$. С учетом этого составим СЛУ:

$$\begin{cases} w_0 + w_1 + w_2 = 0 \\ w_0 = 1 \\ \cancel{0{,}1} \; w_0 + 0{,}5 w_1 + 0{,}5 w_2 = 0{,}5 \end{cases}$$

Запишем матрицу из коэф-ов:

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0{,}5 & 0{,}5 & 0{,}5 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Т.к. в главном ступ. виде матрице 2 лин. независимых строки, а неизвестных 3 => беск. кол-во решений.

б) $\begin{cases} w_0 = 1 \\ w_1 = 0 \\ w_2 = -1 \end{cases}$ $\begin{cases} w_0 = 1 \\ w_1 = -1 \\ w_2 = 0 \end{cases}$ $\begin{cases} w_0 = 1 \\ w_1 = 9 \\ w_2 = -10. \end{cases}$

№4. Чтобы $a(x)$ предсказывал не ~~вер-ти~~ принадлежность к классу, а вер-ти, то прогоним её через сигмоиду $\sigma(z) = \dfrac{1}{1+\exp(-z)}$

Подставим $\langle w, x\rangle$ в сигмоиду и получим:

$$\frac{1}{1+\exp(-\langle w; x\rangle)} = p(y=1|x)$$

Выразим скалярное произведение:

$$\langle w, x\rangle = \log \frac{p(y=+1|x)}{p(y=-1|x)}$$

Подставим это в нашу ф-ию потерь:

$$L(a(x), y) = [y=+1] \log \frac{1}{1+\exp(-\langle w, x_i\rangle)} + [y=-1] \log \frac{\exp(-\langle w, x\rangle)}{1+\exp(-\langle w, x\rangle)} =$$

$$= [y=+1] \log \frac{1}{1+\exp(-\langle w, x_i\rangle)} + [y=-1] \log \frac{1}{1+\exp(\langle w, x\rangle)} =$$

$= \log(1 + \exp(-y_i \langle w, x_i \rangle))$.

№5. Запишем в общем виде функционал, который будем минимизировать:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + c \sum_{i=1}^{d} w_i^2$$

В нашем случае $d = 1$:

$$Q(a, w) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i - y_i)^2 + c w_1^2$$

После диф-ания по весам и нахождения min по w получим:

$$w = (X^T X + cE)^{-1} X^T y.$$

Деревья

№1. Чтобы выбрать правильный первый узел нужно выбрать предикат, чтобы энтропия была минимальна

$X_1 = A1$ $\quad H = -(0{,}75 \ln 0{,}75 + 0{,}25 \ln 0{,}25 + \ln 0{,}5) \approx 1{,}3$

AAAB (0,75; 0,25) $\quad$ AAABBB (0,5; 0,5)

$X_1 = B1$ $\quad H = -(0{,}66 \ln 0{,}66 + 0{,}33 \ln 0{,}33 + 0{,}57 \ln 0{,}57 + 0{,}43 \ln 0{,}43) = 1{,}32$

AAB (0,66; 0,33) $\quad$ AAAABBB (0,57; 0,43)

$X_1 = C1$ $\quad H = -(0{,}33 \ln 0{,}33 + 0{,}66 \ln 0{,}66 + 0{,}71 \ln 0{,}71 + 0{,}29 \ln 0{,}29) = 1{,}24$

ABB (0,33; 0,66) $\quad$ AAAAABB (0,71; 0,29)

$X_2 = A2$ $\quad H = -(0 + 0{,}42 \ln 0{,}42 + 0{,}58 \ln 0{,}58) = 0{,}68$

AAA (1; 0) $\quad$ AAABBBB (0,42; 0,58)

$X_2 = B2$ $\quad H = -(0{,}33 \ln 0{,}33 + 0{,}66 \ln 0{,}66 + 0{,}71 \ln 0{,}71 + 0{,}29 \ln 0{,}29) = 1{,}33$

ABB (0,33; 0,66) $\quad$ AAAAABB (0,71; 0,29)

$X_2 = C_2$ $\quad H = -(\ln 0{,}5 + 0{,}66 \ln 0{,}66 + 0{,}33 \ln 0{,}33) = 1{,}33$

AABB $\quad$ BBAAAA

(0,5; 0,5) $\quad$ (0,66; 0,33)

$X_3 = A_3$ $\quad H = -(0{,}66 \ln 0{,}66 + 0{,}33 \ln 0{,}33 + \ln 0{,}5) = 1{,}33$

AAAABB $\quad$ AABB

(0,66; 0,33) $\quad$ (0,5; 0,5)

$X_3 = B_3$ $\quad H = -(\ln 0{,}5 + 0{,}66 \ln 0{,}66 + 0{,}33 \ln 0{,}33) = 1{,}33.$

AABB $\quad$ AAAABB

(0,5; 0,5) $\quad$ (0,66; 0,33)

Лучшее разбиение $[X_2 = A_2]$, т.к. $H = 0{,}68$.

N2. Критерий информативности: $H(X) = \frac{1}{|X|} \sum_{i \in X} (y_i - \bar{y}(X))^2$

Делаем аналогично. Выбираем признак с минимальной средн. дисперсией $\bar{y}(X) = \frac{1}{|X|} \sum_{i \in X} y_i$

(т.е. с дисперсией в левой части)

N3. Decision Tree 2

N4. Tree 4.

Метрики качества:

N1.

1) a) Т.к. AUC-ROC $= 0{,}1 \Rightarrow b(x)$ — это очень плохой алгоритм, тогда, чтобы его улучшить возьмём обратный $(1 - b(x))$.

2) ? (Не улучшится)?

N2.

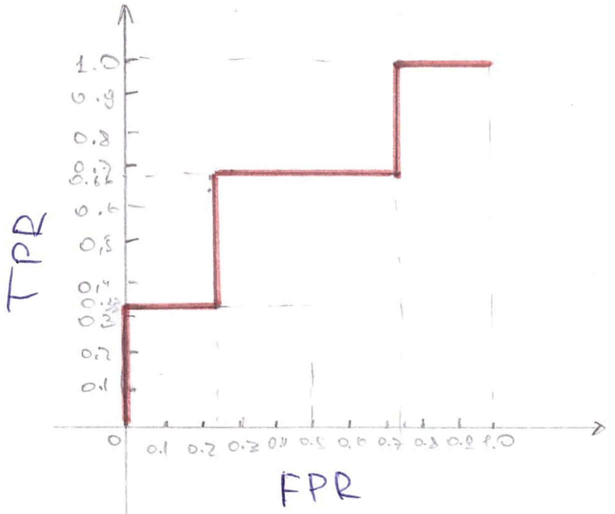
1. $p = [0{,}9,\ 0.1,\ 0{,}75,\ 0.56,\ 0.2,\ 0.37,\ 0.25]$

$y = [+1,\ -1,\ -1,\ +1,\ +1,\ -1,\ -1]$

Отсортируем массив вер-тей

$p = [0{,}9,\ 0{,}75,\ 0.56,\ 0.37,\ 0.25,\ 0.2,\ 0.1]$

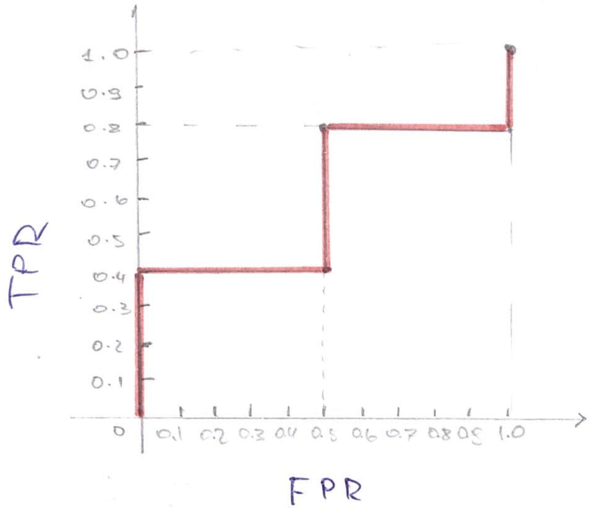
$y = [+1 \quad -1 \quad +1 \quad -1 \quad -1 \quad +1 \quad -1]$



$$AUC-ROC = 0{,}33 \cdot 0{,}25 + 0{,}66 \cdot 0{,}5 + 0{,}25 \cdot 1 = 0{,}0825 + 0{,}33 + 0{,}25 =$$
$$= 0{,}4975 + 0{,}1625 = 0{,}66$$

2. $p = [0{,}9 \quad 0{,}75 \quad 0{,}56 \quad 0{,}37 \quad 0{,}3 \quad 0{,}25 \quad 0{,}2]$

$y = [+1 \quad +1 \quad -1 \quad +1 \quad +1 \quad -1 \quad +1]$



$$AUC-ROC = 0{,}5 \cdot 0{,}4 + 0{,}5 \cdot 0{,}8 =$$
$$= 0{,}2 + 0{,}4 = 0{,}6.$$

№3

$$F_\beta = (1+\beta^2)\frac{precision * recall}{\beta^2 \cdot precision + recall}$$, важнее полнота $\Rightarrow \beta = \frac{1}{3}$

$$F_\beta = (1+\frac{1}{9})\frac{precision\ recall}{\frac{1}{9}precision + recall} = 10\frac{precision\ recall}{precision + 9\ recall}$$

Ответ: 1.