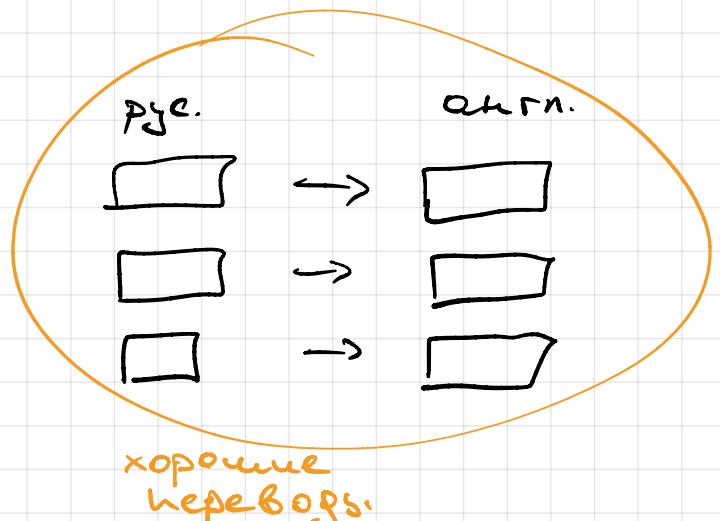


Введение

$f(x, w) \rightarrow y$

↑
текст
на русском

↑
текст
на английском



подберём w , чтобы
ко всем было понятно

Пример: студент \rightarrow оценка за МО-1

↑
Объект (sample)
 x

↑
Целевая переменная (target)
 y

Пр-во объектов

X

Пр-во ответов

$Y : [0; 10], \mathbb{R}$

Опр.: Обучающая выборка

$$X = \{(x_i, y_i)\}_{i=1}^c$$

c - размер обучающей выборки

Опр.: Признаки (факторы, features) - характеристики объектов

$x_i = (x_{i1}, \dots, x_{id})$
↑
объект
число признаков

Какие бывают признаки?

1) числовые ($x_j \in \mathbb{R}$)

2) категориальные ($x_j \in \{c_1, \dots, c_m\}$)

3) порядковые ($x_j \in \{c_1, \dots, c_m\}$)

если он сравнивается

! Тип задачи определяется целевой переменной

I Обучение с учителем (Supervised learning)

- есть целевая переменная

① $Y = \mathbb{R}$ - регрессия

② $|Y| < \infty$ - классификация

$Y = \{0, 1\}$ - бинарная классификация

$Y = \{1, \dots, k\}$ - многоклассовая классификация
multiclass

$Y = \{0, 1\}^k$ - классификация с пересекающимися
классами *multilabel*

II Обучение без учителя (unsupervised learning)

- нет целевой переменной

① кластеризация

\times - надо разбить на группы, чтобы в каждой группе
объекты были не очень

② Оценивание плотности



По объекту можно , мог ли он прийти отсюда

Есть еще: - частичное обучение
- обучение с подкреплением RL

Оп.: Модель (алгоритм) - $a: \mathbb{X} \rightarrow \mathbb{Y}$

Примеры : ① $a(x) = w_0 + w_1 x_1 + \dots + w_d x_d$

признаки

параметры

② нейронные сети

Оп. 1 Семейство моделей - $A = \{a(x, w) | w\}$

Опн.: Функция потерь (Loss Function) - $L: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$

$$L(y, z)$$

правильный
ответ

отвёрт
модели

Примеры: ① $(y-z)^2$

② $|y-z|$

③ $(\log y - \log z)^2$

Опн.: Функционал ошибки - $Q(a, X)$

В большинстве случаев: $Q(a, X) = \frac{1}{C} \sum_{i=1}^C L(y_i, a(x_i))$

Бывает функционал качества

$$\frac{1}{C} \sum_{i=1}^C \underbrace{\delta_i}_{\text{вес}} L(y_i, a(x_i))$$

*: Обучение модели: $Q(a, X) \rightarrow \min_{a \in \mathcal{A}}$

$X \rightarrow$
 $L \rightarrow a_s(x)$ - наименее, это на
 $\mathcal{A} \rightarrow$ выборах данных это
тогда будет работать
хорошо

Ответка: Бывают и иные задачи, но это самая частная
модель

- Этапы решения задачи:
- 1) постановка задачи
 - 2) выбор выборки
 - 3) разработка признаков
 - 4) выбор функции потерь
 - 5) выбор семейства моделей
 - 6) обучение модели
 - 7) валидация модели
 - 8) ML Ops

Линейная регрессия

① Устройство: $\mathbb{Y} = \mathbb{R}$

$\mathbb{X} = \mathbb{R}^d$ - все признаки веществ.

$$x = (x_1, \dots, x_d)$$

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j$$

↓ ↑
 свободный веса
 коэф.
 (сдвиг/
 bias/intercept)

↑
 коэффициенты
 weights

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d, \quad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^d$$

$$a(x) = w_0 + \langle w, x \rangle$$

Предположение: есть коэффициенты признаков $x_i \in \mathbb{R}$

$$\dots + w_1 x_1 + \dots = \dots + w_1 + \dots$$

";"

Тогда: $a(x) = \langle w, x \rangle$

② Помехимость

$$a(x) = w_0 + w_1 \cdot \text{ножки} +$$

$$+ w_2 \cdot \text{стол} +$$

$$+ w_3 \cdot [\text{расстояние}]$$

← не учитываем
важно знать
зависимость

- признаки влияют на прогноз независимо

- признаки влияют линейно

⇒ для линейных моделей данные надо готовить

2.1 Категориальные признаки ($x_j \in \{c_1, \dots, c_m\}$) некодирован.

Например, район

One-hot-encoding (OHE)

$$x_j \rightarrow b_1(x_j), \dots, b_m(x_j)$$

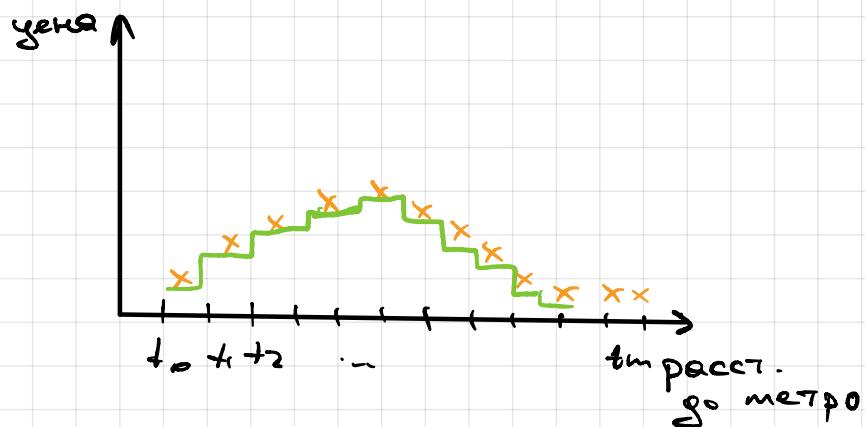
$$b_i(x_j) = [x_j = c_i]$$

Хомячки
Басмачи
Сироги и ко

x	δ	c
0	1	0
0	0	1

$$a(x) = w_0 + w_1 \cdot [хомячки] + w_2 \cdot [басмачи] + \dots$$

2.2 Бинаризация признаков



t_0, \dots, t_m - пороги

$$b_1(x_j), \dots, b_m(x_j)$$

$$b_1(x_j) = [t_0 \leq x_j < t_1]$$

:

$$b_m(x_j) = [t_{m-1} \leq x_j < t_m]$$

$$a(x) = w_0 + w_1 [t_0 \leq x_j < t_1] + \dots + w_m [t_{m-1} \leq x_j < t_m]$$

2.3 Полиномиальные признаки

ненулевые x_1
старт x_2 $\Rightarrow x_1^2, x_2^2, x_1^7 x_3^{10}$
расст. x_3

элгэ: $\sqrt{x_1}$, $\sin(x_2)$, $\log(x_3)$

проблема: неизвестно, что делать

③ Измерение ошибки в задачах регрессии

$$L(y, z)$$

\uparrow \uparrow
нраб. прогноз

$$Q(a, X) = \frac{1}{c} \sum_{i=1}^c L(y_i, a(x_i))$$

$$MSE: L(y, z) = (y - z)^2$$

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

Проблемы:

$$1) \frac{1}{c} \sum_{i=1}^c (a(x_i) - y_i)^2 = 10^7$$

- другие единицы измерения
KB. Рубли

$$RMSE = \sqrt{\frac{1}{c} \sum_{i=1}^c (a(x_i) - y_i)^2}$$

$$2) R^2 = 1 - \frac{\sum_{i=1}^c (a(x_i) - y_i)^2}{\sum_{i=1}^c (y_i - \bar{y})^2}$$

коэф.
детерминации

$$a(x_i) = y_i \rightarrow R^2 = 1$$

$$a(x_i) = \bar{y} \rightarrow R^2 = 0$$

Можно сказать, что для полиномиальных моделей

$$0 < R^2 \leq 1$$

$$MAE: L(y, z) = |y - z|$$

$$Q(a, X) = \frac{1}{c} \sum_{i=1}^c |a(x_i) - y_i|$$

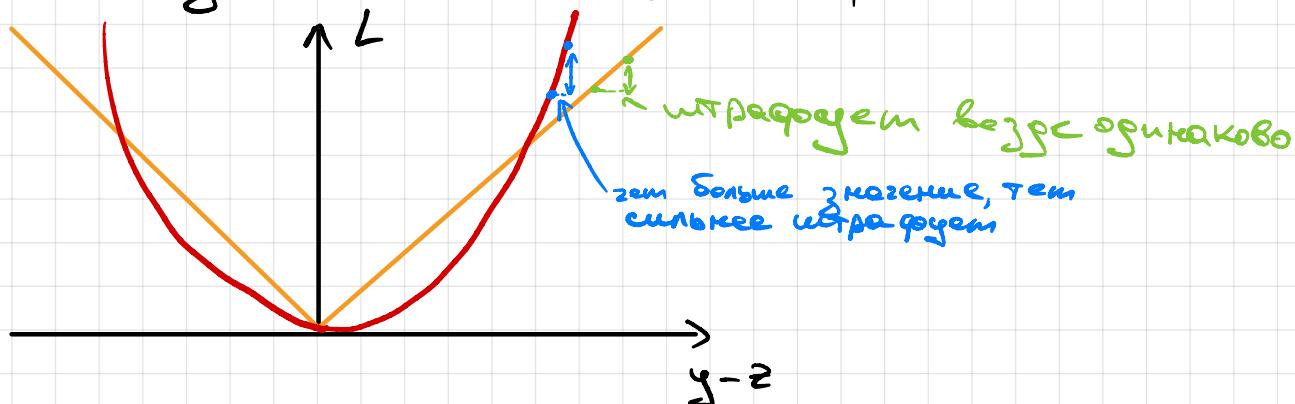
y	$Q_1(x)$	$(Q_1(x) - y)^2$	$ Q_1(x) - y $
1	2	1	1
2	1	1	1
3	2	1	1
4	8	1	1
5	6	1	1
100	7	8649	93
7	6	1	1

$$MSE = 1236 \quad MAE = 14.14$$

y	$Q_2(x)$	$(Q_2(x) - y)^2$	$ Q_2(x) - y $
1	4	9	3
2	5	9	3
3	6	9	3
4	7	9	3
5	8	9	3
100	10	8100	90
7	10	9	3

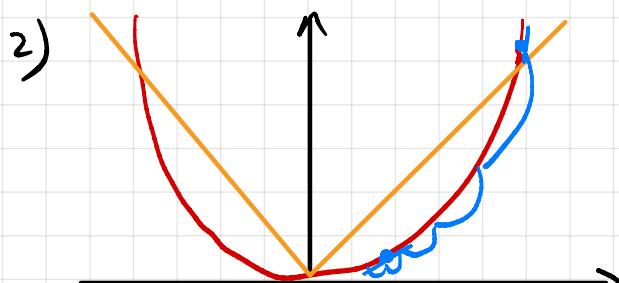
$$MSE = 1164 \quad MAE = 15.13$$

⊕ MSE стимулирует подготовку к выбросам
 MAE лучше игнорирует выбросы



Почему не всегда MAE

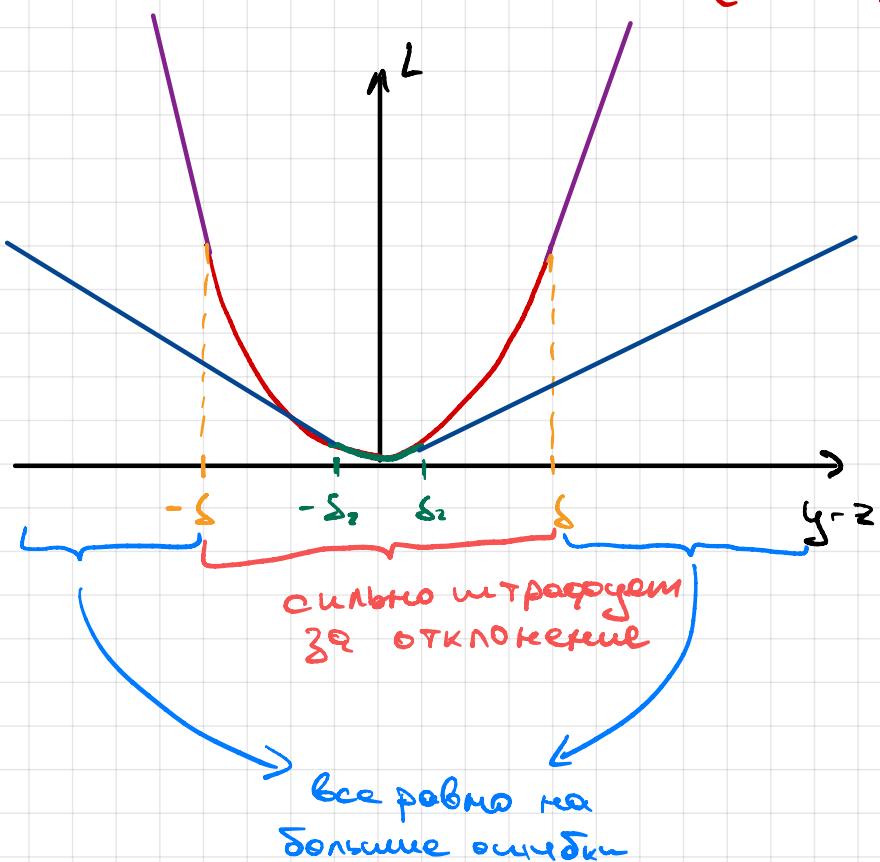
1) М.д. правило не штрафовать за большие ошибки



в MSE не правильный метод нечестно штрафует к экстремуму

и MAE например неизвестно f1
 => способом в оптимизацию

Huber Loss: $L_S(y, z) = \begin{cases} \frac{1}{2}(y-z)^2, & |y-z| < \delta \\ \delta(|y-z| - \frac{1}{2}\delta), & |y-z| \geq \delta \end{cases}$



- ❶ Чем больше δ , тем более сильные ошибки мы не считаем за выбросы
- ❷ Но проблема со второй производн.

Log - Cosh : $L(y, z) = \log \cosh(y-z)$

$$\operatorname{ch} x = \frac{e^x + e^{-x}}{2}$$

Похожая на Huber loss

+ вторая np. непрерывна

MSLE : $y \geq 0, z \geq 0$

$$L(y, z) = (\log(z+1) - \log(y+1))^2$$

Относительные по. нетрв:

$$L(y, z) = \left| \frac{y-z}{y} \right|$$

\leftarrow MAPE

$$Q(a, X) = \frac{1}{C} \sum_{i=1}^C \left| \frac{y_i - Q(x_i)}{y_i} \right|$$

+ интерпретируемость

$L=2 \rightarrow$ ошиблись в 3 раза

+ хороми при разных масштабах y

$$\left. \begin{array}{l} y_1 = 1000 \\ y_2 = 1 \end{array} \right\} \begin{array}{l} 333 \\ 0 \end{array} \quad \begin{array}{l} - \text{кестрочно} \\ - \text{стремно} \end{array}$$

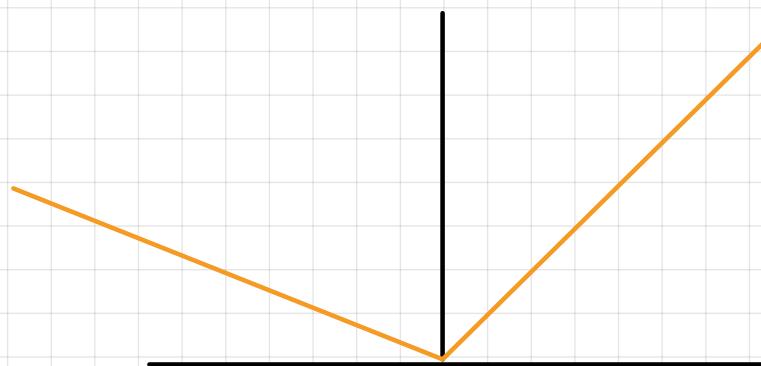
$$MAPE = \frac{1}{L} \sum_{i=1}^L \left| \frac{1}{|y_i|} \cdot (a(x_i) - y_i) \right|$$

бес: тем больше $|y_i|$, тем
менее стрично ошиблись

$$SMAPE: L(y, z) = \frac{|y-z|}{(|y|+|z|)/2}$$

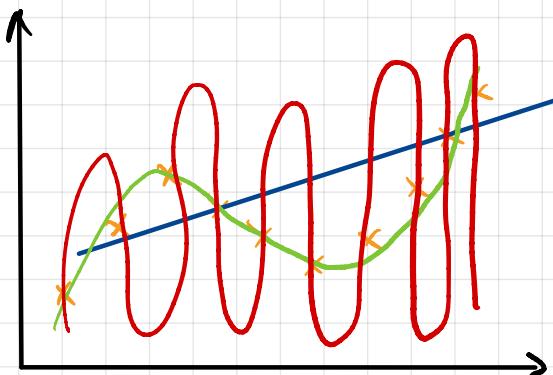
Квадратичные ф. потерь:

$$L(y, z) = (1-\tau) [y-z < 0] (y-z)^2 + \tau [y-z \geq 0] (y-z)^2, \tau \in (0,1)$$



Большие штрафы за перенпрогноз, чем за недопрогноз
(смк недобор)

④ Переобучение



$a(x) = w_0 + w_1 x - \text{недобуз.}$

$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 - \text{недобуз.}$

$a(x) = w_0 + w_1 x + \dots + w_k x^k + \dots - \text{переобуз.}$

- переобуз.

Лекция 3, 22.09.2023

Ошибки на новых данных \rightarrow ошибки на обуз.

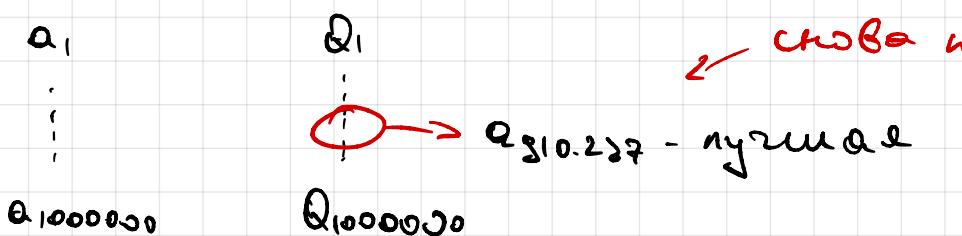
\Rightarrow переобучение (overfitting)

Op. Обобщающая способность (generalization) - ошибка на новых данных такая же, как и на обуз.

Оценка обуз. способности

1) отложенная выборка (hold-out set)

обуз.	тест.
-------	-------



обуз.	валид.	тест.
-------	--------	-------

Какого размера быть отложенному выборку?

Причины: 1) обуз. выборка должна быть как можно больше
2) тестовая выборка должна быть представительной

80:20 / 80:10:10

70:30 / 70:20:10

●	
---	--

всё попадает
в обузжение

2) кросс-валидация (CV)

K-число блоков (Folds)

x_1	x_2	x_3
-------	-------	-------

Объект.

Тест

$x_1 \cup x_2$

x_3

Q_1

$a_1(x)$

$x_1 \cup x_3$

x_2

Q_2

$a_2(x)$

$x_2 \cup x_3$

x_1

Q_3

$a_3(x)$

$$Q = \frac{1}{3} (Q_1 + Q_2 + Q_3)$$

$K = l - LOO$ (leave-one-out)

- Что получим:
- 1) обучить на всей выборке
 - 2) $x \rightarrow \frac{1}{3} (a_1(x) + a_2(x) + a_3(x))$

- Замечания:
- 1) если $l \gg 0$, то CV будет низким
 - 2) CV требует обучения K моделей

⑤ Обучение

$$\frac{1}{l} \sum_{i=1}^l (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_{w \in \mathbb{R}^d}$$

$$X = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ x_{l1} & \dots & x_{ld} \end{bmatrix}, \quad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_l \end{bmatrix}$$

$$Xw = \begin{bmatrix} \langle w, x_1 \rangle \\ \vdots \\ \langle w, x_l \rangle \end{bmatrix}$$

$$Q(w) = \frac{1}{l} \| Xw - y \|_2^2 \rightarrow \min_w$$

$$\nabla Q(w) = 0$$

$$w_* = \underset{d \times d}{(X^T X)^{-1}} X^T y$$

Проблемы: 1) матрица может быть вырожд.

2) Обращение матрицы $O(d^3)$

$d = 10^6$ - много

3) если $L(y, z)$ более хитрая, то
решить $\nabla Q(w) = 0$ не получится

\Rightarrow нужны другие методы

⑥ Градиентные методы обучения

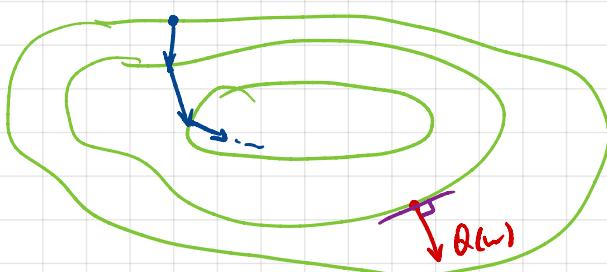
$$Q(w_1, \dots, w_d) \rightarrow \min_{w_1, \dots, w_d}$$

Пусть Q - гладко.

$$\nabla Q(w) = \begin{bmatrix} \frac{\partial Q}{\partial w_1} \\ \vdots \\ \frac{\partial Q}{\partial w_d} \end{bmatrix} \quad - \text{Градиент}$$

Важные свойства: 1) $\nabla Q(w)$ показывает напр. наискор. роста в w
 $\Rightarrow -\nabla Q(w)$ - убывающий

2) $\nabla Q(w)$ ортогональен линии уровня



Градиентный спуск

$w^{(0)}$ - начальное значение (нуль-вектор)

мат: $w^{(k)} = w^{(k-1)} - \eta \nabla Q(w^{(k-1)})$

стк
гиперплоскость
learning rate

остановка критерия: а) $\|w^{(k)} - w^{(k-1)}\| < \epsilon$

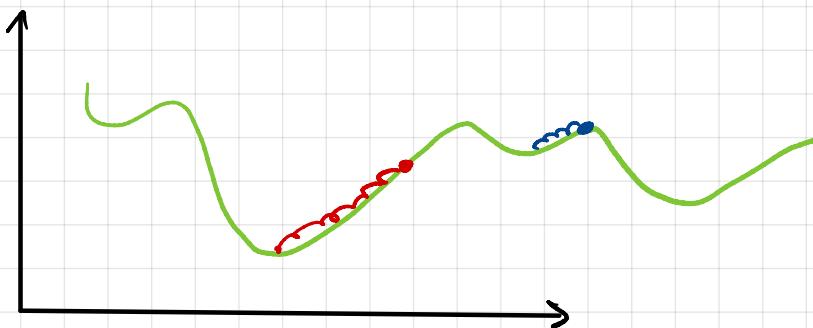
б) $|Q(w^{(k)}) - Q(w^{(k-1)})| < \epsilon$

в) $\|\nabla Q(w^{(k)})\| < \epsilon$

г) $k > N$

г) ошибка на отдельном выборке
перестала уменьшаться

Проблемы: 1) локальные минимумы



"зацикливание" - низкий старт ~ другие эпистрофы

2) много условий сходимости

а) $Q(w)$ - выпукл. и гладк.

б) $\nabla Q(w)$ - линейно убыва

$$\|\nabla Q(w_1) - \nabla Q(w_0)\| \leq L \cdot \|w_1 - w_0\|$$

в) η не очень большой ($\in \frac{1}{L}$)

\Rightarrow гарантируется, что градиентный спуск сойдет к минимуму

3) с лин. моделями и адекватными до. матер.

$Q(w)$ могут всегда быть выпуклые

$$4) Q(w^{(k)}) - Q(w_*) = O(\frac{1}{k})$$

6.1

Оценивание градиента

$$Q(w) = \frac{1}{L} \sum_{i=1}^L \underbrace{L(y_i, \alpha(x_i, w))}_{\varrho_i(w)}$$

$$Q(w) = \frac{1}{L} \sum_{i=1}^L \varrho_i(w)$$

$$\nabla Q(w) = \frac{1}{L} \sum_{i=1}^L \nabla \varrho_i(w)$$

$L = 10^6 \Rightarrow 10^6$ градиентов \Rightarrow GD непрактичный

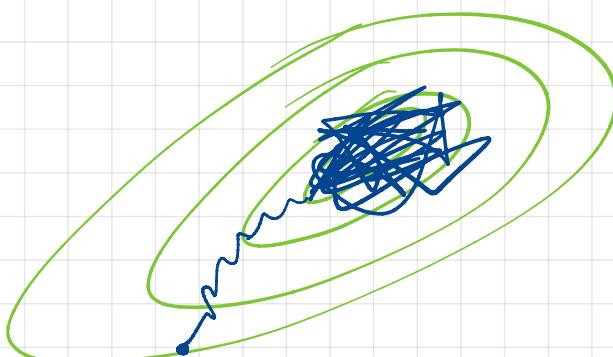
6.1.1

Стochasticеский град. спуск (SGD)

$$\nabla Q \approx \nabla \varrho_i(w)$$

мат SGD: i_k - индекс случайного объекта

$$w^{(k)} = w^{(k-1)} - \eta \nabla \varrho_{i_k}(w^{(k-1)})$$



проблема:

если $\|w^{(k)} - w_*\| \gg 0$, то

$$\nabla \varrho_{i_k}(w) \approx \nabla Q(w)$$

(увеличивает ошибку на всех объектах сразу)

если $\|w^{(k)} - w_*\| \approx 0$, то

$$\nabla \varrho_{i_k}(w) \neq \nabla Q(w)$$

недо: $w^{(k)} = w^{(k-1)} - \eta_k \nabla \varrho_{i_k}(w^{(k-1)})$

$$\begin{aligned} \sum \eta_k &= \infty \\ \sum \eta_k^2 &< \infty \end{aligned}$$

\Rightarrow SGD сойдется к минимуму (если урагаем с $w^{(0)}$)

условия
Роджинса-
Монро

$$\eta_k = \frac{1}{l} \left(\frac{s_0}{s_0 + k} \right)^p - s_1, s_0 \text{ и } p - можно подбирать$$

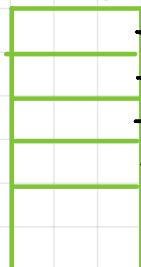
Скорость сходж.: $E(Q(w^{(k)}) - Q(w_*)) = O(\frac{1}{\sqrt{k}})$

Замечания: 1) mini-batch GD

$$\nabla Q(w) \approx \frac{1}{l} \sum_{j=1}^l \nabla q_{i,j}(w)$$

2) SGD хорошо для онлайн-обуч.

HDD



→ генерим по шагу SGD,
последовательно считывая
объекты с диска

но это обуславливается не ограниченных
выборках

6.1.2 SAG (stochastic average gradient)

$$Q(w) = \frac{1}{l} \sum_{i=1}^l q_i(w)$$

$$z_i^{(0)} = \nabla q_i(w^{(0)})$$

Итерация SAG: $i_k \sim \{1, \dots, l\}$ на каждом шаге
пересчитываем только
один объект

$$z_i^{(k)} = \begin{cases} \nabla q_i(w^{(k-1)}), & \text{если } i = i_k \\ z_i^{(k-1)}, & \text{иначе} \end{cases}$$

$$\nabla Q(w^{(k-1)}) \approx \frac{1}{l} \sum_{i=1}^l z_i^{(k)}$$

$$w^{(k)} = w^{(k-1)} - \frac{1}{l} \sum_{i=1}^l z_i^{(k)}$$

$$E(Q(w^{(k)})) - Q(w_*) = O(\frac{1}{k})$$

Почему это гораздо лучше GD?

$$\nabla Q(w^{(k-1)}) = \frac{1}{l} \sum_{i=1}^l z_i^{(k-1)} = G_{k-1}$$

$$\text{безразн. } j: \quad \nabla Q(w^{(k)}) = (G_{k-1} - z_j^{(k-1)} + z_j^{(k)}) \cdot \frac{1}{l}$$

Но такого теряется все $z_j \in \mathbb{R}^d$, $O(l \cdot d)$ памяти

Ко если имеем гено с лин. моделью, то

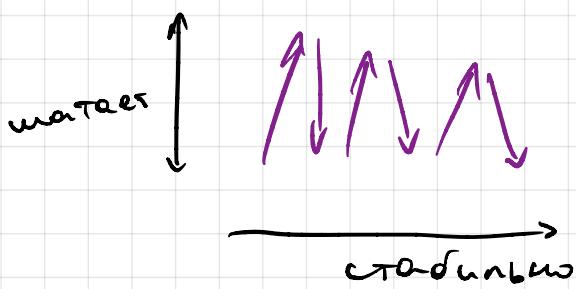
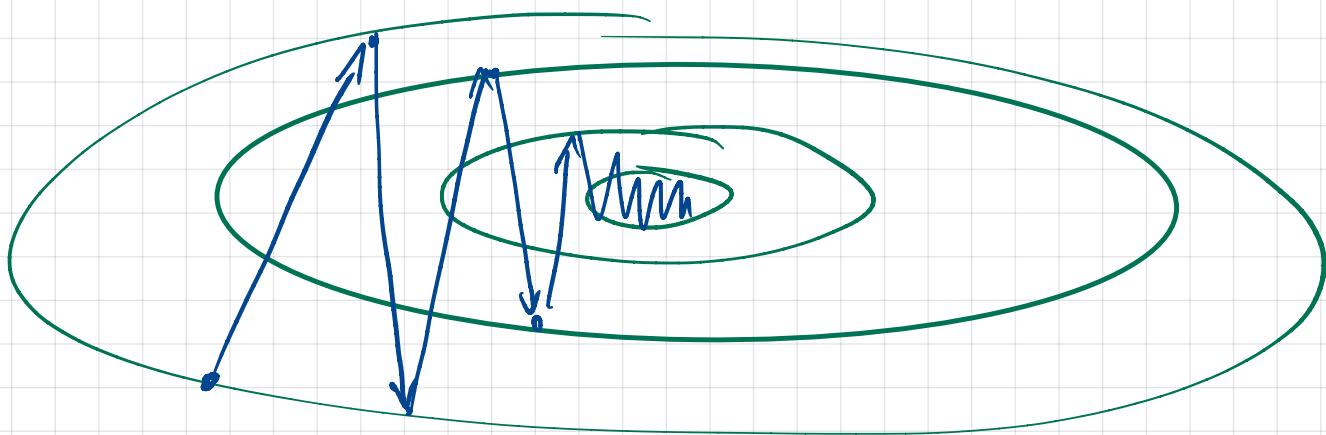
$$q_i(\omega) = L(y_i, \langle \omega, x_i \rangle) = \tilde{q}_i(\langle \omega, x_i \rangle) \Rightarrow$$

\Rightarrow можно хранить только $\langle \omega, x_i \rangle \in \mathbb{R} \rightarrow O(C)$

Лекция 4. 29.08.2023

(6.2) Моделирование (GD)

(6.2.1) Momentum (метод имерзии)



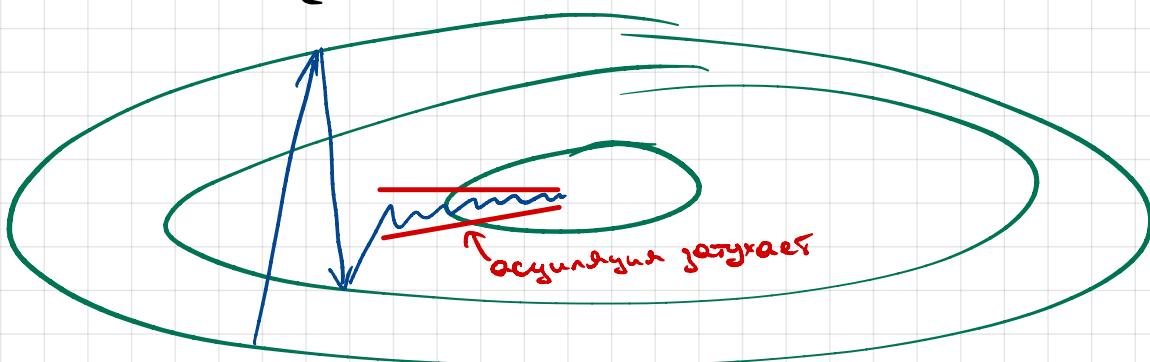
$\omega^{(0)}$ - иниц.

$h_0 = 0$ - Вк. имерзии

шаг:

$$\begin{cases} h_k = \alpha \cdot h_{k-1} + \eta_k \nabla_{\omega} Q(\omega^{(k-1)}) \\ \omega^{(k)} = \omega^{(k-1)} - h_k \end{cases}$$

$$\alpha = 0.9$$



6.2.2. Адабтивный шаг

	1	2	3	4	5
1	0	0	1	0	0
2	1	0	0	0	0
3	1	1	0	0	0
4	:	:	:	:	0
5	:	:	:	:	0
6	:	:	:	:	0
7					:

SGD где $\omega, x \rightarrow h_k$
(batch-size=1)

w_3

w_1

w_1, w_3

w_5

w_5

если значение $x_i = 0$, то
затраты взвешены в пропорции
 $0 \cdot g_i(\omega^{(k)}) = 0$

AdaGrad : $G_{0j} = 0$, j - номер признака

$$G_{kj} = G_{k-1,j} + (\nabla Q(\omega^{(k-1)}))_j^2$$

Несколько сильно мы уже
"одушили" ω_j

$$\omega_j^{(k)} = \omega_j^{(k-1)} - \frac{\eta_k}{\sqrt{G_{kj} + \epsilon}} (\nabla Q(\omega^{(k-1)}))_j$$

G_{kj} - мал. \Rightarrow Дельтое

G_{kj} - бол. \Rightarrow мал.

- Ко G_{kj} может быстро расти (за пару шагов) и мы
перестанем одушияться

Модифик.: RMSProp : $G_{kj} = \alpha G_{k-1,j} + (1-\alpha) (\nabla Q(\omega^{(k-1)}))_j^2$

$$\alpha \in (0, 1)$$

$$\alpha = 0,999$$

6.2.3 Adam = Momentum + AdaGrad (B DL)

7

Регуляризация

- Известный факт:

- лич. модель переодушина \rightsquigarrow большие веса

Почему?

Объяснение 1:

Пусть есть лин. зав. признаки

$\exists \omega \in \mathbb{R}^d : \forall x \in X \quad \langle \omega, x \rangle = 0$

$\omega_* - \text{решение} \quad \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \omega, x_i \rangle - y_i)^2 \rightarrow \min_{\omega}$

$$\langle \omega_* + \lambda \vartheta, x \rangle = \langle \omega_*, x \rangle + \lambda \underbrace{\langle \vartheta, x \rangle}_{=0} = \langle \omega_*, x \rangle$$

т.е. $\omega_* + \lambda \vartheta$ - тоже решение

\Rightarrow решений бесконечно много и можно найти очень
多样性ное

Объяснение 2:

$$Q(x) = 10^8 \cdot \text{площадь} - 10^8 \cdot \text{этажи} + 10^6 \cdot \{\text{хамовники}\}$$

$$10^8 \cdot (\text{площадь} + 0,001) = 10^8 \cdot \text{площадь} + \underbrace{\frac{10^8 \cdot 0,001}{10^5}}_{\text{небольшое влияние}}$$

Гиперчувствительность к изменениям признаков - это соответствует тому, как устроен мир

Идея: запретить большие веса

$$Q(\omega) + \lambda R(\omega) \rightarrow \min$$

$$\textcircled{1} \quad R(\omega) = \|\omega\|_2^2 = \sum_{j=1}^d \omega_j^2 - L_2\text{-регуляризация}$$

$$\textcircled{2} \quad R(\omega) = \|\omega\|_1 = \sum |\omega_j| - L_1\text{-рег.}$$

$$\lambda - \text{коэф. рег.} : \quad \lambda \gg 0 \Rightarrow \omega_* = 0$$

$$\lambda = 0 \Rightarrow \text{нет рег.}$$

λ нужно подбирать по общ. - гиперпараметр

\Rightarrow подбираем по новым данным (отл. вид., CV)

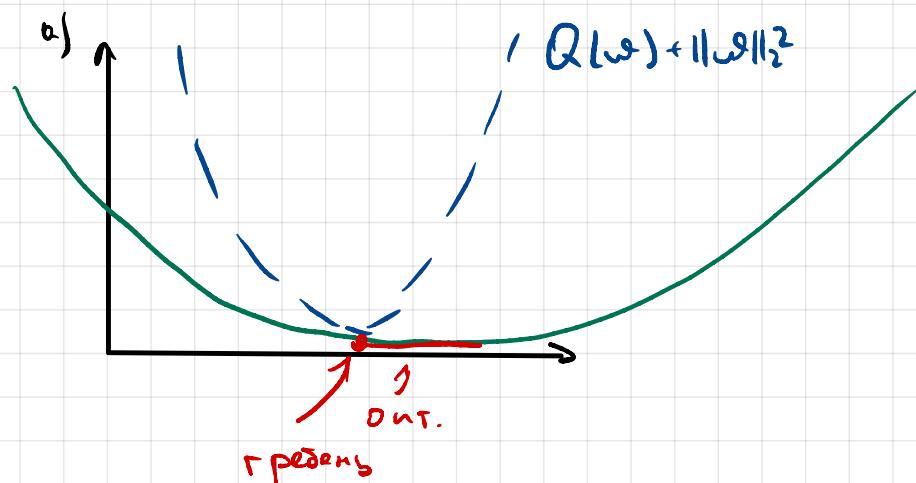
Стратегии подбора: $\textcircled{1}$ Grid Search

$\textcircled{2}$ Random Search

$\textcircled{3}$ AutoML

Критерий: ① $\frac{1}{2} \sum_{i=1}^n (\langle \omega, x_i \rangle - y_i)^2 + \lambda \|\omega\|_2^2 \rightarrow \min \omega$

Ridge - регрессия



$$\delta) \quad \omega_x = \left(\underbrace{\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}}_{\text{C.з.} \geq 0 + \lambda} \right)^{-1} \mathbf{X}^\top \mathbf{y}$$

$\text{C.з.} \geq \lambda$

\Rightarrow только есть решения

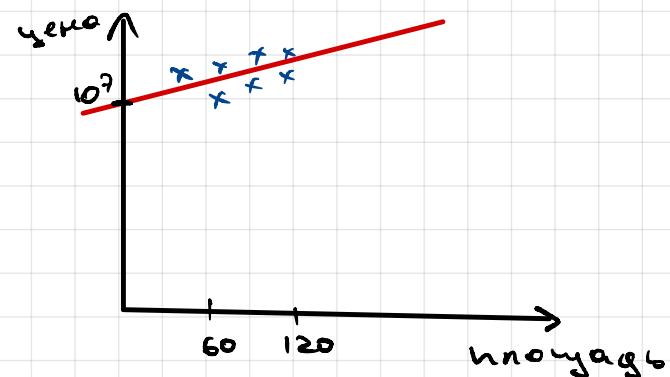
② $\frac{1}{2} \sum_i (\langle \omega, x_i \rangle - y_i)^2 + \lambda \|\omega\|_1 \rightarrow \min \omega$

LASSO

❶

• Замыкает засыпь весов

Замечание 1



$$a(x) = 10^7 + 100 \cdot \text{иподадь}$$

❷

ω нельзя регуляризовать

Большой ω не означает гиперчувствительности к малым изменениям признаков

Замечание 2

Пусть есть признаки разных масштабов

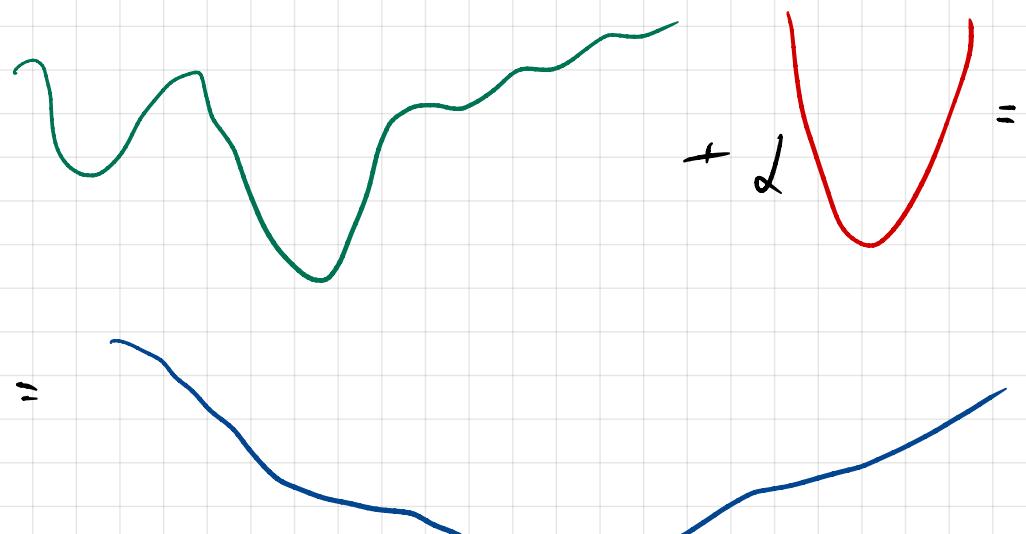
$5-7$ время поиск контакта $\xrightarrow{10^2}$ расстояния (m)
 (мин) $\xrightarrow{10^{-5}}$ минут ≈ 500.000 ионизирующей мощности (мин)

страгоуем сильнее из-за масштаба

$\rightarrow \|w\|$ ведёт себя неадекватно при неодинаковых масштабах

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad x_j = \prod$$

Замечание 3



- "Упрощает" рельеф $Q(\omega) \Rightarrow$ ускоряет GD

⑧ Разреженные модели

Зачем затягивать веса?

- ① Несколько близких признаков всё носит
- ② Ускорение обучения
- ③ $N \ll p$
 $l \ll d$

Решение: l_1 - первая признаки

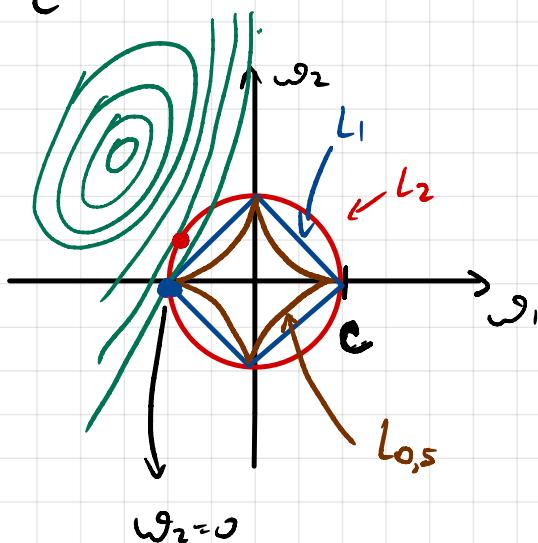
$$Q(\omega) + \lambda \|\omega\|_1 \rightarrow \min_{\omega}$$

Объяснение 1:

$$Q(\omega) + \lambda \|\omega\|_1 \rightarrow \min_{\omega}$$

\uparrow
смкн кнк. С

$$\begin{cases} Q(\omega) \rightarrow \min \\ \|\omega\|_1 \leq C \end{cases}$$



Объяснение 2:

$$0 < \delta < \varepsilon \ll 1$$

$$\omega = (1, \varepsilon)$$

$$\|\omega - (\delta, 0)\|_2^2 = 1 - 2\delta + \delta^2 + \varepsilon^2$$

$$\|\omega - (\delta, 0)\|_1 = 1 - \delta + \varepsilon \quad \checkmark$$

$$\|\omega - (0, \delta)\|_2^2 = 1 - 2\varepsilon\delta + \delta^2 + \varepsilon^2$$

$$\|\omega - (0, \delta)\|_1 = 1 - \delta + \varepsilon$$

$C \rightarrow 3$. L_1 -пер. — ребаразмо, какой козырь уменьшить

$C \rightarrow 3$. L_2 -пер. — ближнее уменьш. Дополнение веса \Rightarrow
 \Rightarrow никогда не тронет вес близких к 0

Объяснение 3:

Проекционный метод — миним. функ. = функ. + Весы
 $Q(\omega) + \lambda \|\omega\|_1$

$$\omega^{(k)} = S_{\eta^k} (\omega^{(k-1)} - \eta \sigma_w Q(\omega^{(k-1)}))$$

$$S_{\eta^k}(\omega_i) = \begin{cases} \omega_i - \eta^k, & \omega_i \geq \eta^k \\ 0, & |\omega_i| < \eta^k \\ \omega_i + \eta^k, & \omega_i \leq -\eta^k \end{cases}$$

речает

на каждом шаге сдвигает веса и затягивает близкие к нулю

Линейная классификация

$\mathcal{Y} = \{1, \dots, k\}$ - многоклассовая кн.

$\mathcal{Y} = \{-1, +1\}$ - бинарная - разбираем её

$\langle \omega, x \rangle \in \mathbb{R}$ - a шаг $\{-1, +1\}$

$$Q(x) = \text{sign} \langle \omega, x \rangle$$

$$Q(x) = \text{sign} (\langle \omega, x \rangle - t)$$

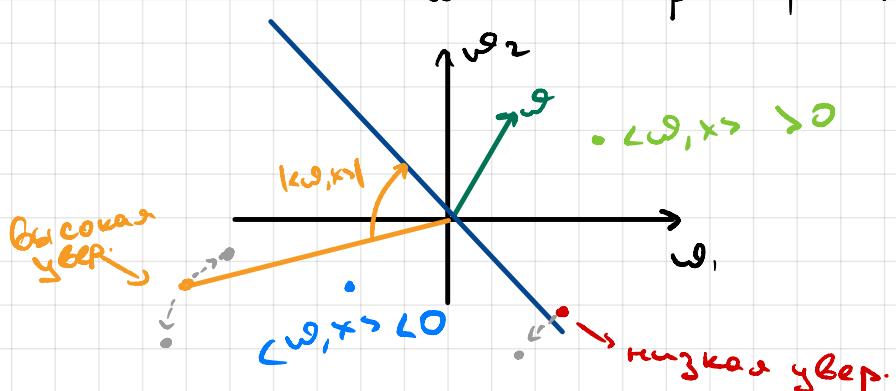
t
порог

но $\text{sign}(0) = 0 \Rightarrow \langle \omega, x \rangle = 0$:

- 1) не вывоет
- 2) отказ от классиф.
- 3) сл. класс

Геометрия: $\langle \omega, x \rangle = 0$ - ур-е гиперплоскости

ω - вектор нормали



т.е. лин. классиф.
разделяет классы
гиперплоскостью

$|\langle \omega, x_i \rangle|$ - тем дальше, тем дальше x от гиперпл.
 говорит об уверенности

Фундаментальная ошибка

$$Q(\omega) = \frac{1}{l} \sum_{i=1}^l [\alpha(\omega, x_i) \neq y_i] - \text{error rate}$$

$$\frac{1}{l} \sum_{i=1}^l [\text{sign } \langle \omega, x_i \rangle \neq y_i] \rightarrow \min_{\omega} - \text{не проходит мажи}$$

$$= \frac{1}{l} \sum_{i=1}^l [y_i \cdot \langle \omega, x_i \rangle < 0]$$

$$y_i \cdot \langle \omega, x_i \rangle > 0 \Rightarrow y_i = \text{sign}(\langle \omega, x_i \rangle)$$

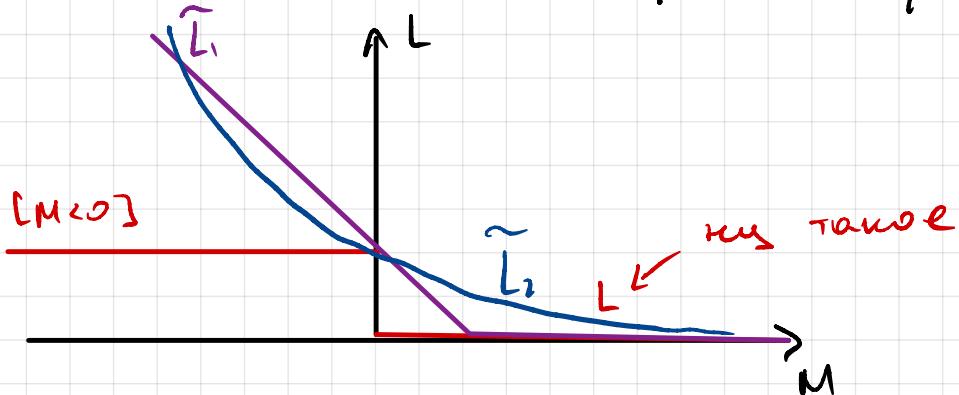
$$y_i \cdot \langle \omega, x_i \rangle < 0 \Rightarrow y_i \neq \text{sign}(\langle \omega, x_i \rangle)$$

$$M_i = y_i \cdot \langle \omega, x_i \rangle - \text{отрыв (margin)}$$

$\text{sign } M_i$ - корректность кл.

$|M_i|$ - уверенность

$$L(M) = [M < 0] - \text{нормальная ф. потерь}$$



Но есть: $[M < 0] \leq \tilde{L}(M)$ - эндо. верхняя оценка

$$0 \leq \frac{1}{l} \sum_{i=1}^l [y_i \cdot \langle \omega, x_i \rangle < 0] \leq \frac{1}{l} \sum_{i=1}^l \tilde{L}(y_i \cdot \langle \omega, x_i \rangle) \rightarrow \min$$

если хорошо проанализ.,
то и левая будет мал.

$$\textcircled{1} \quad \tilde{L}_1(M) = \max(0, 1-M) - \text{hinge loss}$$

$$\textcircled{2} \quad \tilde{L}_2(M) = \log(1 + \exp(-M)) - \text{логистическая}$$

$$③ \tilde{L}_3(M) = e^{-M}$$

$$④ \tilde{L}_4(M) = \frac{2}{1+e^M}$$

$$⑤ \tilde{L}_5(M) = \frac{\arctan(-M)}{\pi} + 1$$

Разные оценки дают разные сб-ва

$$\begin{cases} \sum_{i=1}^l \tilde{L}(y_i \cdot \langle \omega, x_i \rangle) + \alpha R(\omega) \rightarrow \min \end{cases}$$

Дальше всё стандартно

Лекция 6. 13.10.2023

Метрики качества классификации

① Доля верных ответов (accuracy) - не точность

$$\frac{1}{l} \sum_{i=1}^l [\alpha(x_i) = y_i]$$

Проблема с дисбалансом классов

$$\begin{array}{ll} +1: 50 & \alpha(x) = +1 \Rightarrow acc = 85\% \\ -1: 850 & \end{array}$$

Мораль: вместе с accuracy надо смотреть на баланс классов

- Offtop

Когда мы обучаем модель, есть 2 способа оценить прирост:

r_1 - одна ош. 80
 r_2 - две

$ r_1 - r_2 $	r_1, r_2	n.n. адс.	погрешн. ошк.
1)	$20\%, 10\%$	100%	50%
2)	$50\%, 25\%$	25%	50%
	$0.1\%, 0.01\%$	0.005%	50%

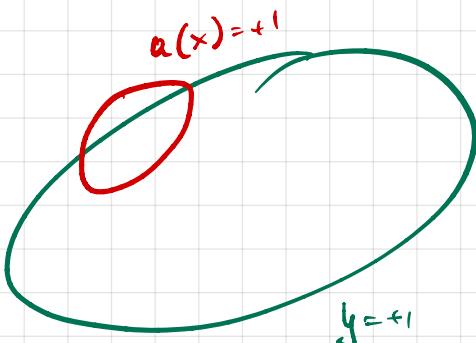
② Матрица ошибок

	$y=+1$	$y=-1$
$a=+1$	TP	FP
$a=-1$	FN	TN

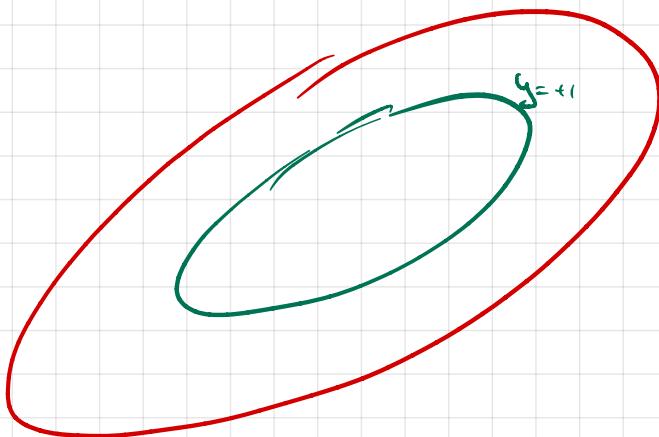
$$acc = \frac{TP + TN}{TP + FP + TN + FN}$$

$$precision = \frac{TP}{TP + FP} \rightarrow \text{сколько можно gob. модели, если } a(x) = +1$$

$$recall = \frac{TP}{TP + FN} \rightarrow \text{сколько модель корректно знает класс}$$



precision ↑
recall ↓



precision ↓
recall ↑

Как балансировать между точностью и полнотой?

$$a(x) = \text{sign}(b(x)) , \quad b(x) = \langle w, x \rangle - \text{уверенность}$$

sort по $b(x)$

$\langle w, x \rangle$

105	+1
85	+1
87	-1
⋮	
-80	+1
-80	-1

precision ↑
recall ↓

Обычно точность и полноту регулируют выбором порога

$$a(x) = \text{sign}(b(x) - t)$$

порог - гиперпараметр

precision ↓
recall ↑

После обуздания модели начнем ее подбирать.

$$\begin{cases} \text{precision} \rightarrow \max \\ \text{recall} \geq 0,8 \end{cases}$$

или

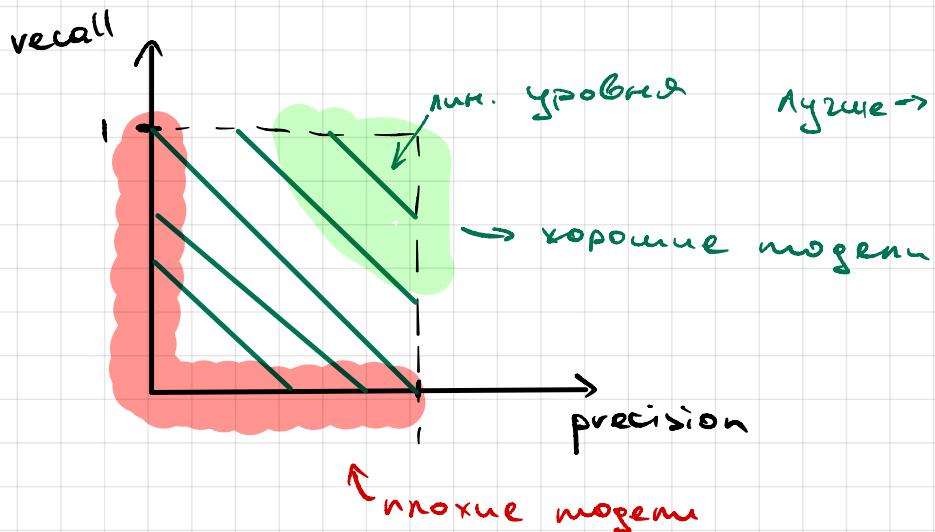
$$\begin{cases} \text{recall} \rightarrow \max \\ \text{precision} \geq 0,6 \end{cases}$$

Проблема: показателей 2 - как оценить?

(3) Прокси метрика

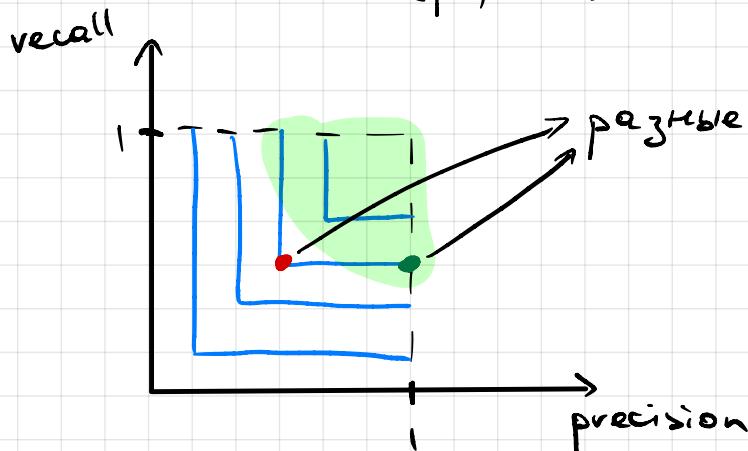
1) Ср. арифм.: $f = \frac{\text{pr} + \text{rec}}{2}$

$$\begin{array}{l} \text{pr} = 0,05 \xrightarrow{\text{worse than const}} \\ \text{rec} = 0,8 \end{array} \rightarrow f = 0,475$$



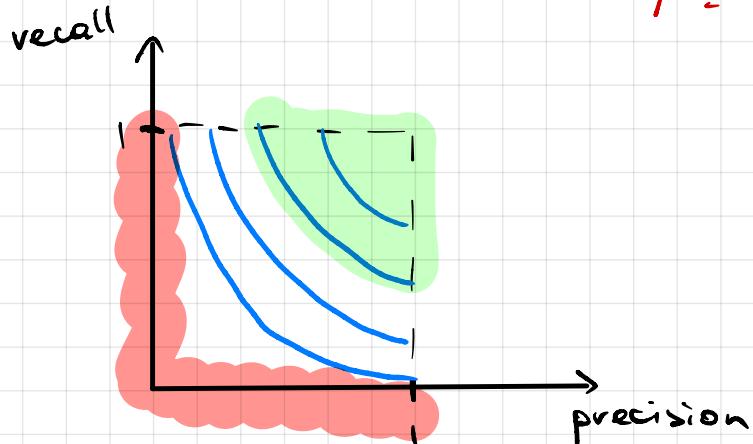
2) Максимум

$$M = \min(\text{pr}, \text{rec})$$

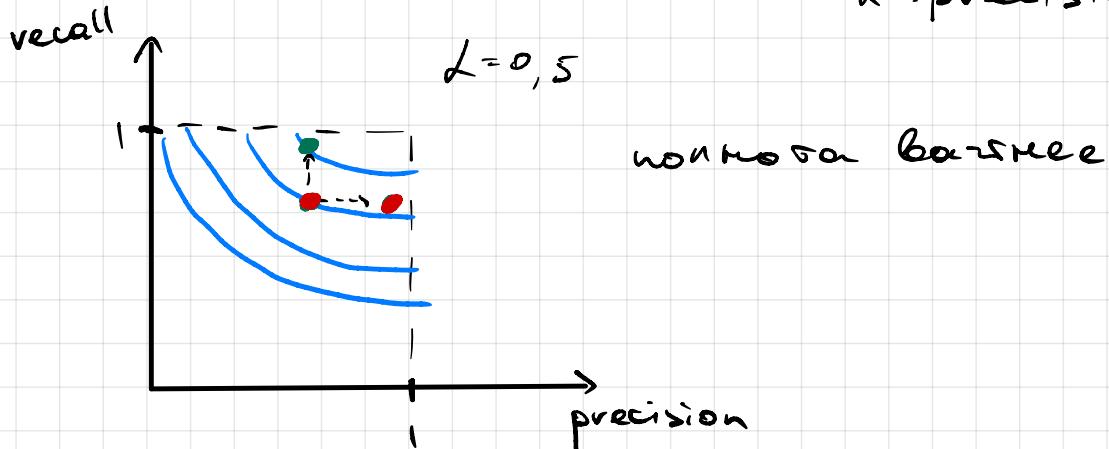


3) Среднее гармоническое

$$F = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



$$\text{Модифицированное: } F = \frac{(1+\alpha^2) \text{precision} \cdot \text{recall}}{\alpha^2 \cdot \text{precision} + \text{recall}}$$



4) Среднее геометрическое

$$G = \sqrt{\text{precision} \cdot \text{recall}} - \text{лучше чем } F$$

$$\begin{array}{ll} \text{Pr} = 0,3 & F = 0,18 \\ \text{rec} = 0,1 & G = 0,3 \end{array}$$

$$G \geq F$$

5) DFFtop

$$\text{Lift} = \frac{\text{precision}}{(\text{TP} + \text{FN})/l} = \frac{\text{precision}}{l/c}$$

Модель предсказания оттока

$$\text{Lift} = 1 \Rightarrow \text{модель} = \text{идеал}$$

$$\text{Lift} < 1 \Rightarrow \text{:()$$

Если правильных
запросов при исп. модели
→
Если правильных
запросов при случайном
объекте
(или если из запросов
все дают ответ)

④

Площадь под кривыми
(метрики качества реш.)

$$Q(x) = \text{Sign} (b(x) - +), \quad b(x) - \text{уверенность } B + 1$$

$b(x)$	y	y^*
100	+1	+1
93	+1	-1
87	+1	+1
⋮	⋮	⋮
1	+1	-1
-2	-1	+1
⋮	⋮	⋮
-30	-1	+1

Как измерить качество $b(x)$ в целом?

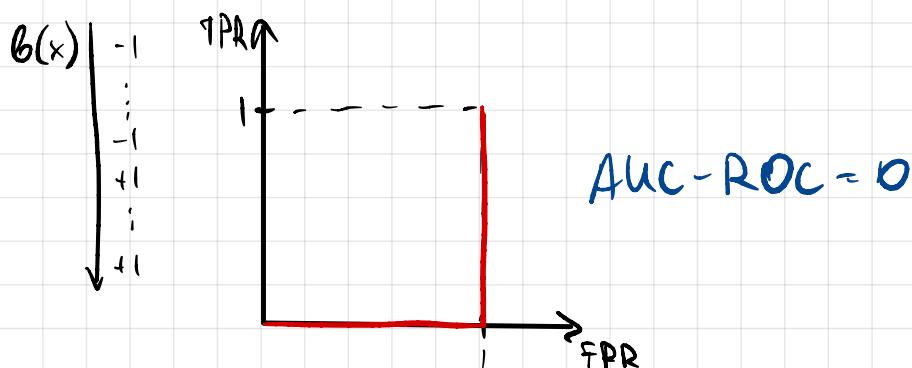
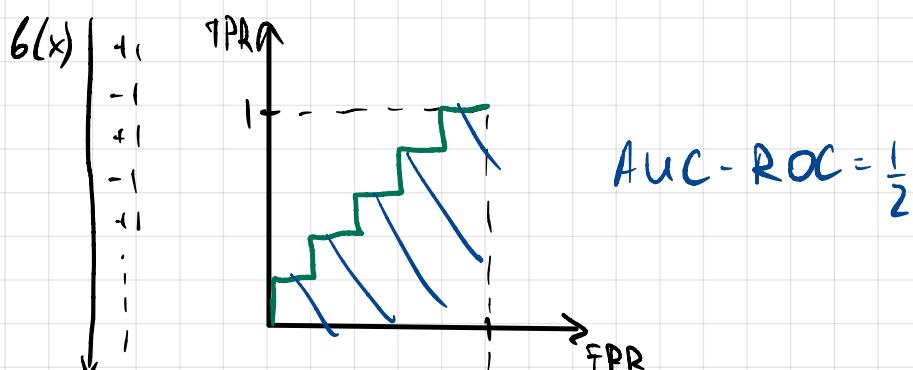
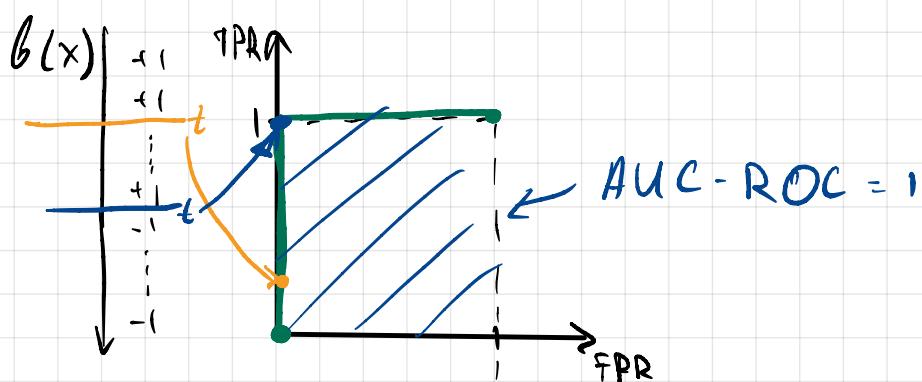
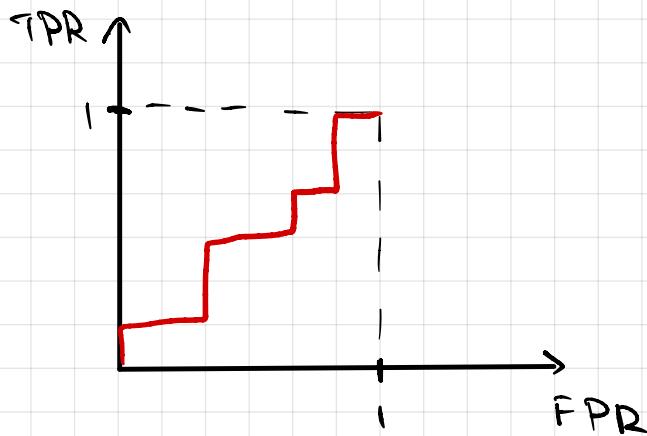
4.1 ROC-кривая

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{C_-}$$

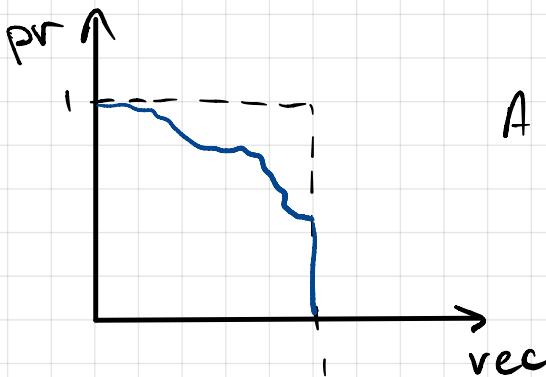
$$TPR = \frac{TP}{TP + FN} = \frac{TP}{C_+} = \text{recall}$$

Перебираем все пороги для каждого приема

TPR и FPR



4.2 PR-крубы

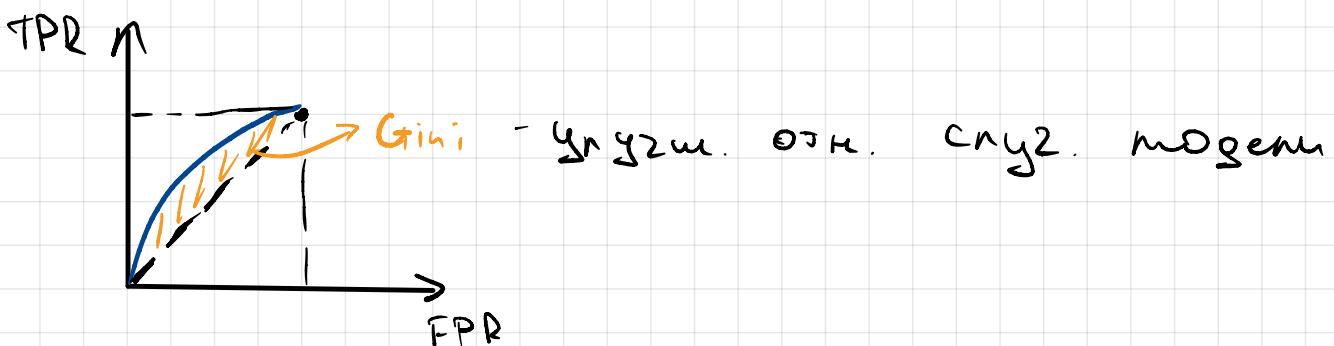


AUC-PRC

$t \downarrow \rightarrow$ vec \uparrow
 \rightarrow pr?

Замечание 1 Недекс Азуми

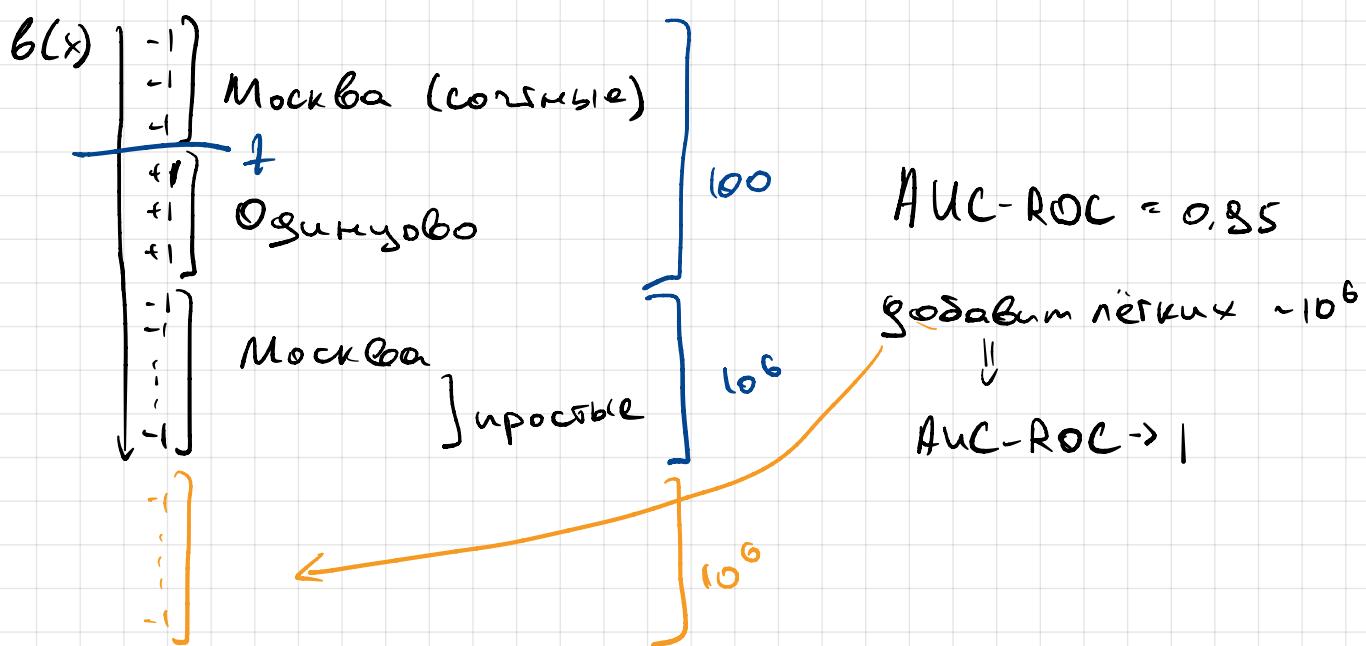
$$G_{ini} = 2 \text{AUC-ROC} - 1$$

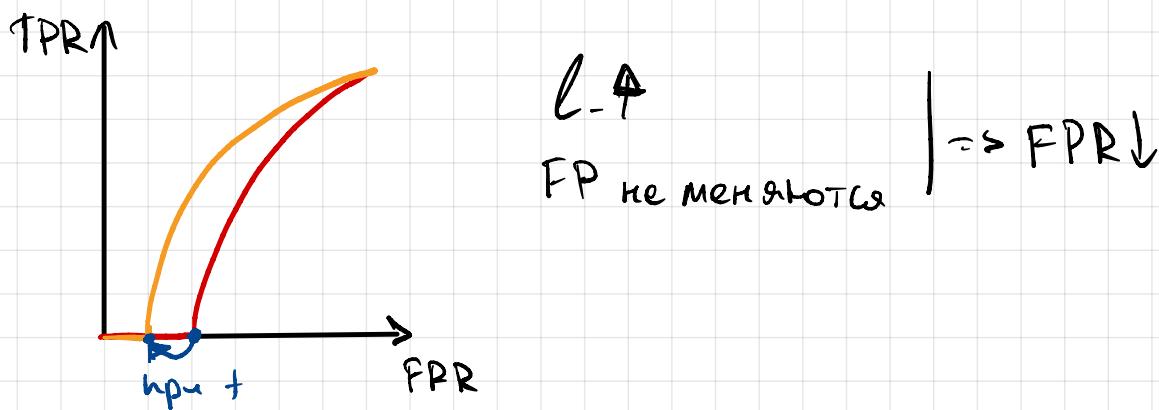


Sammlung 2

AIC-R0C может впасть в заблуждение, если много лёгких объектов

Задача: Отличить Москву от Одинцово по фото





но AUC - PRC будет идти синхронно и не изменится
 \Rightarrow нужно учитывать гидбаланс

Лекция 7. 20.10.2023

Методы классификации

$$\frac{1}{\ell} \sum_{i=1}^{\ell} [\text{sign}(\omega, x_i) \neq y_i] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(y_i, \text{sign}(\omega, x_i)) \rightarrow \min_{\omega}$$

$$\tilde{L}_1(M) = \log(1 + e^{-M})$$

$$\tilde{L}_2(M) = \max(0, 1 - M)$$

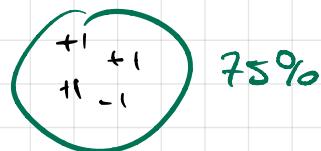
:

① Логистическая регрессия

①.1 Хочем оценивать вер. классов (уверенность модели в своём ответе)

$a(k) = +1$ — модель уверена на 80%

Интерпретация: Если взять все объекты с ув. 80%, то из них 80% будут иметь $y = +1$



② А зачем нам это?

1) $b(x)$ — вер. тб, что клиент уйдёт

$$b(x) \geq 0.8 \Rightarrow \text{заболт}$$

среди них $\leq 10\%$ не уходит (логично номоз.)

т.е. можем искать подбором не от баллов

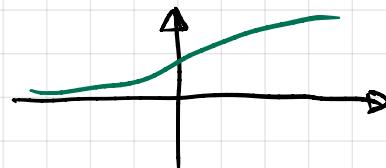
2) Баннерная реклама

	$b(x)$	$C(x)$	мат. ож.
1	0.1	1	0.1
2	0.5	0.01	0.005
3	0.01	500	5

1.3 Формализация

$b(x)$ - выдаёт вер. нологистического класса

$$(b(x) = \sigma(\langle w, x \rangle) \in (0, 1))$$



$b(x)$ - корректно оценивает вер., если для вер-ти p среди всех объектов $x \in \mathbb{X}$ с $b(x) = p$ одна нологистическая равна p

Пусть выборка из однотипных объектов:

$$x_1 = x_2 = \dots = x_n \Rightarrow b(x) = \dots = b(x_n) = b = \text{const}$$

y_1, y_2, \dots, y_n - разные

Было бы логично выдавать $b = \frac{1}{n} \sum [y_i = +1]$

$$\underset{b \in [0, 1]}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n L(y_i, b) = \frac{1}{n} \sum_{i=1}^n [y_i = +1] \quad \begin{array}{l} \text{- тогда ф. номинально} \\ \text{требует корректного} \\ \text{оценивания вер.} \end{array}$$

$n \rightarrow \infty :$

$$P(y = +1 | x)$$

$$\frac{1}{n} \sum_{i=1}^n [y_i = +1] \cdot L(+1, b) + \frac{1}{n} \sum_{i=1}^n [y_i = -1] \cdot L(-1, b) =$$

$$= L(+1, b) \cdot \frac{\sum [y_i = +1]}{n} + L(-1, b) \cdot \frac{\sum [y_i = -1]}{n} \rightarrow$$

$$P(y = +1 | x)$$

$$P(y = -1 | x)$$

$$\rightarrow E_y [L(y, b) | x]$$

$$n \rightarrow \infty : \underset{b \in [0,1]}{\operatorname{argmin}} E_y[L(y, b) | x] = p(y=+1 | x) \quad *$$

Хочим, чтобы мак. оцв. фн. натерп. давала нам вер. номог. класса

Почему это работает?

Чтобы модели не переобучались, мы сильно ограничиваем их сложность
 ↓

Если x_1 и x_2 близки, то на них прогнозы тоже будут близки $b(x_1)$ и $b(x_2)$

↓

Будут возникать группы объектов с идентичн. прогнозами

$$b(x_1) \approx b(x_2) \approx \dots \approx b(x_n) \approx b = \text{const}$$

↓

$b(x)$ будет стремиться выдавать на них корректные вер-ти в силу установленного нами требования

1.4 Для каких $L(y, z)$ выполнено (*)?

$$\textcircled{1} L(y, z) = (y - z)^2 \rightarrow (b - [y=+1])^2$$

$$E_y[(b - [y=+1])^2] = \underbrace{p(y=+1 | x)}_{P} \cdot (b-1)^2 + \underbrace{(1 - p(y=+1 | x))}_{P} \cdot (b-0)^2$$

$$\frac{\partial}{\partial b} = 2p(b-1) + 2(1-p)b = 2pb - 2p + 2b - 2pb = 0$$

$$b = p(y=+1 | x)$$

т.е. MSE будет пытаться корректно оценивать вер-ти

но это так себе идея (

$$\frac{1}{l} \sum_{i=1}^l \left(\underbrace{b(x_i)}_{\Delta(\omega, x_i)} - [y_i=+1] \right)^2 \rightarrow \min_{\omega}$$

$$\nabla_{\omega} (\nabla(\langle \omega, x_i \rangle) - [y_i = +1])^2 = 2(\dots) \cdot \nabla'(\langle \omega, x_i \rangle) \cdot \dots$$

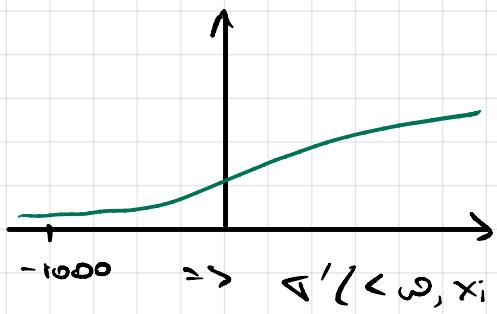


График застремвається до межкласи функції

② $L(y, b) = |b - [y_i = +1]| \rightarrow (\Rightarrow)$ не виконуємо

- ③ Кожен з об'єктів y_i є б. с. з розр. багатул.
- $b(x_i)$ - оцінка вер. $p(y_i = +1 | x_i)$

Запишем наведене:

$$\prod_{i=1}^l b(x_i)^{[y_i = +1]} \cdot (1 - b(x_i))^{[y_i = -1]} \rightarrow \max_b | -\log$$

$$\sum_{i=1}^l (-[y_i = +1] \cdot \log b(x_i) - [y_i = -1] \cdot \log(1 - b(x_i))) \rightarrow \min_b$$

log-loss / binary cross-entropy loss

(\Rightarrow) виконуємо

$$y \text{ нас } b(x) = \nabla(\langle \omega, x \rangle)$$

$$\frac{1}{l} \sum_{i=1}^l \left(-[y_i = +1] \log \frac{1}{1 + e^{-\langle \omega, x_i \rangle}} - [y_i = -1] \log \left(1 - \frac{1}{1 + e^{-\langle \omega, x_i \rangle}} \right) \right)$$

$$\dots \approx \frac{1}{l} \sum_{i=1}^l \log \left(1 + \exp(-y_i \cdot \langle \omega, x_i \rangle) \right) \rightarrow \min_{\omega}$$

$$L(y, z) = \log(1 + e^{-yz})$$

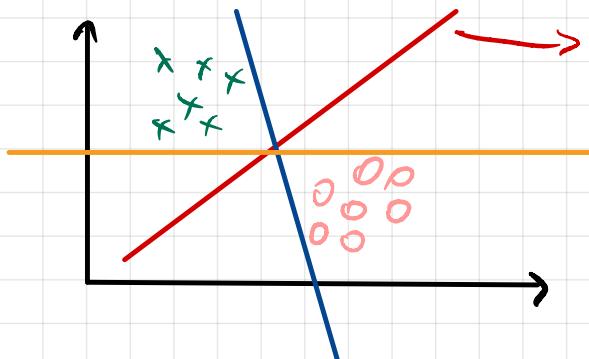
$b(x) = \Phi(\langle \omega, x \rangle)$ - норм-регресия

\downarrow ф. розр. $\mathcal{N}(0, 1)$ probability-unit

$b(x) = \nabla(\langle \omega, x \rangle)$ - лог-регресия

② Метод подобных векторов (SVM)

$$\sum_i (M_i) = \max(0, 1 - M) - \text{норма?}$$



- выглядит лучше всего
- генерирует минимальную предположенную о грешности
- максимальное удаление от грешности

2.1 Линейно разделимый случай

$$\exists \omega : y_i \cdot \langle \omega, x_i \rangle > 0 \quad \forall i=1 \dots l$$

$$a(x) = \text{sign}(\langle \omega, x \rangle + \omega_0) = \begin{cases} 1 & \cdot \omega \geq 0 \\ -1 & \cdot \omega < 0 \end{cases}$$

$= \text{sign}(\langle \omega, x \rangle + \omega_0) \Rightarrow$ ищем нормированные как хотим

$$\min_{x_i \in X} |\langle \omega, x_i \rangle + \omega_0| = 1 \quad (*)$$

- можно добиться масштабированием параметров

Далее будем считать, что $\omega_0 = 1$ Верно

Рассмотрим отклонение x от прямой с вектором нормали ω :

$$p(a, x) = \frac{|\langle \omega, x \rangle + \omega_0|}{\|\omega\|}$$

$$\min_{x_i \in X} p(a, x_i) = \min_{x_i \in X} \frac{|\langle \omega, x_i \rangle + \omega_0|}{\|\omega\|} =$$

$$= \frac{1}{\|\omega\|} \cdot \underbrace{\min_{x_i \in X} |\langle \omega, x_i \rangle + \omega_0|}_{1} = \frac{1}{\|\omega\|} \rightarrow \max_{\omega}$$

$$\left\{ \begin{array}{l} \frac{1}{\|\omega\|} \rightarrow \max_{\omega, \omega_0} \\ \text{или} \end{array} \right.$$

$$y_i \cdot (\langle \omega, x_i \rangle + \omega_0) \geq 0 \quad i=1 \dots l$$

$$\min_{x_i \in X} |\langle \omega, x_i \rangle + \omega_0| = 1$$

- нунуны

$$\begin{cases} \min_{\omega, \omega_0} \|\omega\|^2 \\ \text{such that } y_i \cdot (\langle \omega, x_i \rangle + \omega_0) \geq 1, \quad i=1..l \end{cases}$$

если $\min_{\omega, \omega_0} \|\omega\|^2 > 1$, то
 $\|\omega\|$ можно уменьшить

Задача SVM

2.2 Линейно неразделимый случай

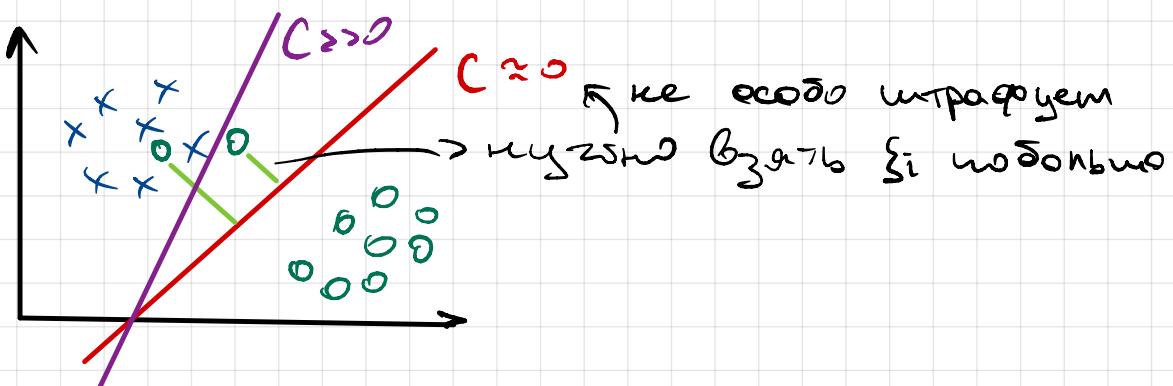
$$\begin{cases} \min_{\omega, \omega_0} \sum_{i=1}^l \xi_i \\ \text{such that } y_i (\langle \omega, x_i \rangle + \omega_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{cases}$$

$\frac{1}{2} \|\omega\|^2 + C \cdot \sum_{i=1}^l \xi_i \rightarrow \min_{\omega, \omega_0, \xi_i}$

нормализация, чтобы избежать переполнения

ограничение

SVM при обучении



Берём $\sum \xi_i$, а не $\sum \xi_i^2$, потому что хотим, чтобы ошибка была разрезаемой (т.е. чтобы можно было дальше ее разбить на две части).

Попробуем выразить ξ_i :

$$\begin{cases} \xi_i \geq 1 - y_i (\langle \omega, x_i \rangle + \omega_0) \\ \xi_i \geq 0 \end{cases} \Rightarrow \xi_i = \max(0, 1 - y_i (\langle \omega, x_i \rangle + \omega_0))$$

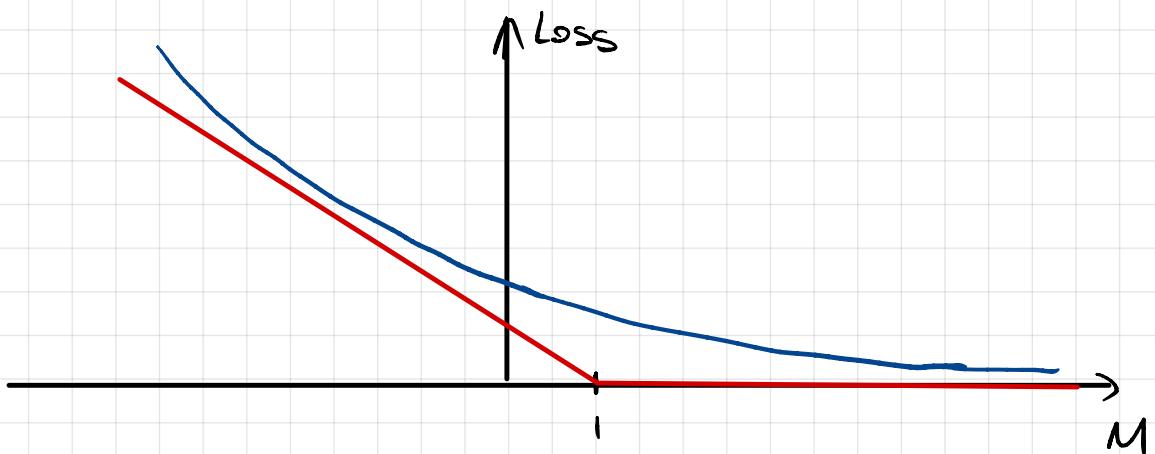
$$\frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l \max(0, 1 - y_i (\langle \omega, x_i \rangle + \omega_0)) \rightarrow \min_{\omega, \omega_0}$$

регуляризатор

$$L(y, z) = \max(0, 1 - yz)$$

Функционал ошибки

2.3 Оценка LR и SVM



при $M < 0$ поведение ~ ожидаемое

при $M > 0$: SVM достоверно, требуя $M \geq 1$
LR ходит, требуя $M \rightarrow +\infty$

SVM приходит к наилучшему значению метрик качества классификации

LR - стремится корректно оценивать вероятности

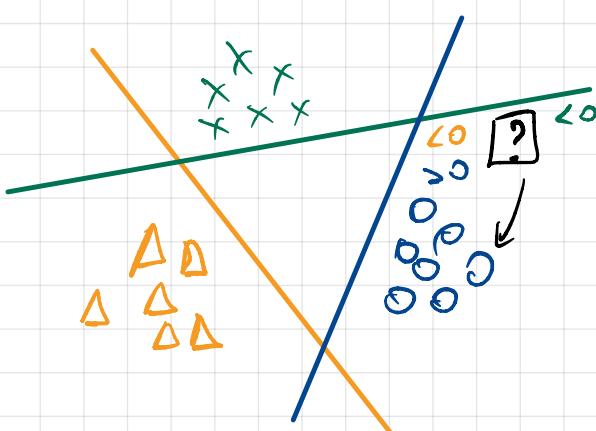
В твоё здешнее условие на как можно лучше классифицировать

Многоклассовая классификация

$$\mathbb{Y} = \{-1, +1\} \rightarrow \mathbb{Y} = \{1, \dots, K\}$$

Сведение к серии бинарных задач

II. One-vs-all



$$b_k(x) = \langle \omega_k, x \rangle + \omega_{0k}$$

↑
однозначно на $X_k = \{x_i | y_i = k\}_{i=1}^l$

$$q(x) = \operatorname{argmax}_{k=1, \dots, K} b_k(x)$$

+ легко генерируется

- много классов \Rightarrow много параметров: $K(d+1)$

$\Rightarrow b_k$ не знатят друг о друге при обучении

$$|b_1(x)| \propto 10^6$$

$$|b_2(x)| \approx 10^1 \rightarrow \text{недависимые}$$

1.2 all - vs - all

C^2_K классификаторов

$$q_{ij}(x) - \text{однозначно на } X_{ij} = \{(x_n, y_n) | y_n \in \{i, j\}\}$$

$$q(x) = \operatorname{argmax}_{k=1, \dots, K} \sum_{j \neq k} [a_{kj}(x) = k]$$

предполагаем независимые классы и смотрим, сколько раз он подходит

+ нет проблем с калибровкой весов
- $\sim K^2$ классификаторов

2 Прямые подходы

Много классовых логист. регрессия

$$b_k(x) = \langle \omega_k, x \rangle + \omega_{0k}$$

$$x \rightarrow b(x) = (b_1(x), \dots, b_K(x))$$

$$\text{Softmax } (z_1, \dots, z_K) = \left(\frac{e^{z_1}}{\sum e^{z_k}}, \dots, \frac{e^{z_K}}{\sum e^{z_k}} \right)$$

$$P(y=k|x) = \frac{\exp(\langle \omega_k, x \rangle + \omega_{0k})}{\sum_{j=1}^K \exp(\langle \omega_j, x \rangle + \omega_{0j})}$$

$$\sum_{i=1}^l \log P(y=y_i|x_i) \rightarrow \max_{\omega_k, \omega_{0k}}$$

③ Метрики качества в многоклассовой класс.

$$\text{accuracy} = \frac{1}{n} \sum_{i=1}^n [\alpha(x_i) = y_i]$$

k -ий класс: TP_k, FP_k, FN_k, TN_k
(где отнесение k -го класса от основных)

Макроусреднение

$$1) \text{precision}_k = \frac{TP_k}{TP_k + FP_k}$$

$$2) \text{precision} = \frac{1}{k} \sum_{k=1}^K \text{precision}_k$$



Все классы имеют равный вклад

Микроусреднение

$$1) \overline{TP} = \frac{1}{K} \sum_{k=1}^K TP_k$$

$$2) \text{precision} = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}$$



Большее влияние крупных классов

Классификация с пересек. классами (multi-label classification)

$$\mathbb{Y} = \{0, 1\}^K$$

① binary relevance

$$b_k(x) - \text{указм на } (x_i, y_{ik})_{i=1}^n$$

Решающие деревья

Линейные модели:

- + линейные
- + линейные
- линейные \rightarrow решаются в DL

но не все задачи такие:

$$\text{сортировка массива } (x_1, \dots, x_n) \xrightarrow{a} (\tilde{x}_1, \dots, \tilde{x}_n)$$

Будем призывать линейн. модели

Как оценить студента на 100%?



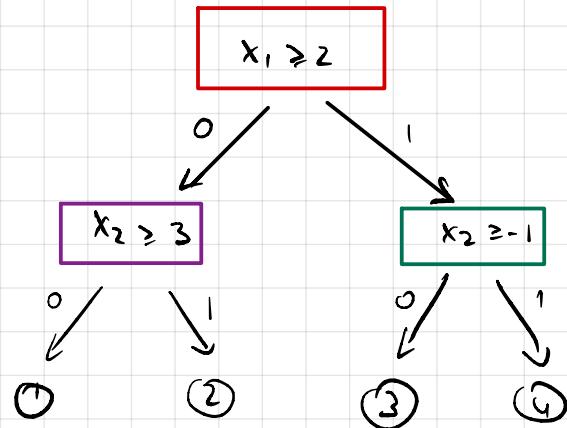
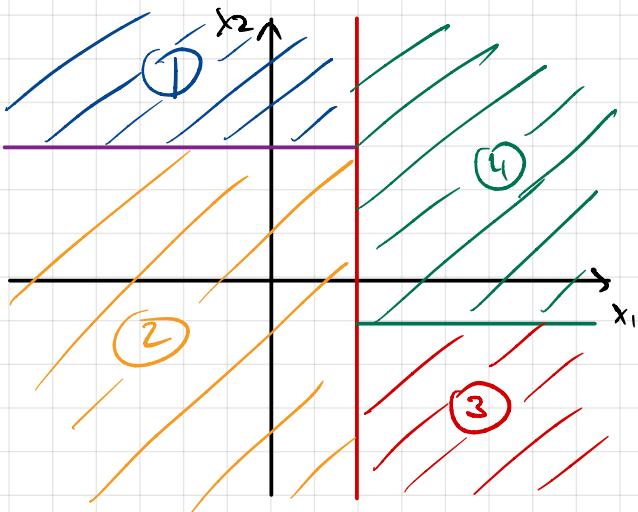
① Структура модели

Решающее дерево - бинарное дерево, где:

- 1) \mathcal{D} - вектор. Вершина $\Rightarrow \beta_{\mathcal{D}} : \mathbb{X} \rightarrow \{0, 1\}$
предикат
- 2) \mathcal{D} - лист $\Rightarrow C_{\mathcal{D}} \in \mathbb{Y}$ - прогноз

Обычно multi-way дерево, то есть ограничено и есть риски переобуч.

- Предикаты:
- 1) $\beta_{\mathcal{D}}(x) = [x_j \geq t]$, j, t — самая норм
 - 2) $\beta_{\mathcal{D}}(x) = [\langle \omega, x \rangle \geq t]$ $\omega \in \mathbb{R}^d$, t
 - 3) $\beta_{\mathcal{D}}(x) = [g(x, x_0) \geq t]$



Какие обычные прогнозы? ① регр. $\Rightarrow C_{\mathcal{D}} \in \mathbb{R}$

② класс. $\Rightarrow C_{\mathcal{D}} \in \{1, \dots, K\}$

$$\cdot C_{\mathcal{D}} \in \mathbb{R}^K : C_{k_{\mathcal{D}}} \geq 0 \\ \sum C_{k_{\mathcal{D}}} = 1$$

② Общее

ЧБ.: если $B \times X$ нет $x_i = x_j$, $y_i \neq y_j$, то \exists дерево с культивой ошибкой не обуз. ветвление

Решения: если из всех корректных на X деревьев ветвь самое мал. (по глубине / по числу вершин/...), то оно не будет перебором.



не перебор, т.к. NP-hard



важник

Вершина

объекты из X , которые попали в неё

Split Node (m, R_m):

if Критерий останова:

$C_m = \dots \rightarrow$ средний ответ R_m
 самый частый класс
 один классов в R_m

return C_m

$j, + = \operatorname{argmax}_{j=1, \dots, d} Q(R_m, j, +)$

Split Node (l, X)

- фундаментал. критерий качества предиката
 (критерий информативности пред.)

полный перебор

$$R_l = \{(x, y) \in R_m \mid [x_j < +] = 1\}$$

$$R_r = \{(x, y) \in R_m \mid [x_j \geq +] = 1\}$$

Split Node (l, R_l)

Split Node (r, R_r)

Критерий останова:

- макс. глубина
- мин. число объектов в листе
- одна ошибка в листе
- ...

③ Критерии качества предиката

$H(R)$ - impurity (хаотичность) - показывает разнородность объектов (семинар)

Примеры: ретр. $\Rightarrow H(R_m) = \frac{1}{|R_m|} \sum_{(x_i, y_i) \in R_m} (y_i - \bar{y}_m)^2$ - дисперсия

$$\text{класс.} \Rightarrow p_k = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k]$$

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

макс. при $H[a, b]$
 0 при 1 шаг.

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

Однако когда K построено H(R):

$$H(R) = \min_{C \in \mathcal{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c)$$

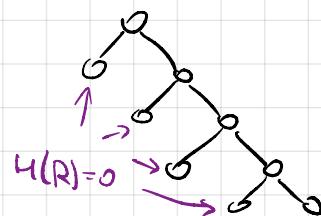
Ошибка линейного классификации

Если константой можно добиться низкой ошибки, то impurity низкая

$$Q(R_m, j, +) = H(R_m) - \frac{|R_{el}|}{|R_m|} H(R_e) - \frac{|R_{vl}|}{|R_m|} H(R_v) \rightarrow \max$$

самое интересное, если $|R_{vl}| \gg |R_{el}|$

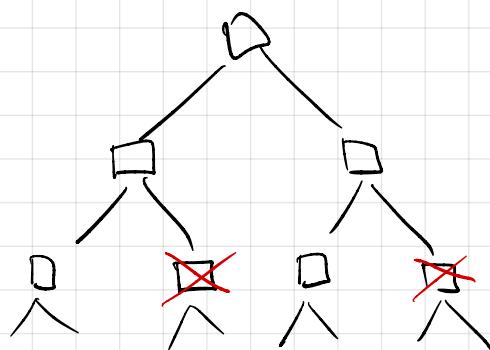
здесь не получится:



$$+ H(R_e) + H(R_v) может оказаться > H(R_m)$$

где критерий: $\frac{|R_{el}|}{|R_m|} H(R_e) + \frac{|R_{vl}|}{|R_m|} H(R_v) \rightarrow \min_{j,+}$

④ Глубина деревьев (pruning)



Но техника старая

Лекция 10. 15. 11.2023

⑤ Обработка пропусков

Вариант 1

Объяснение: б) Рассмотрим признак j есть пропуск как подсчитать $Q(R, j, t)$?

Рассмотрим $V_j(R)$ - объекты, у которых x_j неизвестно

$$Q(R, j, t) = \frac{|R \setminus V_j(R)|}{|R|} Q(R \setminus V_j(R), j, t)$$

$H(R_e), H(R_u)$ ищутся методами

Если в выражении предикат с j -м нр., то отбрасываем объекты из $V_j(R)$ влево, вправо с весами $\frac{|R_e|}{|R|}$ и $\frac{|R_u|}{|R|}$

Тест: $Q_{mk}(x)$ - прогноз вер. класса k в верн. m

$$Q_{mk}(x) = \begin{cases} \alpha_k(x), & \beta_m(x) = 0 \\ \alpha_k(x), & \beta_m(x) = 1 \\ \frac{|R_e|}{|R_m|} \alpha_k(x) + \frac{|R_u|}{|R_m|} \alpha_k(x), & \beta_m(x) - близкий к нулю \\ \text{смк}, & m - настоящая верн. \end{cases}$$

Вариант 2 - Суррогатные предикаты

Выбираем лужиний предикат без учёта пропусков $\beta(x)$

Ищем несколько суррогатных предикатов (которые дают идентичные разбиения) $\beta'(x), \beta''(x)$

$$x \rightarrow \beta(x) \xrightarrow{\text{не нулев.}} \beta'(x) \xrightarrow{\text{не нулев.}} \beta''(x)$$

⑥ Работа с категориальными признаками

x_j - катер., $x_j \in Q = \{u_1, \dots, u_g\}$

Привильный вариант: 1) $[x_j = q] \Leftrightarrow \text{ONE}$

2) multi-way splits $u_1 / \dots / u_g$

вариант ненужный: $Q = Q_1 \cup Q_2$ (2^g разбиений)

$$\beta(x) = [x_j \in Q_i]$$

$$Rm(u) = \{(x, y) \in Rm \mid x_j = u\}$$

$$Nm(u) = |Rm(u)|$$

$$\Psi = \{-1, +1\}$$

sout:

$$\frac{1}{N_m(u_{(1)})} \sum_{(x_i, y_i) \in \text{Rm}(u_{(1)})} [y_i = +1] \leq \dots \leq \frac{1}{N_m(u_{(q)})} \sum_{(x_i, y_i) \in \text{Rm}(u_{(q)})} [y_i = +1]$$

$$u_{(1)}, \underbrace{u_{(2)}}, \dots, u_{(q)}$$

\rightarrow среди этих разбиений $Q = Q_1 \cup Q_2$ обязательно будет оптимальное

⑦ Методы построения деревьев

ID3: энтр. крит., кр. останова

C4.5: Gain Ratio, стратегия, пропуски

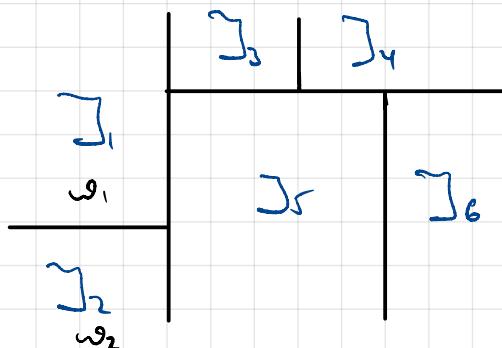
CART: ...

⑧ Дерево с лин. модели

$$X = J_1 \cup \dots \cup J_n$$

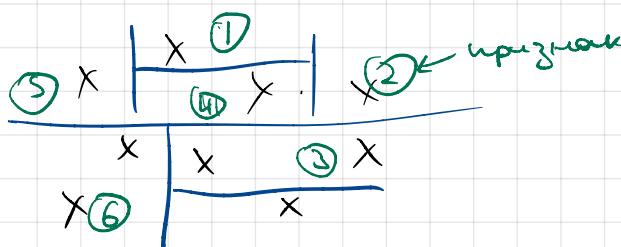
w_1, \dots, w_n - признаки

$$a(x) = \sum_{j=1}^n \underbrace{w_j}_{\text{признак}} [\underline{x \in J_j}]$$

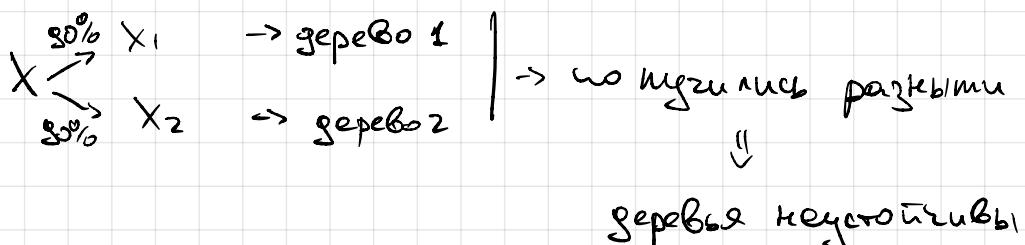


T.e. дерево как бы находит новые лекальные признаки, а затем строит лин. модель

Идея: можно построить дерево со служебными предикатами, и использовать номера листьев как новые признаки



Композиции могут (ensembles)



Но что же это если усреднить прогнозы неустойчивых деревьев?

Все будет очень сплошно!

$$X = (x_i, y_i)_{i=1}^L, y \in \mathbb{R}$$

Будет так: Генерации подвыборки из X выбором L объектов с возвращением

$$\{1, 2, 3, 4\} \rightarrow \{1, 2, 2, 4\} \sim \frac{2}{3} \text{ уникальных}$$

$X \rightarrow x_1 \dots x_N$ — подвыборки
 $\downarrow \quad \downarrow$
 $b_1(x) \quad b_N(x)$

$y(x)$ — истинный ответ

$p(x)$ — идентичность на x

$$\varepsilon_j(x) = b_j(x) - y(x)$$

$$E_x[(b_j(x) - y(x))^2] = E_x[\varepsilon_j^2(x)] = E_1$$

Предполож.: $E \varepsilon_j(x) = 0$ — несмещённость

$E \varepsilon_i(x) \cdot \varepsilon_j(x) = 0$, $i \neq j$ — некорр. ошибок

то будем говорить неправда

$$a(x) = \frac{1}{N} \sum_{j=1}^N b_j(x)$$

$$\begin{aligned}
 E\left[\left(\frac{1}{N} \sum_{j=1}^N b_j(x) - y(x)\right)^2\right] &= E\left[\left(\frac{1}{N} \sum b_j(x) - \frac{1}{N} \sum y(x)\right)^2\right] = E\left[\left(\frac{1}{N} \sum_{j=1}^N \varepsilon_j(x)\right)^2\right] = \\
 &= \frac{1}{N^2} E\left[\sum_{j=1}^N \varepsilon_j^2(x) + \underbrace{\sum_{i \neq j} \varepsilon_i(x) \varepsilon_j(x)}_0\right] = \frac{1}{N^2} \sum_{j=1}^N E[\varepsilon_j^2(x)] = \frac{1}{N^2} \cdot N \cdot E_1 = \frac{1}{N} E_1
 \end{aligned}$$

Как перенести в задачи?

Bias-Variance Decomposition (BVD)

$$X = (x_i, y_i)_{i=1}^l, y \in \mathbb{R}$$

$p(x, y)$ из $\mathbb{X} \times \mathbb{Y}$

$$L(y, a) = (y - a)^2$$

Среднеквадратичный риск:

$$R(a) = E_{x,y} (y - a(x))^2 = \iint_{\mathbb{X} \times \mathbb{Y}} (y - a(x))^2 \cdot p(x, y) dx dy$$

$$\text{Можно показать: } a_s(x) = E[y|x] = \int_{\mathbb{Y}} y \cdot p(y|x) dy$$

Метод обуления:

$$\mu: \underbrace{(\mathbb{X} \times \mathbb{Y})^l}_{\substack{\text{нр-во} \\ \text{обул. выборок}}} \rightarrow \underbrace{\mathcal{A}}_{\substack{\text{семейство} \\ \text{моделей}}} \\ \text{как бы на всем тестовом объектам}$$

$$L(\mu) = E_x E_{x,y} (y - \underbrace{\mu(X)}_{\substack{\text{модель} \\ \text{прогноз}}})^2$$

- основка метода обуления
средний среднеквадратичный риск

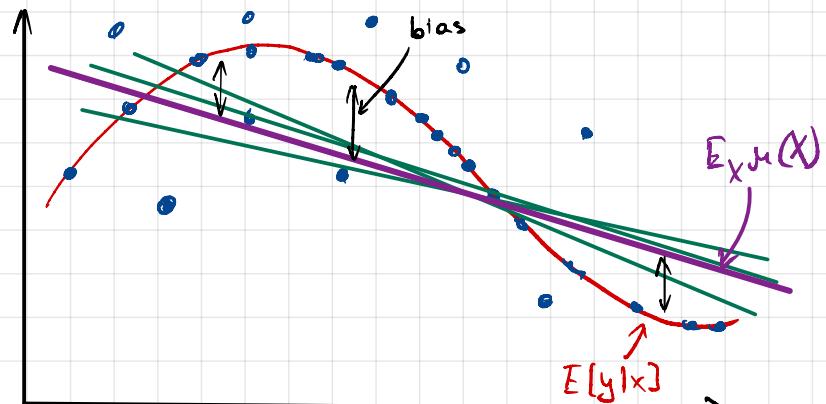
$$L(\mu) = E_{x,y} [(y - \underbrace{E[y|x]}_{\substack{\text{лучшая} \\ \text{модель}}})^2] \leftarrow \text{шум (noise) - ошибка лучшей модели, текущая оценка на ошибку модели}$$

$$+ E_x \left[(E_x [\mu(X)] - \underbrace{E[y|x]}_{\substack{\text{средняя модель} \\ \text{на всем } X}})^2 \right] \leftarrow \text{смещение (bias) - отклонение средней модели от лучшей "находящейся" семейства моделей из-за языка}$$

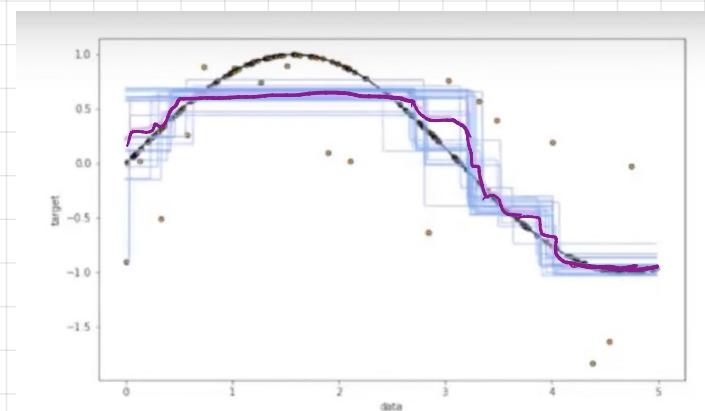
$$+ E_x [E_x [(\mu(X) - E_x[\mu(X)])^2]] \leftarrow \text{разброс (variance) - показатель устойчивости метода обуления (характер переобученности)}$$

Лекция II, 24.11.2023

интуиция: bias \uparrow
модель variance \downarrow

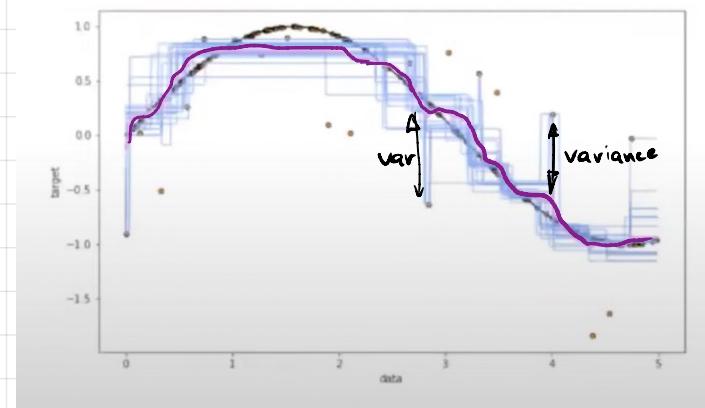


геометрия:



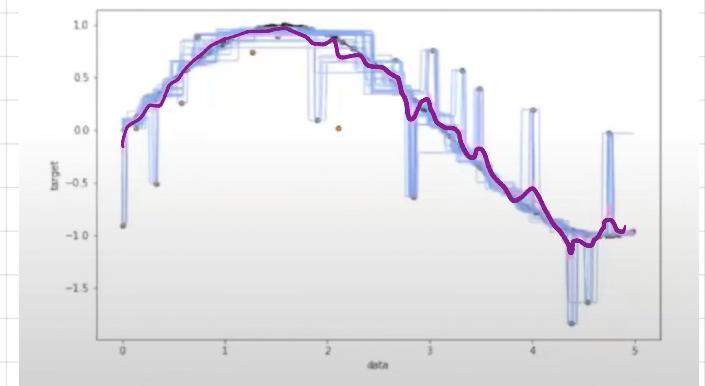
Графика 2

bias ↑
var ↓



Графика 5

bias ↓
var ↑

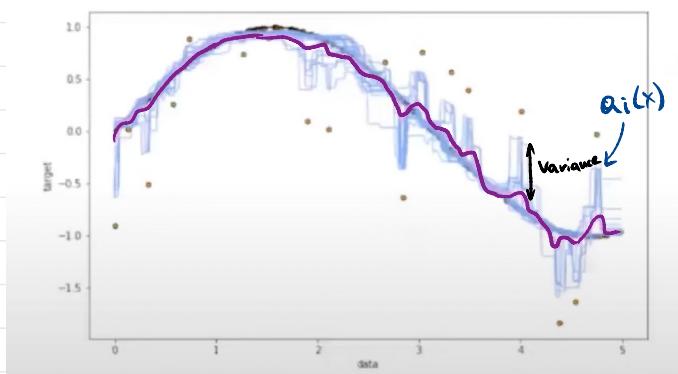


Графика 8

bias ≈ 0
var ≈ 0

$b_1(x), \dots, b_N(x)$ - геометрия гн. 8

$$a(x) = \frac{1}{N} \sum_{n=1}^N b_n(x)$$



curv. kp. - огнико $a(x)$

grav. kp. - средняя комбинация

bias ≈ 0

variance ↓

Бэггинг (bagging)

$\mu(x)$ - метод обучения

\tilde{x} - слу. подборка (ген. бутстр.ом)

$$\left. \begin{array}{l} b_1(x) = \mu(\tilde{x}_1) \\ \vdots \\ b_N(x) = \mu(\tilde{x}_N) \end{array} \right\} \text{- базовые модели}$$

$$a_N(x) = \frac{1}{N} \sum_{n=1}^N b_n(x) = \frac{1}{N} \sum_{n=1}^N \mu(\tilde{x}_n) \quad \text{- композиция (бэггинг)}$$

bootstrap aggregating

метод обучения

Можно показать: 1) $\text{bias}(a_N) = \text{bias}(b_n) \Rightarrow$ если базовые модели слабые, то a_N тоже слабая
 $\Rightarrow b_n(x)$ неизвестные

$$\begin{aligned} 2) \text{var}(a_N) &= \frac{1}{N} \text{var}(b_n) + \frac{N(N-1)}{N^2} \cdot \underline{\text{cov}(b_n(x), b_m(x))}, \\ &= E_{x,y} E_x [(\mu(\tilde{x}_n) - E \mu(\tilde{x}_n)) (\mu(\tilde{x}_m) - E \mu(\tilde{x}_m))] \end{aligned}$$

\Rightarrow тем менее коррелированные базовые модели, тем лучше

Random Forest

1) Глубокие: $\text{min_samples_leaf} = 3$

2) Бутстр. подборок

3) в каждой верн. б $[x_j \geq +]$ из выбираем из слу. подбр. признаков
 размера k

~~4) сбои моделей признаков для каждого дерева~~

$\Rightarrow b_i(x)$ не нулевое важнейшее признак

$\Rightarrow \text{bias}(b_3) \uparrow \Rightarrow \text{bias}(a_N) \uparrow$

классиф.: $\underset{y \in \mathbb{Y}}{\operatorname{argmax}} \sum_{n=1}^N [b_n(x) = y]$

регр.: $\frac{1}{N} \sum_{n=1}^N b_n(x)$

$$\begin{aligned} \text{перп.} : k &= \sqrt{d} \\ \text{класс.} : k &= \lfloor \frac{d}{3} \rfloor \end{aligned}$$

• RF - самый универсальный метод в ML

Гиперпараметр: N , ... и т.д.



- Подсказки:
- 1) ошибка долго обнуждается
 - 2) если $\text{bias}(b_n) \gg 0$, то RF будет плохим

RF: out-of-bag

$b_n(x)$ обнуж. на $X_n \subset X$

$$OOB = \frac{1}{C} \sum_{i=1}^C L(y_i, \underbrace{\frac{1}{\sum_{n \neq i} b_n(x_n)} \sum_{i=n}^N [x_i \notin X_n] b_n(x_i)}_{\text{прогноз для } i\text{-го объекта}}) \approx LOO$$

(CV с $k=C$)

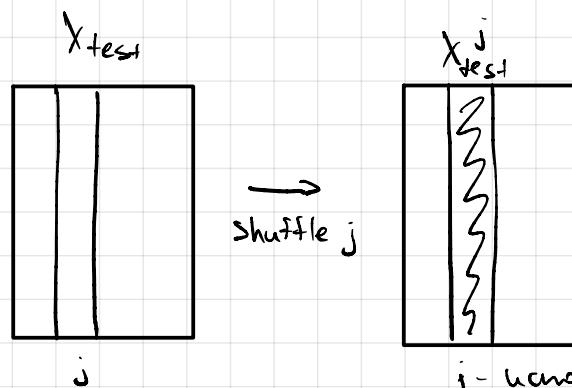
но дерево есть, которое на i -м не обнужилось

RF: важности признаков (перекрестковые, не только для RF)

$a(x)$

$$Q(a, X_{test}) = Q_{test}$$

$$Q(a, X_{test}^j) = Q_{test}^j$$



j - непрерывный

Важность признака: $q_j = Q_{test}^j - Q_{test}$

$q_j \approx 0 \Rightarrow$ неважен

$q_j > 0 \Rightarrow$ важен

$q_j < 0 \Rightarrow$ $\frac{|q_j|}{Q_{test}} \times 100$

Градиентный бустинг

(gradient boosting machine, GBM)

каждая следующая модель корректирует ошибки предыдущих

① Бустинг для MSE

$$\frac{1}{C} \sum_{i=1}^N (a(x_i) - y_i)^2 \rightarrow \min_a$$

$$a_N(x) = \sum_{n=1}^N \delta_n b_n(x), \quad \delta_n \in \mathbb{R}$$

$$b_1(x) : \frac{1}{C} \sum_{i=1}^N (b_1(x_i) - y_i)^2 \rightarrow \min_{b_1(x)} \quad | \text{ фиксируем } \delta_1 = 1$$

$$b_2(x) : b_1(x_i) + b_2(x_i) = y_i \rightarrow b_2(x_i) = y_i - b_1(x_i) = s_i \in \mathbb{R}$$

$$\frac{1}{C} \sum_{i=1}^N (b_2(x_i) - s_i)^2 \rightarrow \min_{b_2}$$

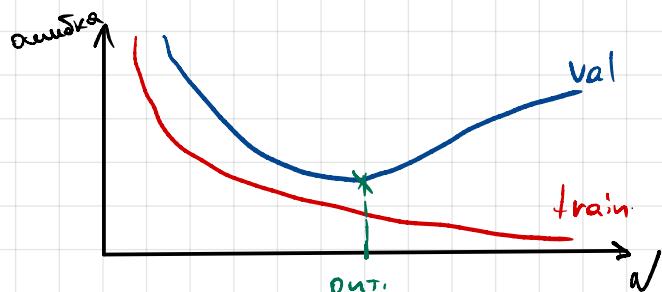
$$\delta_2 = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{C} \sum_{i=1}^N (b_1(x_i) + \delta b_2(x_i) - y_i)^2$$

$$b_3(x) : s_i = y_i - b_1(x_i) - \delta_2 b_2(x_i)$$

$$\frac{1}{C} \sum (b_3(x_i) - s_i)^2 \rightarrow \min$$

Когда останется вибрация?

- ГД неподвижна
- сдвиг на val



② Общий случай

$L(y, z)$ - гип. фн. н.

$$a_N(x) = \sum_{n=1}^N b_n(x)$$

нотом сделаем лучше

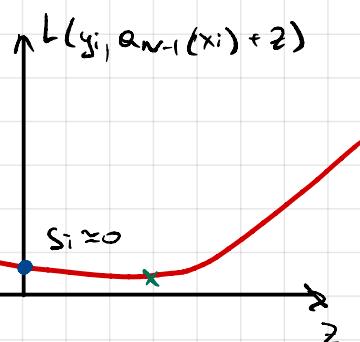
Как обуздать? $b_N(x) : \frac{1}{C} \sum_{i=1}^N L(y_i - a_{N-1}(x_i), b_N(x_i))$

ноч - не зас - перес

$b_1(x)$ - оцениваем как умеем

$b_N(x) : b_1(x), \dots, b_{N-1}(x)$ - уже есть

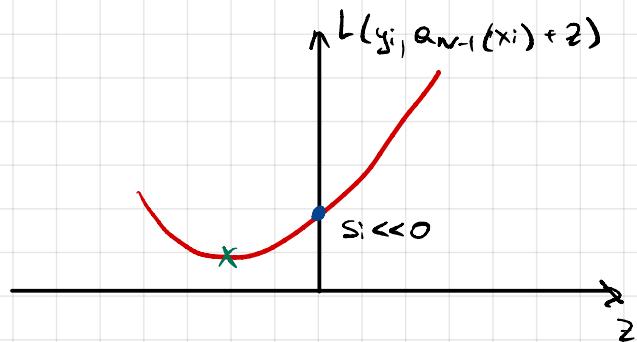
$$\frac{1}{C} \sum_{i=1}^C L(y_i, a_{N-1}(x_i) + b_N(x_i)) \rightarrow \min_{b_N}$$



$$s_i = -\frac{\partial}{\partial z} L(y_i, z) \Big|_{z=a_{N-1}(x_i)}$$

$$\frac{1}{C} \sum_{i=1}^C (b_N(x_i) - s_i)^2 \rightarrow \min_{b_N}$$

$$a_N(x) = a_{N-1}(x) + b_N(x)$$



- сдвиг

градиентный
метод

$$Q(z_1, \dots, z_C) = \frac{1}{C} \sum_{i=1}^C L(y_i, a_{N-1}(x_i) + z_i)$$

$$(s_1, \dots, s_C) = \nabla_z Q(z)$$

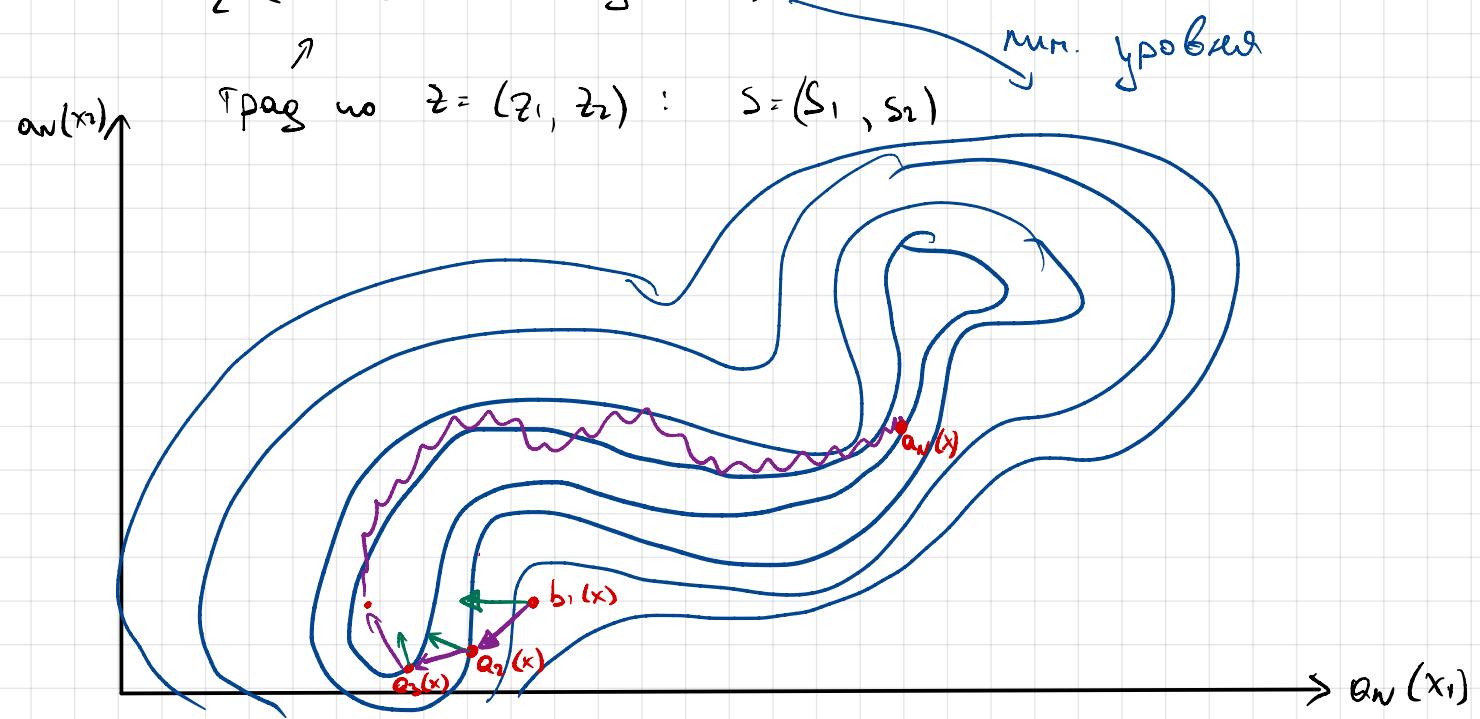
Лекция 12 . 01.12.2023

Пример: $X = ((x_1, y_1), (x_2, y_2))$

$$\frac{1}{2} (L(y_1, z_1) + L(y_2, z_2)) \rightarrow \min$$

таким же $z = (z_1, z_2) : s = (s_1, s_2)$

мин. попытка



Проф. бустинг - граф. смысл в кр-ве прогнозов композиции на train

Маг делает не по градиенту, а по его аппроксимации

Много загадок:

1) Почему нельзя просто считать $a_n(x_i) = a_{n-1}(x_i) + s_i$

- не имеет прогноза на $x \notin X$

2) Почему бv обуз. на MSE?

③ Регуляризация

Можно показать, что в бустинге (экспериментально):

	bagging	boosting
bias	-	↓
variance	↓	↑

⇒ базовые модели должны быть простыми - неглубокие деревья

Проблемы: 1) если базовые модели простые, то качество так себе
⇒ они могут испортить композицию
2) если базовые модели сложные, то они будут перебуждаться

⇒ бv (x) нельзя доверять

Сокращение шага: $a_n(x) = a_{n-1}(x) + \sum b_i \epsilon(0, 1)$

$$\sum b_i \in (0, 1]$$

$$\sum b_i \Rightarrow \text{онт. } N$$

Стохастический градиентный бустинг:

модели простые ⇒ могут вобрать мало информации о данных

чтд: обуз. $b_N(x)$ на $X_N \subset X$
↑ подвыборка

④ Руками модели

4.1 Регрессия

$$L(y, z) = \frac{1}{2} (y - z)^2$$

$$s_i = - \left. \frac{\partial L}{\partial z} (y_i, z) \right|_{z=a_{N-1}(x_i)} = y_i - a_{N-1}(x_i)$$

$$L(y, z) = |y - z|$$

$$s_i = -\text{sign}(a_{N-1}(x_i) - y_i)$$

4.2 Классификация

$a_N(x) \in \mathbb{R}$ - уверенность в принадлежности к классу

$\text{sign } a_N(x)$ - прогноз

$$L(y, z) = \log(1 + \exp(-yz))$$

$$\frac{1}{C} \sum_{i=1}^n \left(b_N(x_i) - \underbrace{\frac{y_i}{1 + \exp(y_i a_{N-1}(x_i))}}_s \right)^2 \rightarrow \min_{b_N}$$

$$s_i = y_i \cdot y_i$$

$y_i \cdot a_{N-1}(x_i) \gg 0 \Rightarrow w_i \approx 0 \Rightarrow$ не трогаем такие
 $s_i \approx 0$ обзоры

$y_i \cdot a_{N-1}(x_i) \ll 0 \Rightarrow w_i \approx 1 \Rightarrow$ стараемся корректировать
 (всюду) $s_i \approx y_i$ в сторону правильного обзора

$y_i \cdot a_{N-1}(x_i) \approx 0 \Rightarrow w_i = \frac{1}{2} \Rightarrow$ корректируем умеренно
 $s_i = \frac{1}{2} y_i$

Логист. фн. переводит частоты в вероятности
 (против e^{-yz})

5 Прогностический бюджет на деревьями

$$b_N(x) \geq \sum_{j=1}^{J_N} b_{Nj} [x \in R_{Nj}]$$

\uparrow \uparrow
 прогноз б. обнаруж. X ,
 j-м месте соотв. j-му месту

$$Q_N(x) = a_{N-1}(x) + \sum_{j=1}^{J_N} b_{Nj} [x \in R_{Nj}]$$

Идея: подберём b_{Nj} , чтобы были одн. с 7.2. $L(y, z)$

$$\sum_{i=1}^n L(y_i, \alpha_{n-1}(x_i) + \sum_{j=1}^{J_n} \delta_j [x_i \in R_{nj}]) \rightarrow \min_{\delta_1, \dots, \delta_{J_n}}$$

$$= \sum_{j=1}^{J_n} \sum_{(x_i, y_i) \in R_{nj}} L(y_i, \alpha_{n-1}(x_i) + \delta_j) \rightarrow \min_{\delta_1, \dots, \delta_{J_n}}$$

Задача сводится к нахождению J_n и δ_j :

$$\sum_{(x_i, y_i) \in R_{nj}} L(y_i, \alpha_{n-1}(x_i) + \delta_j) \rightarrow \min_{\delta_j \in \mathbb{R}} \rightarrow \text{степеней вида}$$

Например: $L(y, z) = \log(1 + \exp(-yz))$

$$\sum_{(x_i, y_i) \in R_{nj}} \log(1 + \exp(-y_i(\alpha_{n-1}(x_i) + \delta_j))) \rightarrow \min_{\delta_j}$$

Сгенерируем линию метода Ньютона - Радемахера для $\delta = 0$:

$$\delta_j = -\frac{\sum_{(x_i, y_i) \in R_{nj}} s_i}{\sum_{(x_i, y_i) \in R_{nj}} |s_i| \cdot (1 - |s_i|)}$$

Линия линия не отмечается

⑥ Равнение

6.1 Взвешивание

6 класс. задача Берга: $L(y, z) = \tilde{L}(yz)$

$$\text{тогда } s_i = y_i \left(-\frac{\partial \tilde{L}}{\partial u} (u) \Big|_{u=y_i \alpha_{n-1}(x_i)} \right)$$

w_i

6.2 Устойчивость к беспорядкам

$$L(y, z) = e^{-yz}$$

$$s_i = y_i \underbrace{\exp(-y_i \alpha_{n-1}(x_i))}_{w_i}$$

$$y_i \alpha_{n-1}(x_i) \ll 0 \Rightarrow w_i \rightarrow +\infty$$