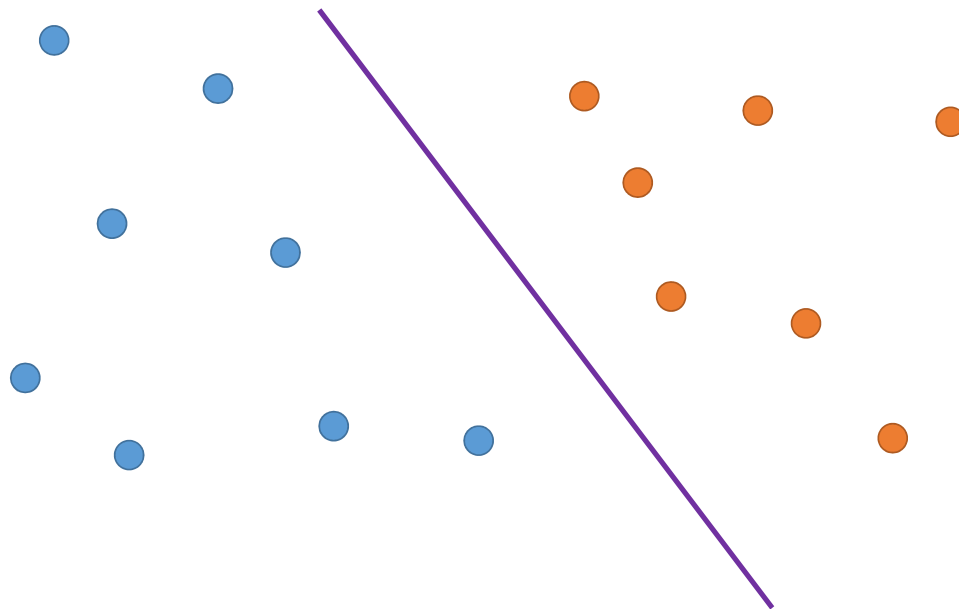


Многоклассовая классификация

# Бинарная классификация



- Бинарная

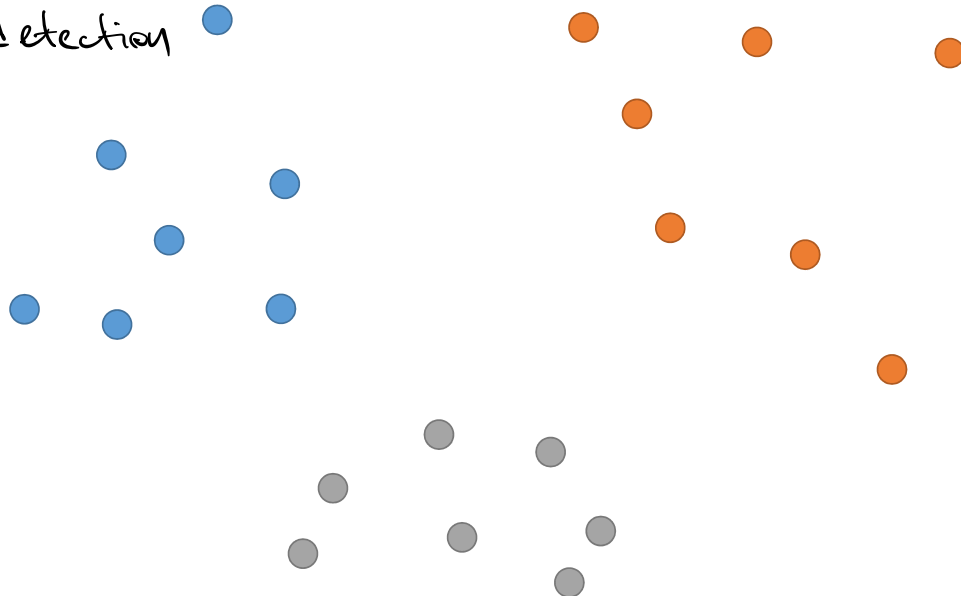
- Многоклассовая классификация

- Одноклассовая классификация

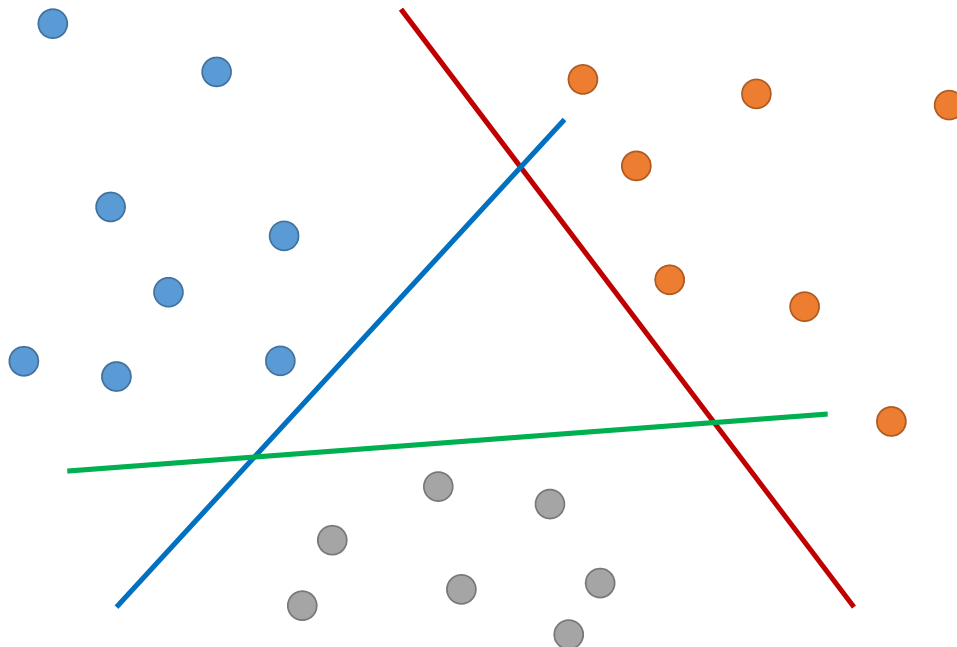
~ Anomaly detection

(1.) попробовать  
обобщить алгоритм  
=> для  $K$  классов

(2.)  $K$  классов  
=> несколько  
ex - классов  
заданы

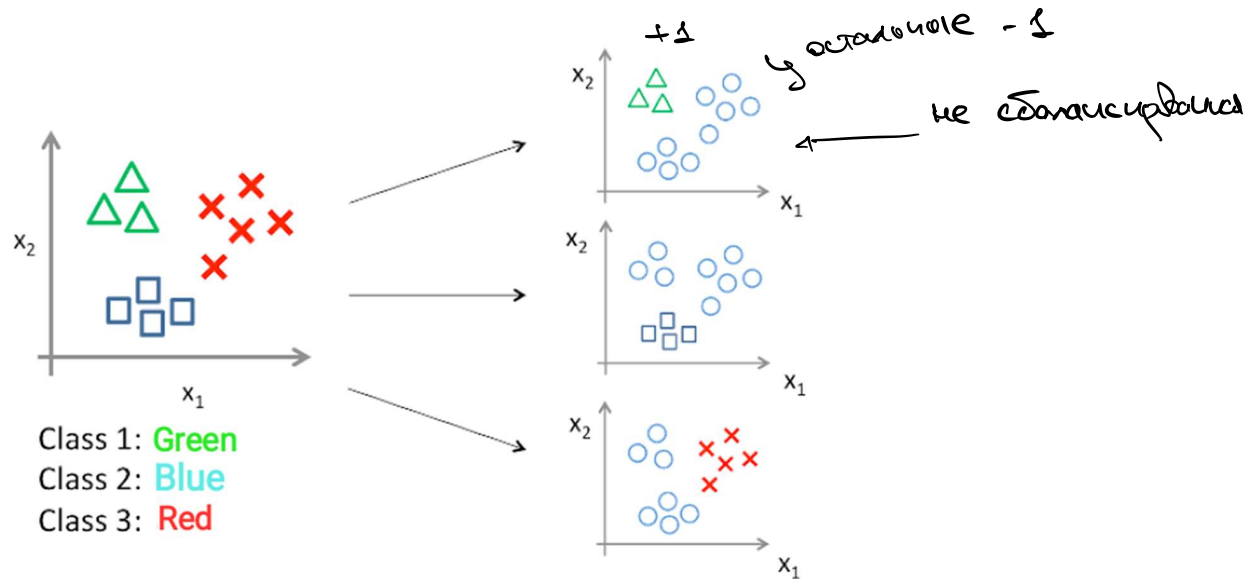


# Многоклассовая классификация



$K$  классов  $\Rightarrow K$  классификаторов

## One-vs-All (One-vs-Rest)



$$(y_i, x_i) \mapsto a(x_i) \in \mathbb{R}$$

$$a(x_i) \mapsto \log \text{Reg.}$$

①. Нужно смотреть  
поверх каждого

②. Калибровка  
(prob. calibration  
sketch)

# One-vs-All

$$a_1 \in \tilde{\mathcal{P}}_V(x \in 1)$$

$$a_2 \in \tilde{\mathcal{P}}_V(x \in 2)$$

$$a_3 \in \tilde{\mathcal{P}}_V(x \in 3)$$

• K классов:  $\mathcal{Y} = \{1, \dots, K\} \in \tilde{\mathcal{P}}_V(x \in 3)$

•  $X_k = (x_i, [y_i = k])_{i=1}^{\ell}$

• Обучаем  $a_k(x)$  на  $X_k, k = 1, \dots, K$

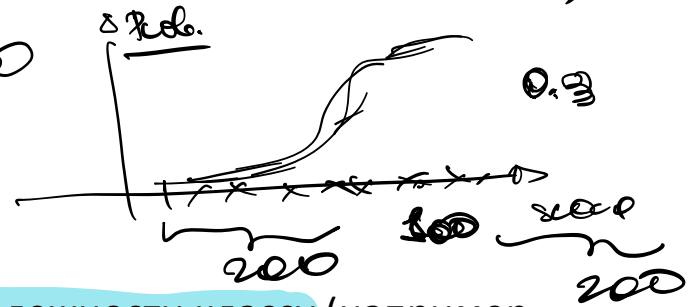
•  $a_k(x)$  должен выдавать оценки принадлежности классу (например,  $\langle w, x \rangle$  или  $\sigma(\langle w, x \rangle)$ )

• Итоговая модель:

• логическая  
регрессия  
 $\in \mathcal{P}$

$$a(x) = \arg \max_{k=1, \dots, K} a_k(x)$$

$$\left\lfloor \frac{\langle w, x \rangle}{\|w\|} \right\rfloor \begin{matrix} (2, 1) \\ (10, 5) \end{matrix} = (0, 1)$$



①.  $\log \text{Reg.}$

②. SVM

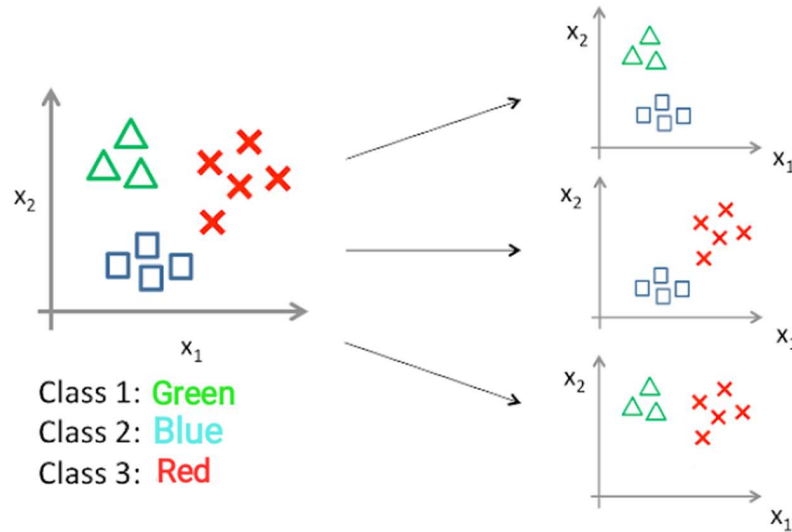
③. Tree

# One-vs-All

- Модель  $a_k(x)$  при обучении не знает, что её выходы будут сравнивать с выходами других моделей
- Нужно обучать  $K$  моделей

$\binom{K}{2}$  моделей

## All-vs-All (One-vs-One)





GA

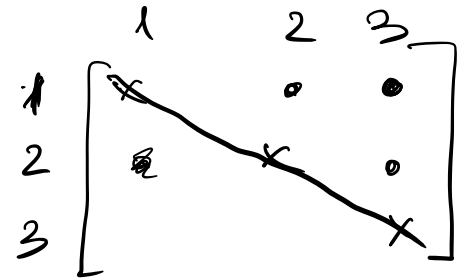
## All-vs-All

- $X_{km} = \{(x_i, y_i) \in X \mid y_i = k \text{ или } y_i = m\}$
- Обучаем  $a_{km}(x)$  на  $X_{km}$
- Итоговая модель:

$$\textcircled{1} \quad a(x) = \arg \max_{k \in \{1, \dots, K\}} \sum_{m=1}^K [a_{km}(x) = k]$$

↗  $\leftarrow$  голосование

$\left\{ \textcircled{2} \right.$  процедура  
скорее  
верно.



$\textcircled{1}$  vs  $2, 3$   
 $\begin{bmatrix} 0.51 & 0.49 \end{bmatrix}$   
 $\begin{bmatrix} 0.9 & 0.1 \end{bmatrix}$  x  
 $\begin{bmatrix} 0.51 & 0.49 \end{bmatrix}$

# All-vs-All

- Нужно обучать порядка  $K^2$  моделей
- Зато каждую обучаем на небольшой выборке

# Доля ошибок

- Функционал ошибки — доля ошибок (error rate)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

- Нередко измеряют долю верных ответов (accuracy):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

- Подходит для многоклассового случая!

confusion  
matrix

## Общие подходы

### Микро-усреднение

Вычисляем  $TP_k, FP_k, FN_k, TN_k$  для каждого класса

Суммируем по всем классам, получаем TP, FP, FN, TN

Подставляем их в формулу для precision/recall/...

$$\text{Precision} = \frac{\sum_k TP_k}{\sum_k TP_k + \sum_k FP_k}$$

$C_{ij}$  = число объектов  
которым присвоили  $j$ ,  
а они  $i$

### Макро-усреднение

Вычисляем нужную метрику для каждого класса (например,  $\text{precision}_1, \dots, \text{precision}_K$ )

Усредняем по всем классам

$$\text{Precision} = \frac{\sum_k \text{Precision}_k}{K}$$

# Общие подходы

Микро-усреднение

Крупные классы вносят большой вклад

Макро-усреднение

Игнорирует размеры классов

- one-hot  $N$  категорий  $\rightarrow [0 \dots 1_N \dots 0]$  \*  $\begin{cases} \text{ai индекс} \\ \text{работает} \end{cases}$   
 $=$  очень много  $< 10$  + метрируемость  
+ не нормуется
- Label Encoder  $N$  категорий  $\rightarrow \{1 \dots N\}$   
+ & строка | - порядок следов.  
+ число | - не метрику
- метки

## Работа с категориальными признаками

# Кодирование категориальных признаков


Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

# Label encoding

- Значения признака «район»:  $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо  $x_j$ : каждая категория заменяется числом от 0 до  $m-1$
- **Label encoding**



# Label encoding

Район		Район
ЦАО		0
ЮАО		1
ЦАО		0
САО		2
ЮАО		0

# Label encoding

- Label encoding может плохо работать для категориальных признаков, но хорошо – для порядковых

# One-hot encoding

- Значения признака «район»:  $U = \{u_1, \dots, u_m\}$
- Новые признаки вместо  $x_j$ :  $[x_j = u_1], \dots, [x_j = u_m]$
- **One-hot encoding**

# One-hot encoding

Район	ЦАО	ЮАО	САО
ЦАО	1	0	0
ЮАО	0	1	0
ЦАО	1	0	0
САО	0	0	1
ЮАО	0	1	0

# One-hot encoding

- One-hot encoding может плохо работать в случае большого числа категориальных признаков с большим числом категорий

Счетчик.

# Mean encoding

целевая  
переменная

модель  
вещ.  
функция  
которая  
уже ada

Район	Цена
ЦАО	10.000.000
ЮАО	4.000.000
ЦАО	9.000.000
САО	7.000.000
ЮАО	5.000.000

# Mean encoding

- Не хотим сильно увеличивать размер выборки только из-за кодирования признаков
- Хотим передать информацию о целевой переменной в данные – это может позволить ускорить обучение
- **Mean encoding (target encoding)**

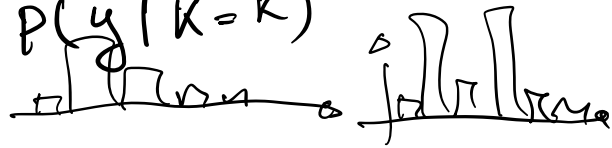
# Mean encoding

- Значения признака  $x_j$ :  $U_j = \{u_1, \dots, u_m\}$
- Посчитаем все категории в обучающей выборке:

$$\text{count}(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p]$$



$$P(y | K=k)$$



размер

- mean
  - mode
  - quantile  $[0.25, 0.75, 0.9]$
  - std
  - min, max
- размер

Mean encoding  $\rightarrow$  предикт

- Значения признака  $x_j$ :  $U_j = \{u_1, \dots, u_m\}$
- Для регрессии посчитаем суммарный ответ в категории:

$\forall k \in K$



$$\text{target}(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p] y_i$$

$(K_1) K_2 \rightarrow K_3$   
 $K \rightarrow \text{best}$   
 $K \rightarrow \text{max}$

$$F[K=k] \in \mathbb{R}$$

$\downarrow$  число  
 значение

а также берем

# Mean encoding

- Значения признака  $x_j$ :  $U_j = \{u_1, \dots, u_m\}$
- Для классификации посчитаем классы в категории:

$$\text{target}_k(j, u_p) = \sum_{i=1}^{\ell} [x_{ij} = u_p] [y_i = k]$$

# Mean encoding

- Задача регрессии
- Заменим категориальный признак на числовой:

$$\widetilde{x_{ij}} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$$

# Mean encoding

- Задача классификации
- Заменим категориальный признак на  $K$  числовых:

$$\widetilde{x}_{ij} = \left( \frac{\text{target}_1(j, x_{ij})}{\text{count}(j, x_{ij})}, \dots, \frac{\text{target}_K(j, x_{ij})}{\text{count}(j, x_{ij})} \right)$$

# Mean encoding

Район	Цена
ЦАО	10.000.000
ЮАО	4.000.000
ЦАО	9.000.000
САО	7.000.000
ЮАО	5.000.000



Район	Счётчик	Цена
ЦАО	9.500.000	10.000.000
ЮАО	4.500.000	4.000.000
ЦАО	9.500.000	9.000.000
САО	7.000.000	7.000.000
ЮАО	4.500.000	5.000.000

# Mean encoding

- В отличие от label encoding, где мы кодируем признак случайными категориями, тут намного больше смысла
- Однако, раз мы добавляем информацию о целевой переменной в данные, то можно легко переобучиться

ошб +  $10^2 \cdot \text{pr. random.normal}(0, 1, K)$

## Борьба с переобучением в счётчиках

- Решение 1: добавление шума +  $N(0, ?)$   
- 10 10

Район	Счётчик	Цена
ЦАО	9.500.000	10.000.000
ЮАО	4.500.000	4.000.000
ЦАО	9.500.000	9.000.000
САО	7.000.000	7.000.000
ЮАО	4.500.000	5.000.000



Район	Счётчик	Цена
ЦАО	9.130.000	10.000.000
ЮАО	4.023.000	4.000.000
ЦАО	10.124.000	9.000.000
САО	7.942.000	7.000.000
ЮАО	4.728.000	5.000.000

# Борьба с переобучением в счётчиках

- Решение 2: добавление априорных величин в счётчики (сглаживание)

$$\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij}) + a}{\text{count}(j, x_{ij}) + b}$$

- Например:

$$\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij}) + w * \text{mean}(y)}{\text{count}(j, x_{ij}) + w}$$



# Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

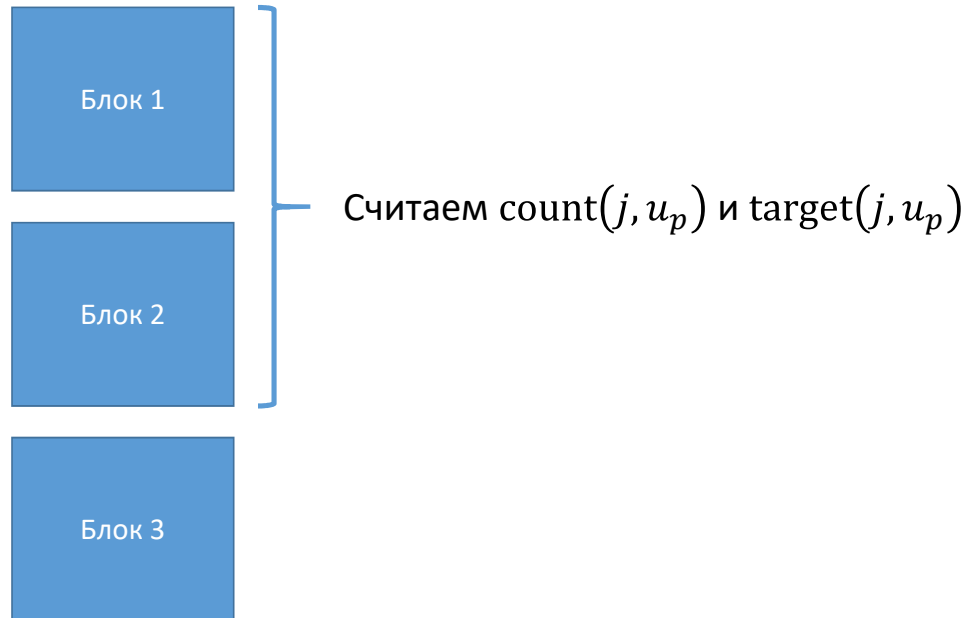
Блок 1

Блок 2

Блок 3

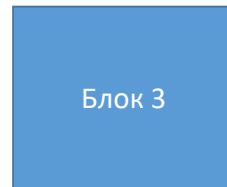
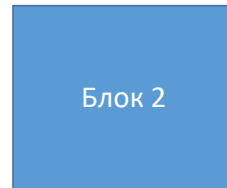
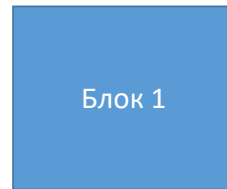
# Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



# Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

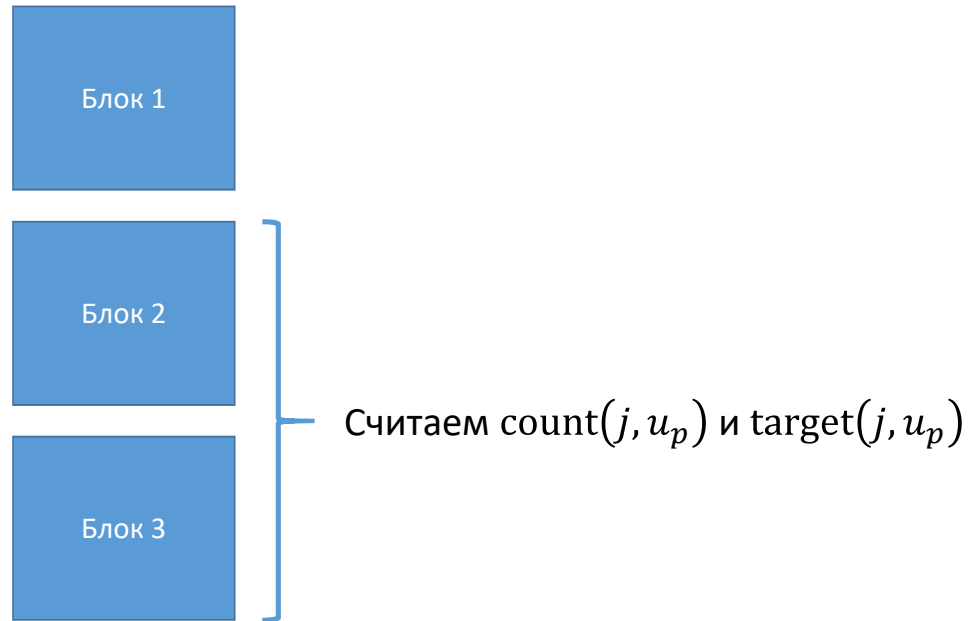


Считаем  $\text{count}(j, u_p)$  и  $\text{target}(j, u_p)$

Вычисляем признаки:  $\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$

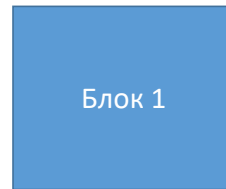
# Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков

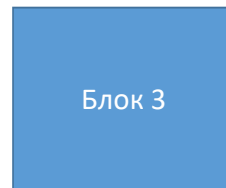
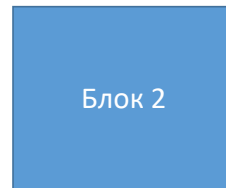


# Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Вычисляем признаки:  $\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij})}{\text{count}(j, x_{ij})}$



Считаем  $\text{count}(j, u_p)$  и  $\text{target}(j, u_p)$

# Mean encoding

- Mean encoding позволяет заменить категориальный признак на один числовой
- Могут привести к переобучению
- Можно бороться с ним через добавление шума, априорных значений или кросс-валидацию



# Методы Балансировки

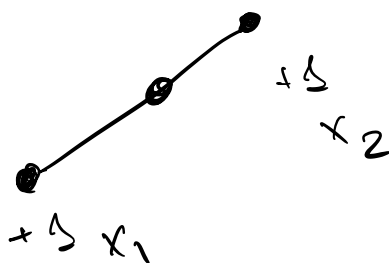
①  $\alpha$  Bootstrap  $\rightarrow$  oversampling minority  
 $\sim$  искусственные веса для объектов

$\sum_{i=1}^n f(x_i)$   $\rightarrow$   $\sum_{i=1}^n f(x_i) \cdot \alpha_i$

$\rightarrow$  взвешенный класс в классе  
 $\hookrightarrow$  как выбрать тот же

② undersampling  $\oplus$   
 $\downarrow$  уменьшил majority  
 $\downarrow$  оставил hard examples

## ③ SMOTE (sklearn)



$x \in [0, 1]$   $\alpha x_1 + (1 - \alpha)x_2 = x_{new}$

③ модификация loss критерия оптимизации

как получить много классов сразу?

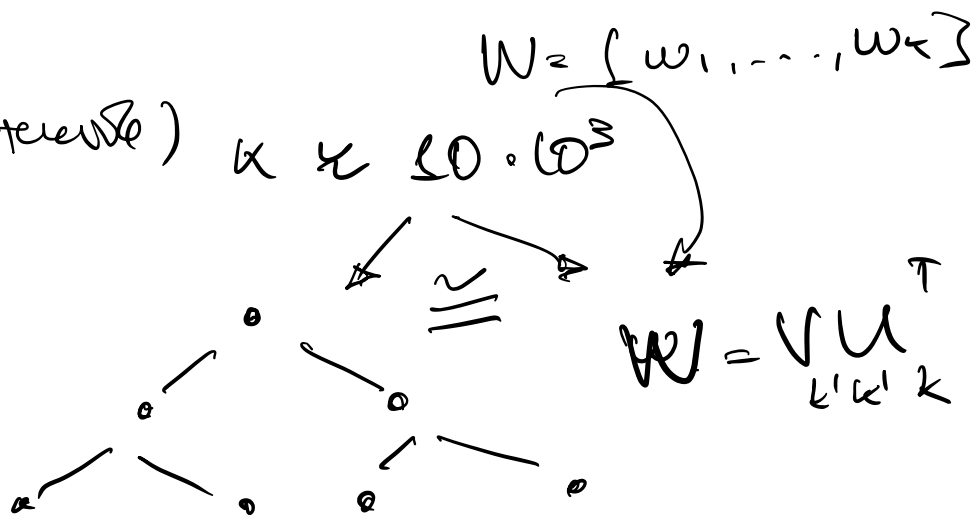
$\in \mathbb{R}^K$

③. Логистическая регрессия  $\Rightarrow \frac{\exp(\langle x, w_i \rangle)}{\sum_{i=1}^K \exp(\langle x, w_i \rangle)}$

когда  $K \leq 10$  ✓

Тестами (разрабатываем тесты)  $K \leq 10 \cdot 10^3$

100 ①  
010 ②  
111 ③



$x \in \mathbb{R}^d$

$f(x) \in \mathbb{R}^K$

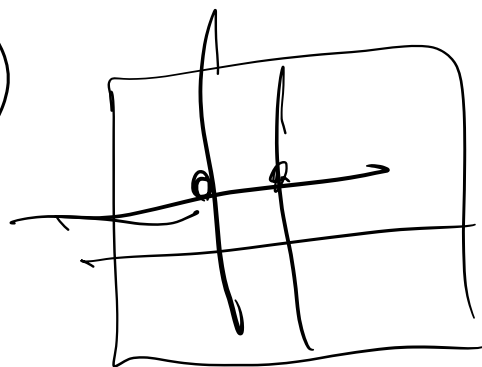
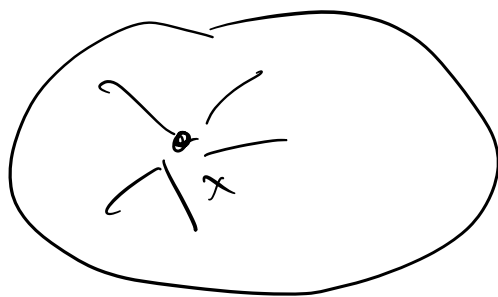
kd tree

$W \times K \times d$

Memory Time

②. KNN

↓  
дом  
классов  
вызвѣт



③. SVM



④ One class - SVM | Isolation tree

• ~~output~~ typical, we around

