IAD

Банда цумовского катка

Осень 2019 – Весна 2020

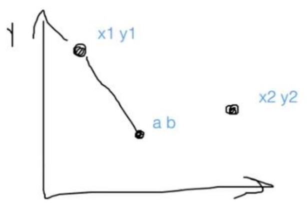
Оглавление

Глава 1

Метрические методы, knn

1.1 Задача 1

Пусть на плоскости дано два объекта двух разных классов. Покажите, что если выполнять классификацию методом 1-NN с евклидовым расстоянием, то разделяющей границей между классами будет прямая линия.



Пусть есть две точки X_1 и X_2 с координатами (x_1, y_1) и (x_2, y_2). В каком случае мы причислим новую точку (a, b) к классу 1, а в каком - к классу 2? Если расстояние от X_1 до нее меньше, чем расстояние от X_2.

$$p(x, X_1) < p(x, X_2)$$

Тогда условие принадлежности новой точки к классу 1:

$$\sqrt{(a - x_1)^2 - (b - y_1)^2} < \sqrt{(a - x_2)^2 - (b - y_2)^2}$$

Раскроем скобки.

$$2b(y_2 - y_1) < 2a(x_2 - x_1) - (x_1^2 - x_2^2) - (y_2^2 - y_1^2)$$

Получаем неравенство:

$$b > \frac{2a(x_2 - x_1) - (x_1^2 - x_2^2) - (y_2^2 - y_1^2)}{2(y_2 - y_1)}$$

$$b > a \times \underbrace{\frac{2(x_2 - x_1)}{2(y_2 - y_1)}}_{\text{coef}} - \underbrace{\frac{(x_1^2 - x_2^2)}{2(y_2 - y_1)}}_{\text{const}} - \underbrace{\frac{(y_2^2 - y_1^2)}{2(y_2 - y_1)}}_{\text{const}}$$

Мы видим, что условие принадлежности точки к классу 1 определяется **линейной функцией** $f(a)$ (при этом, знак неравенства нам неважен, поскольку он зависит от того, как расположены точки в пространстве согласно своим координатам, и влияет только на знак линейной функции, но не на саму линейность).

Либо проводим серединный перпендикуляр между точками, и он является ГМТ, равноудаленных от X_1 и X_2, соответственно условие принадлежности новой точки к одному из классов - ее место на плоскости по отношению к перпендикуляру, который является прямой.

1.2 Задача 2

Пусть даны следующие точки в одномерном пространстве:

$$X = [1, 2, 4, 8, 16, 32]$$

с соответствующими метками классов:

$$y = [1, 2, 2, 1, 2, 1]$$

Обозначим через x^* объект, для которого необходимо выполнить классификацию, а через $a(x)$ - алгоритм, в соответствии с которым выполняется классификация.

Найдите и выпишите границы классов, если $a(x)$ это

1 1-NN,

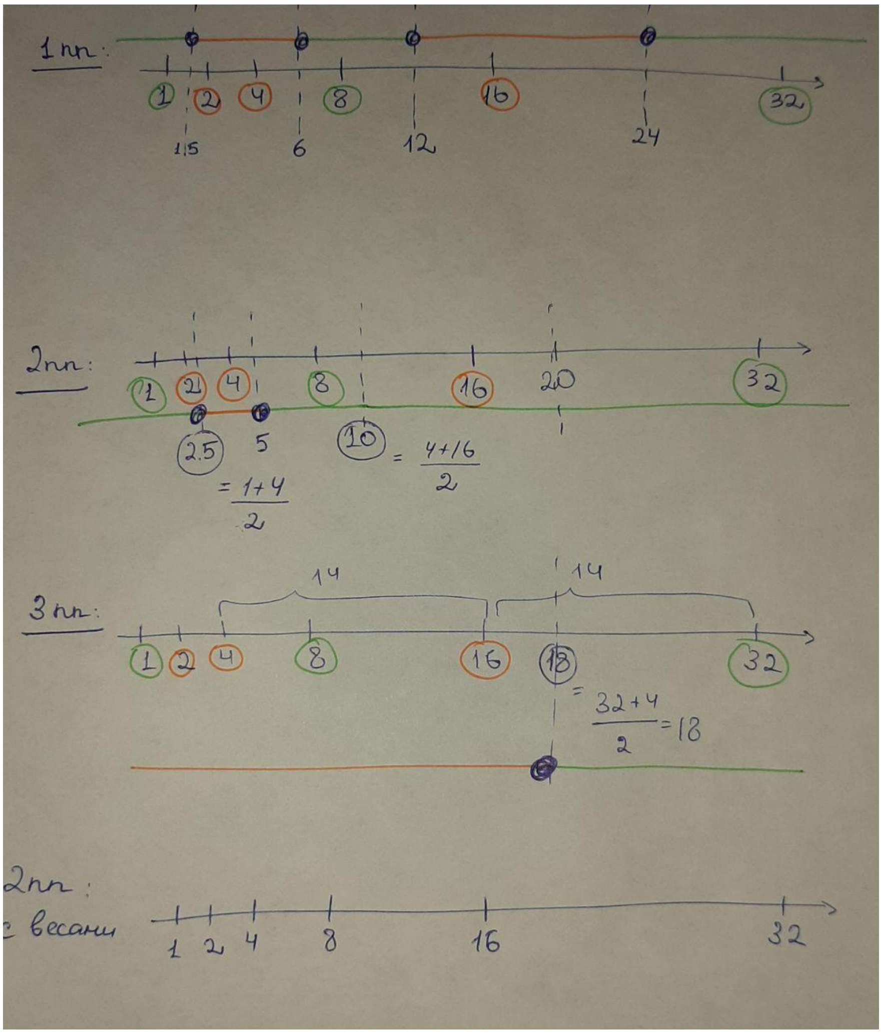
2 2-NN,

3 2-NN, с весами обратно пропорциональными расстоянию до ближайших соседей,

4 3-NN

Для всех $a(x)$ мера близости - евклидова. В случае равнозначности, выставляется класс с наименьшей меткой.

Объект X относится к классу y_k, если большинство ближайших к нему объектов по метрике принадлежат классу k.



1.3 Задача 3

Перечислите все числовые гиперпараметры метода k-NN и определите как они влияют на переобучение/недообучение

В алгоритме KNN присутствует всего один гиперпараметр -> k. Если $k = 1$ - переобучение, а если $k = l$ -недообучение (l - размер выборки)

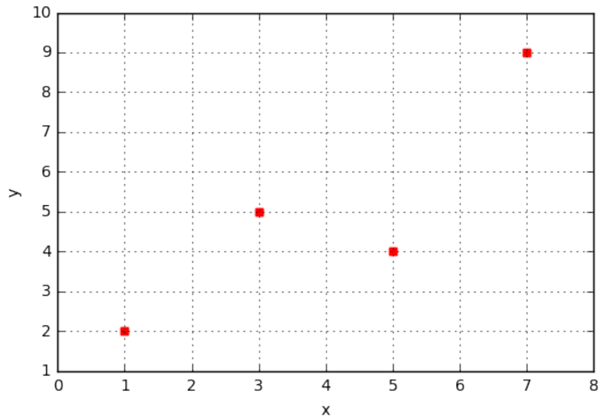
1.4 Задача 4

По графику ниже выполните регрессию точки с координатой в $x = 3.5$ с помощью взвешенного метода k-NN, где $k = 3$, расстояние - евклидово, а вес i-го ближайшего соседа определяется

как, "

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

.



$$a(x) = \frac{\sum_{i=1}^{3} w_i y_i}{\sum_{i=1}^{3} w_i} - -\text{алгоритм 3-NN (взвешенный)}$$

Считаем координаты y для точки $x = 3.5$. Берем 3 ближаших соседа (из условия).

x	w_i	result
3	$\frac{3-1+1}{3}$	$= 1$
5	$\frac{3-2+1}{3}$	$= \frac{2}{3}$
1	$\frac{3-3+1}{3}$	$= \frac{1}{3}$

Тогда:

$$a(3{,}5) = \frac{5 * 1 + 4\frac{2}{3} + 2\frac{1}{3}}{2} = \frac{25}{6}, \text{ где}-(5,4,2)-\text{значения по } y \text{ у } x \text{ координат.}$$

Глава 2

Линейные методы

2.1 Задача 1

У Вас есть набор данных из 1000 объектов, описанных 10 признаками. Вы обучаете модель линейной регрессии с константным признаком. Вы рассматриваете 2 модификации модели: (а) включение Lasso-регуляризации и (б) добавление к исходным признакам квадратов каждого признака. Таким образом, у Вас может получиться 4 модели: () без регуляризации с обычными признаками, (а), (б) и (аб).

В модели линейной регрессии с константным признаком мы имеем $a(x) = w_0 + \sum_{i=1}^{l}(w_i x_i)$. Соответственно, в регрессии без регулярицзации с обычными признаками мы настраиваем то, что записано в изначальной формулировке модели: $w_0 w_1$,...,$w_1 0 = 11$. Добавляя квадраты признаков, мы увеличиваем количество в два раза $= 22$ признака.

В формуле для линейной регрессии без регуляризации отсутствует опция настройки гиперпараметра, поэтому для () и (б) мы его настроить не можем $= 0$.

Если рассмотреть модель с регуляризацией, количество параметров от этого не изменится, так как количество признаков не зависит от гиперпараметра. Соответственно, настраиваем 11 чисел для модели с исходными признаками, 22 - для исходных + квадратов признаков.

В модель с регуляризацией добавляется одна переменная - гиперпараметр, который мы можем предварительно настроить. Соответственно, вне зависимости от конфигурации признаков, настраиваем один гиперпараметр.

Сколько параметров нужно настроить (сколько чисел)?*

() 11
(а) 11
(б) 22
(аб) 22

Сколько гиперпараметров нужно настроить (сколько чисел)?

() 0
(а) 1
(б) 0
(аб) 1

2.2 Задача 2

Вы решаете задачу бинарной линейной классификации в трехмерном признаковом пространстве. Соответственно, решающее правило имеет вид:

$$a(x) = sign(w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3)$$

У Вас есть следующие объекты:

x_1	x_2	x_3	y
0.2	0.4	0	+1
0.5	0.9	0	+1
0.3	0.3	0	+1
0.1	0.8	1	-1
0.5	0.7	1	-1
0.9	0.9	1	-1
0.1	0.3	1	-1

(а) Сколько существует наборов коэффициентов (w0,w1,w2,w3), при которых задача будет решена идеально (с accuracy=1)?

Значение x_3 перед определяет наличие влияния коэффициента w_3 на модель. Мы видим, что, когда влияние есть ($x_3 = 1$), y классифицируется как -1, в противном случае — +1. Соответственно, нам нужен такой w_3, чтобы по модулю он превосходил решающее правило в такой его вариации, в которой $w_0 + w_1 x_1 + w_2 x_2$ принимает наибольшее значение. Среди имеющихся объектов этому условию удовлетворяет следующий набор x: (0.9;0.9;1).

Соответственно, по модулю w_3 должен быть больше суммы взвешенных признаков, однако принимать отрицательное значение, чтобы все выражение принимало отрицательное значение и классифицировало y как -1.

Таким образом, единственное необходимое и достаточное условие — любой $-|w_3|$, где $|w_3| > w_0 + 0.9w_1 + 0.9w_2$. Под такое условие подходит бесконечное количество наборов коэффициентов (w0,w1,w2,w3).

(б) Если ответ в п. (а) 0, пропустите это задание. Если ответ в п. (а) - 1 или 2, запишите этот набор коэффициентов (или 2 набора). Иначе приведите хотя бы три таких набора коэффициентов.

1) (0.1;0;0;-1);
2) (1;1;1;-3);
3) (-1;2;0;-1).

2.3 Задача 3

В пункте а достаточно составить систему уравнений следующего вида:

$$\begin{cases} 0 = & w_0 + w_1 + w_1 \\ 1 = & w_0 \\ 0.5 = & w_0 + 0.5w_1 + 0.5w_2 \end{cases}$$

Как видно,у данной системы бесконечное число решений, которому достаточно удовлетворять условию: $W_1 + w_2 = -1$

б) Такие решения -
$x_1 = -0.5; x_2 = -0.5$
$x_1 = -0.7; x_2 = -0.3$
$x_1 = -0.2; x_2 = -0.8$

2.4 Задача 4

Для того, чтобы наш классификатор (линейный) предсказывал вероятность принадлежности, нам нужно прогнать $a(x) =< w,x >$ через сигмоиду. Тем самым, мы переведем получаемые нами значения на отрезок [0,1]

$\sigma(w,x) = \dfrac{1}{1+exp(-<w;x>)} = P(1|x)$

Скалярное произведение:

$<w,x>= log(\dfrac{P(y=+1|x)}{P(y=-1|x)})$

Подставим в нашу функцию потерь:

$L(a(x),y) = [y=+1]log(\dfrac{1}{1+exp(-<w;x>)})+[y=-1]log(\dfrac{exp(-<w;x>)}{1+exp(-<w;x>)}) =$

y=+1

$log(\dfrac{1}{1+exp(-<w;x>)})+[y=-1]log(\dfrac{1}{1+exp(<w;x>)}) = log(1+(exp(-y\times <w;x>)$

2.5 Задача 5

Краткое решение (оно правильное, мне добавит нечего) смотрите в картине, приложенной в папке images Чуть более подробное с дифференциацией (картинка - подробное решение)

Глава 3

Решающие деревья

3.1 деревушки

3.1.1 Задача 1. Классификация по категориальным признакам

Используя таблицу ниже, по какому признаку следует формировать первый узел решающего дерева, если мы хотим предсказать Y? В качестве критерия информативности использовать энтропию, в качестве критериев разделения - индикаторы $[x_j = a]$

x1	x2	x3	y
A1	A2	A3	A
B1	A2	A3	A
C1	C2	A3	A
A1	A2	B3	A
C1	B2	A3	B
B1	C2	B3	A
A1	B2	A3	A
C1	C2	B3	B
B1	B2	B3	B
A1	C2	A3	B

Вспомним формулу энтропии:

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k.$$

А теперь просто все считаем, а именно нужно понять, где энтрапия минимальна, **однако в данном варианте рассчитана только одна энтрапия по критерию. Нужно не забыть рассчитать и вторую и затем с весами посчитать по формуле:**

$$Q(R_m, j, s) = \frac{|R_\ell|}{|R_m|}H(R_\ell) + \frac{|R_r|}{|R_m|}H(R_r). -> min$$

тоже простите за фотки...

3.1.2 Задача 2. Регрессия по числовым признакам

Определите признак, по которому следует произвести первое ответвление в решающем дереве для данных, приведенных ниже? В качестве критерия информативности использовать дисперсию, в качестве критериев разделения - индикаторы $[x_j < a]$

x1	x2	x3	y
2	1	2	1
2	3	3	2
5	3	1	2
5	6	4	4
6	5	3	4
7	5	5	3
8	7	2	6

Здесь $H(R)$ — это *критерий информативности* (impurity criterion), который оценивает качество распределения целевой переменной среди объектов множества R. Чем меньше разнообразие целевой переменной, тем меньше должно быть значение критерия информативности — и, соответственно, мы будем пытаться минимизировать его значение. **Для регресси это следующий вид:**

$$H(R) = \frac{1}{|R|} \sum_{(x_i,y_i)\in R} \left(y_i - \boxed{\frac{1}{|R|} \sum_{(x_i,y_i)\in R}} y_i \right)^2 ., \text{где } \boxed{\frac{1}{|R|} \sum_{(x_i,y_i)\in R} y_i} = \bar{y}(R)$$

,

Мы получили, что информативность вершины измеряется её дисперсией — чем ниже разброс целевой переменной, тем лучше вершина.

Посчитаем теперь для одного случая (остальные аналогично делатся и потом сраниваются):

x1 $<$ 4:

$$\bar{y}(R_r) = \frac{1}{|R_r|} \sum_{(x_i,y_i)\in R} y_i = \frac{1+2}{2} = 1.5$$

$$\bar{y}(R_\ell) = \frac{1}{|R_\ell|} \sum_{(x_i,y_i)\in R} y_i = \frac{2+4+4+3+6}{5} = 3.8$$

$$H(R) = \frac{1}{|R|} \sum_{(x_i,y_i)\in R} (y_i - \bar{y}(R))^2 ., \textbf{Подставляем сюда } \bar{y}(R_\ell) \textbf{ и } \bar{y}(R_r) \textbf{ и } y_i \textbf{ по ним}$$

В итоге получаем $H(R_\ell)$ **и** $H(R_r)$. Затем по формуле с весами смотрим на значение. Там, где это будет минимально - значит, самое лучшее для нас

$$Q(R_m, j, s) = \frac{|R_\ell|}{|R_m|} H(R_\ell) + \frac{|R_r|}{|R_m|} H(R_r). -> min$$

Глава 4

Метрики качества

4.1 Вопросы по ROC кривую.

4.1.1

У алгоритма $b(x)$ AUC-ROC равен 0.1. Предложите способ построить алгоритм, имеющий лучшее качество.

Можно просто предложить алгоритм c(x) = 1 - b(x). AUC-ROC < 0.5 говорит о том, то наш алгоритм путает класс 1 и класс 0 (вместо вероятности принадлежности к классу 1, выдает вероятность принадлежности к классу 0). Если вычитать все значения b(x) из единицы, все значения TPR и FPR при переборе порогов будут также вычитаться из единицы. Так, например, если для первого порога TPR была равна 1/5 (внизу оси y), она станет 4/5 (вверху оси y).

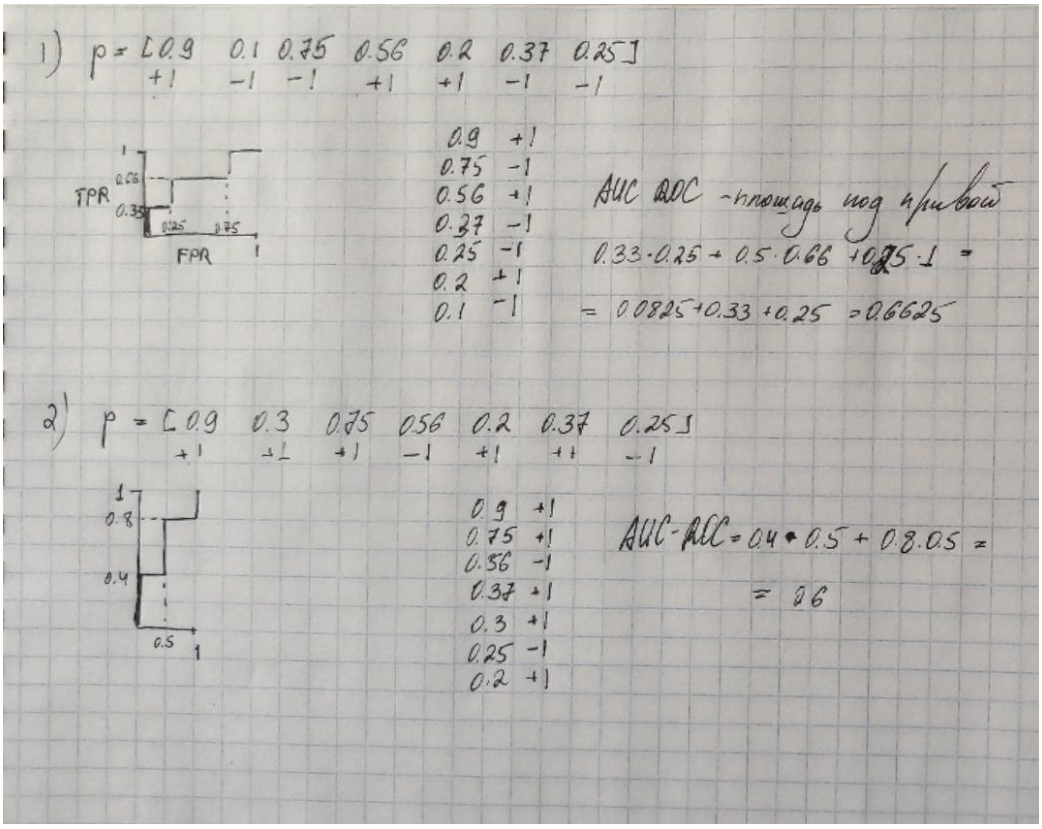
4.1.2

Вася построил алгоритм b(x), AUC-ROC которого 0.63. Петя построил алгоритм, $c(x) = b(x)/3$. Чему равен AUC-ROC для него?

Так же 0.63. Для решения таких задач нужно запомнить интуицию, что если мы как-то преобразуем оагоритм, но он все еще выдает вероятность попадания в класс 1 выше, чем вероятность попадания в класс 0, и ранжирование объектов по вероятнсоти попадания в класс 1 никак не меняется, то ROC никак не изменится вообще.

4.2 Задачи на рисование ROC-кривых и подсчёт AUC-ROC.

извините за фотки друзья.....



4.3 Задачи про F_β меру

$\beta < 1$ – важнее точность
$\beta > 1$ – важнее полнота

Если точность в три раза больше, чем полнота, то $\beta = \frac{1}{3}$

$$F_\beta = (1+\beta^2) * \frac{precision * recall}{\beta^2 * precision + recall} = (1+\frac{1}{9}) * \frac{precision * recall}{\frac{1}{9} * precision + recall} =$$
$$\frac{10}{9} * \frac{precision * recall}{\frac{1}{9} * (precision + 9 * recall)} = 10 * \frac{precision * recall}{precision + 9 * recall}$$

Ответ: 1

4.4 Задачи про точность, полноту, F1-меру

Посчитайте точность, полноту и F1-меру для алгоритма, если

1. TP = 8 , FP = 2 , FN = 16 , TN = 4
2. TP = 5 , FP = 4 , FN = 6 , TN = 7

$Precision = \dfrac{TP}{TP + FP}$

$Recall = \dfrac{TP}{TP + FN}$

$F_1 = \dfrac{2 * precision * recall}{precision + recall}$ это просто F-мера с $\beta = 1$

Соответственно:

1. P = $\frac{8}{8+2}$ = 0.8, R = $\frac{8}{8+16}$ = 0.33, F = $\frac{2*0.8*0.33}{0.8+0.33}$ = 0.46

2. P = $\frac{5}{5+4}$ = 0.55, R = $\frac{5}{5+6}$ = 0.45, F = $\frac{2*0.55*0.45}{0.55+0.45}$ = 0.495

4.5 Ещё одна задача про точность, полноту и F1-меру

Логистическая регрессия, вектор предсказанных вероятностей принадлежности к классу +1:
p = [0.9 0.1 0.75 0.56 0.2 0.37 0.25]

Вектор правильных ответов:

y = [+1 -1 -1 +1 +1 -1 -1]

Бинаризуйте ответ по порогу t и посчитайте точность, полноту и F1-меру.

1. t = 0.3
2. t = 0.8

1. TP = 2, FP = 2, FN = 1, TN = 2
P = $\frac{2}{4}$ = 0.5, R = $\frac{2}{3}$ = 0.67, F = $\dfrac{2 * 0.5 * 0.67}{0.5 + 0.67}$ = 0.57

2. TP = 1, FP = 0, FN = 2, TN = 4
$\mathrm{P} = \frac{1}{1} = 1$, $\mathrm{R} = \frac{1}{3} = 0.33$, $\mathrm{F} = \dfrac{2 * 1 * 0.33}{1 + 0.33} = 0.496$