A decorative graphic on the left side of the slide consists of a grid of colored squares. The grid is 4 columns wide and 4 rows high. The colors of the squares are: Row 1: Teal, Orange, Brown, Teal; Row 2: Orange, Brown, Tan, Tan; Row 3: Orange, Teal, Tan, Tan; Row 4: Tan, Orange, Orange, Brown.

Многоклассовая классификация, работа с категориальными признаками

Повторение

Классификация

- $Y = \{-1, +1\}$
- -1 – отрицательный класс
- +1 – положительный класс
- Алгоритм $a(x)$ должен возвращать одно из двух чисел

Линейный классификатор

$$a(x) = \text{sign}(w_0 + \sum_{j=1}^d w_j x_j) - \text{выдает знак вещественного числа}$$

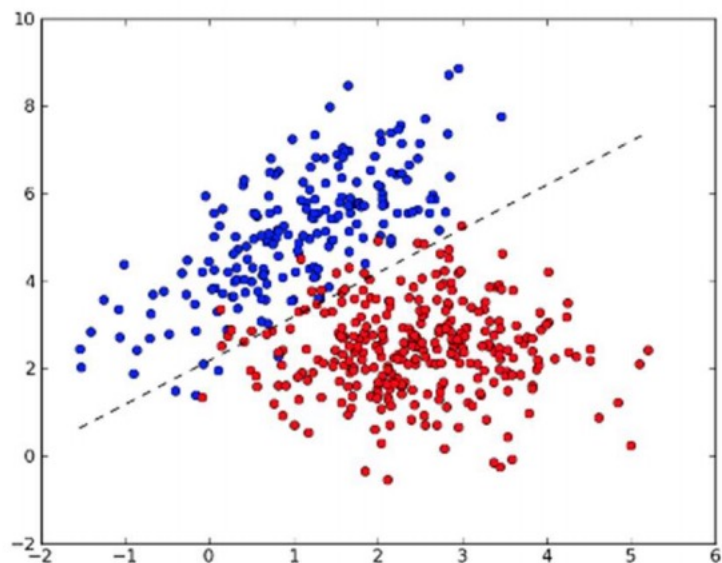
Свободный
коэффициент

веса

признаки

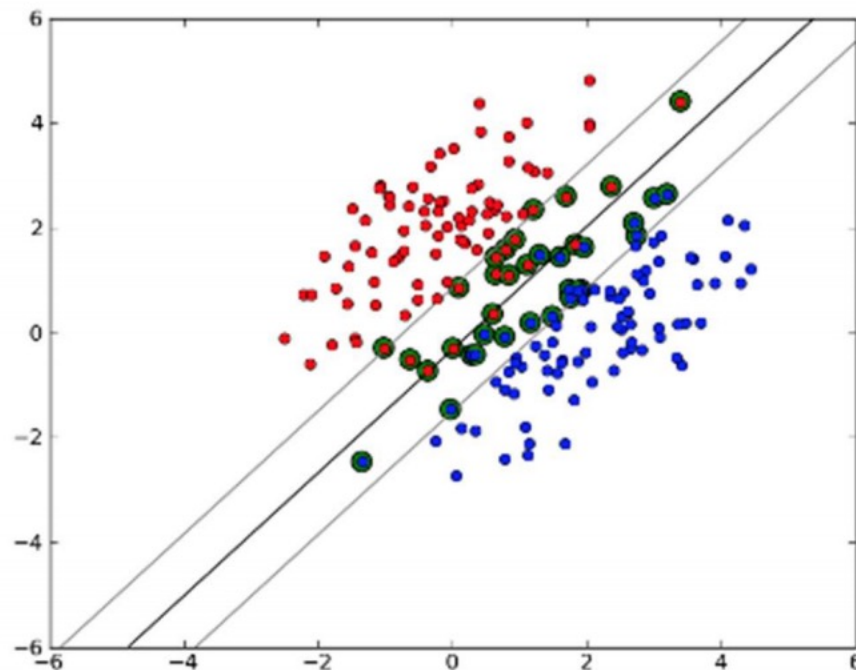
Геометрия

- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$ – объект слева от нее
- $\langle w, x \rangle > 0$ – объект справа от нее



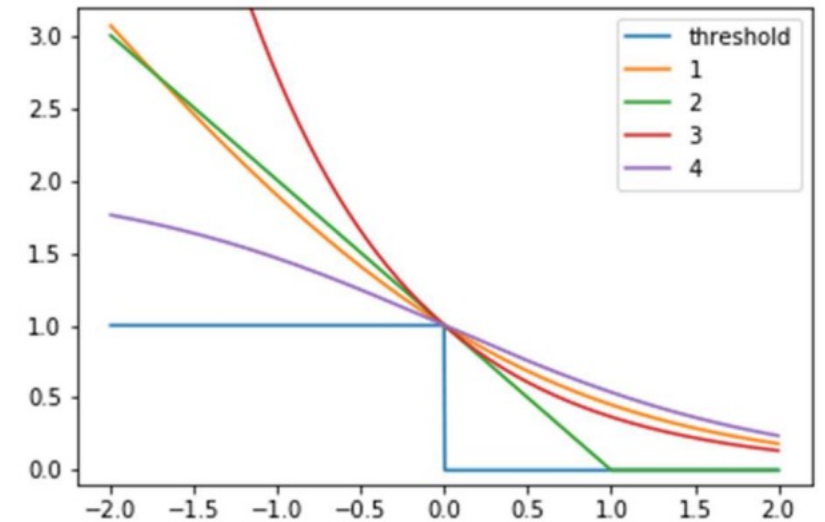
Отступ

- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$ - классификатор дает верный ответ
- $M_i < 0$ - классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



Примеры верхних оценок

- $L^{\sim}(M) = \log(1 + e^{-M})$ – логистическая
- $L^{\sim}(M) = \max(0, 1 - M)$ – кусочно – линейная
- $L^{\sim}(M) = e^{-M}$ – экспоненциальная
- $L^{\sim}(M) = \frac{2}{1+e^M}$ - сигмоидная



Матрица ошибок

	$Y = 1$	$Y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

Логистическая регрессия

- Решаем задачу бинарной классификации: $\mathbb{Y} = \{-1, +1\}$

- Минимизация верхней оценки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log(1 + \exp(-y_i \langle w, x_i \rangle)) \rightarrow \min_w$$

Линейный классификатор

$$a(x) = \text{sign } \langle w, x \rangle$$

- Обучим как-нибудь — например, на логистическую функцию потерь:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \log(1 + \exp(-y_i \langle w, x_i \rangle)) \rightarrow \min_w$$

- Может, $\langle w, x \rangle$ сойдёт за оценку?

Линейный классификатор

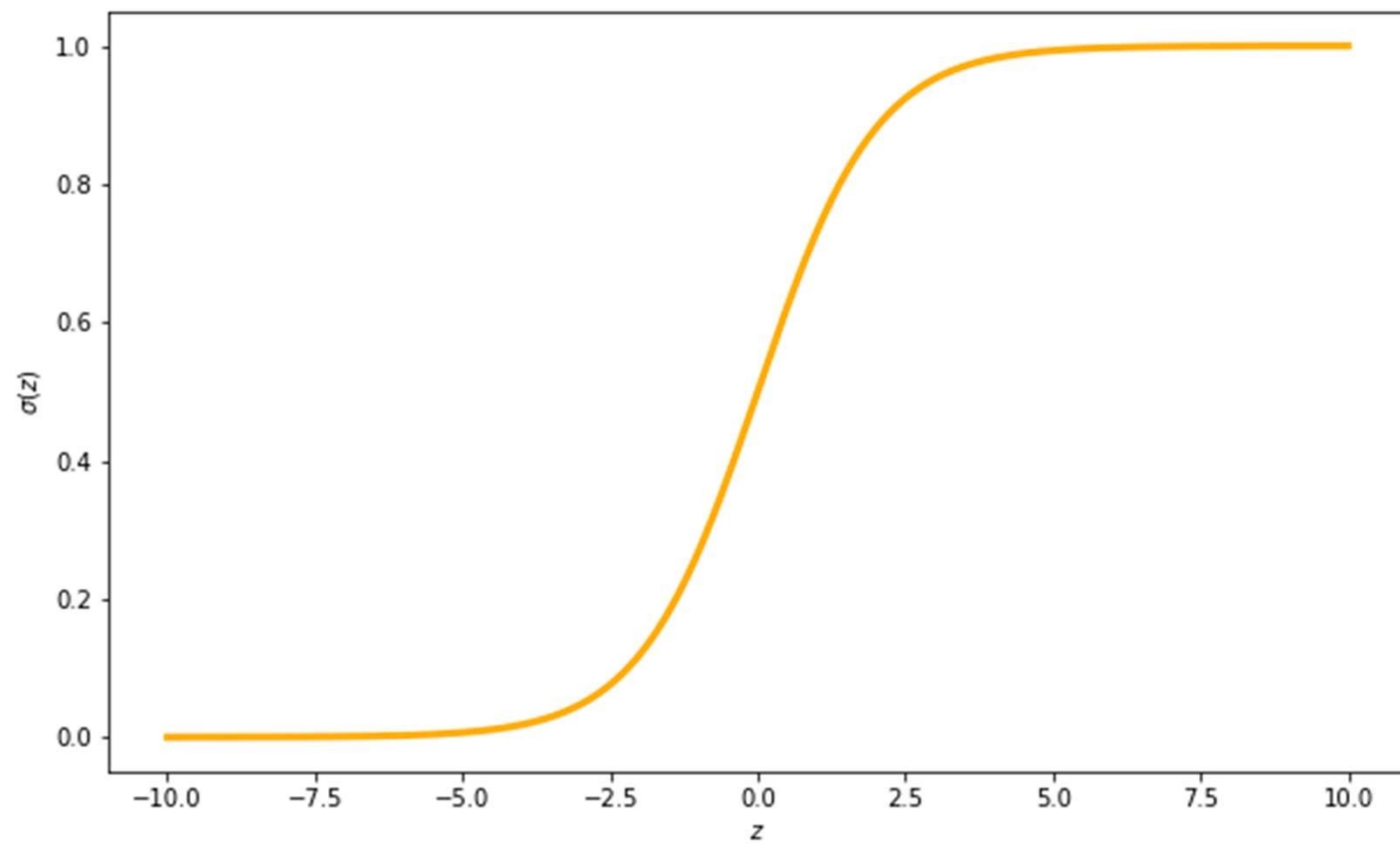
- Переведём выход модели на отрезок $[0, 1]$
- Например, с помощью сигмоиды¹:

$$\sigma(\langle w, x \rangle) = \frac{1}{1 + \exp(-\langle w, x \rangle)}$$

¹<https://sebastianraschka.com/faq/docs/logistic-why-sigmoid.html>

Сигмоида

...



Предсказание вероятностей

$$-\sum_{i=1}^{\ell} \{ [y_i = 1] \log \sigma(\langle w, x_i \rangle) + [y_i = -1] \log(1 - \sigma(\langle w, x_i \rangle)) \} \rightarrow \min_w$$

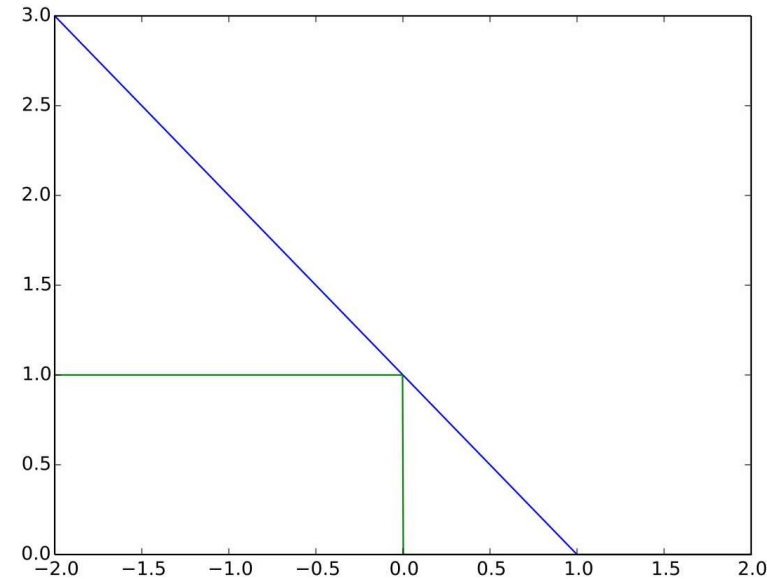
- Если $y_i = +1$ и $\sigma(\langle w, x_i \rangle) = 0$, то штраф равен $-\log 0 = +\infty$
- Достаточно строго
- Функция потерь называется **log-loss**

$$L(y, z) = -[y = 1] \log z - [y = -1] \log(1 - z)$$

Hinge loss

- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$
- Минимизация верхней оценки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} \max(0, 1 - y_i \langle w, x_i \rangle) \rightarrow \min_w$$



Метод опорных векторов (SVM)

$$\begin{cases} \|w\|^2 \rightarrow \min_{w, w_0} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 \end{cases}$$

Метод опорных векторов

$$\begin{cases} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, w_0, \xi_i} \\ y_i(\langle w, x_i \rangle + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

- Объединим ограничения:

$$\xi_i \geq \max(0, 1 - y_i(\langle w, x_i \rangle + w_0))$$

Метод опорных векторов

$$C \sum_{i=1}^{\ell} \max(0, 1 - y_i(\langle w, x_i \rangle + w_0)) + \|w\|^2 \rightarrow \min_{w, w_0}$$

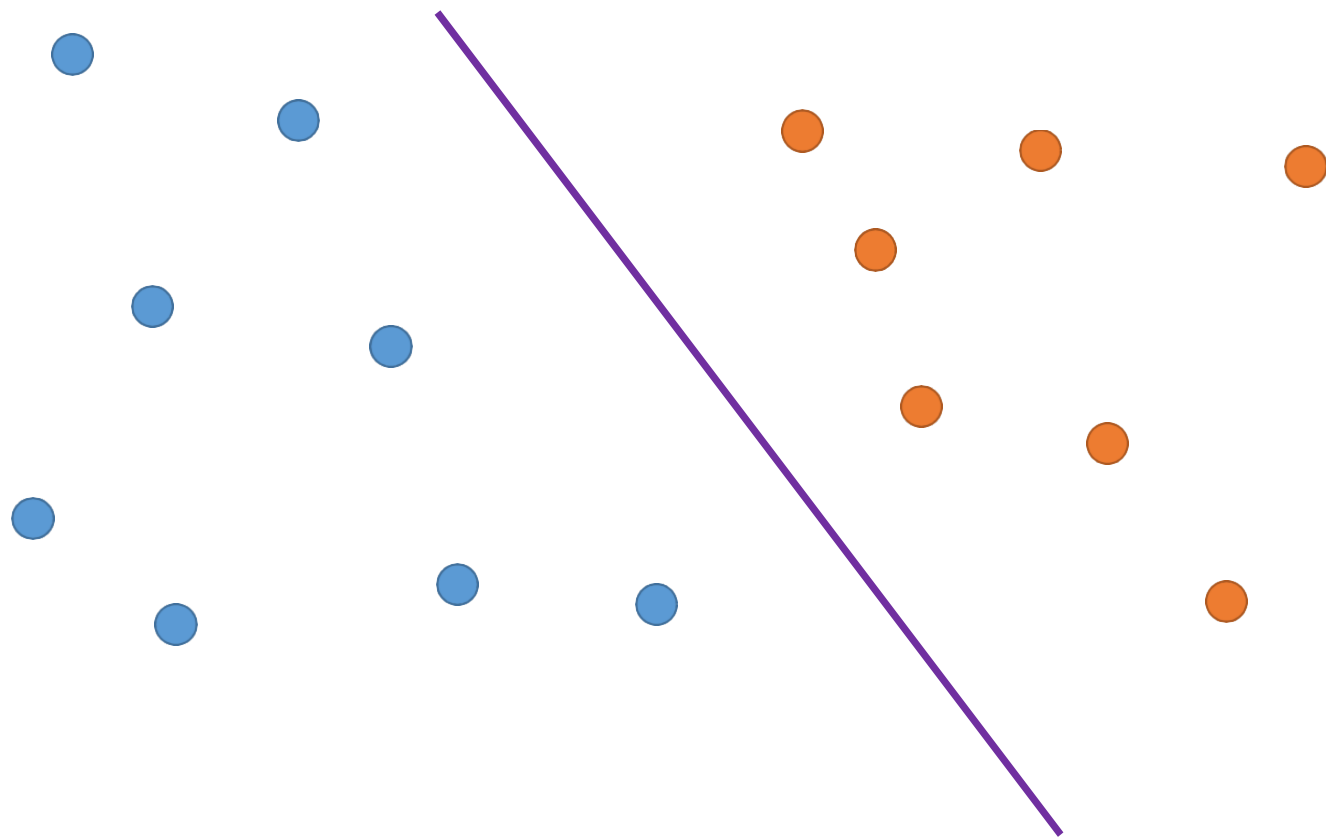
- Функция потерь (hinge loss) + регуляризация

Резюме

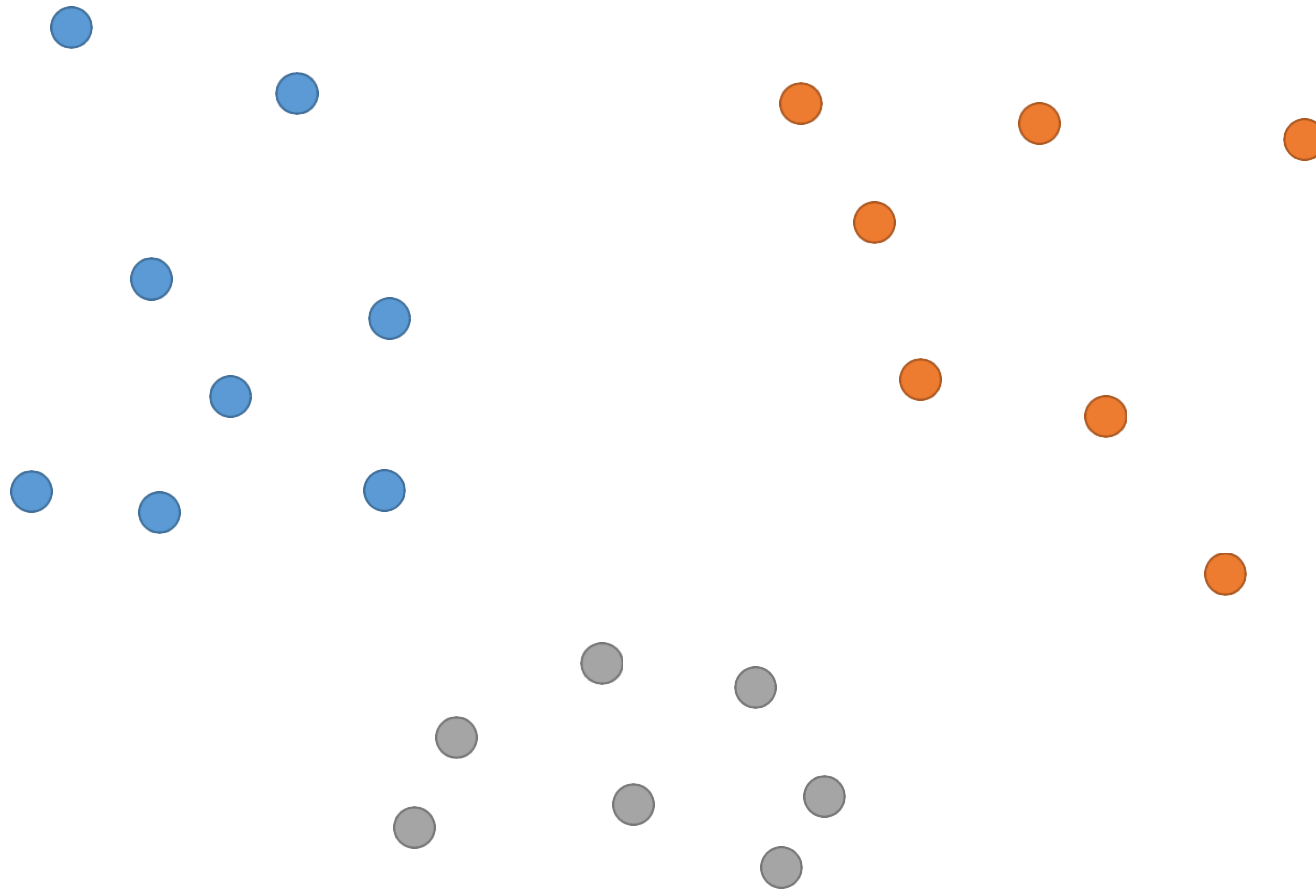
- Логистическая регрессия минимизирует логистические потери
- Метод опорных векторов основан на идее максимизации отступа классификатора

Многоклассовая классификация

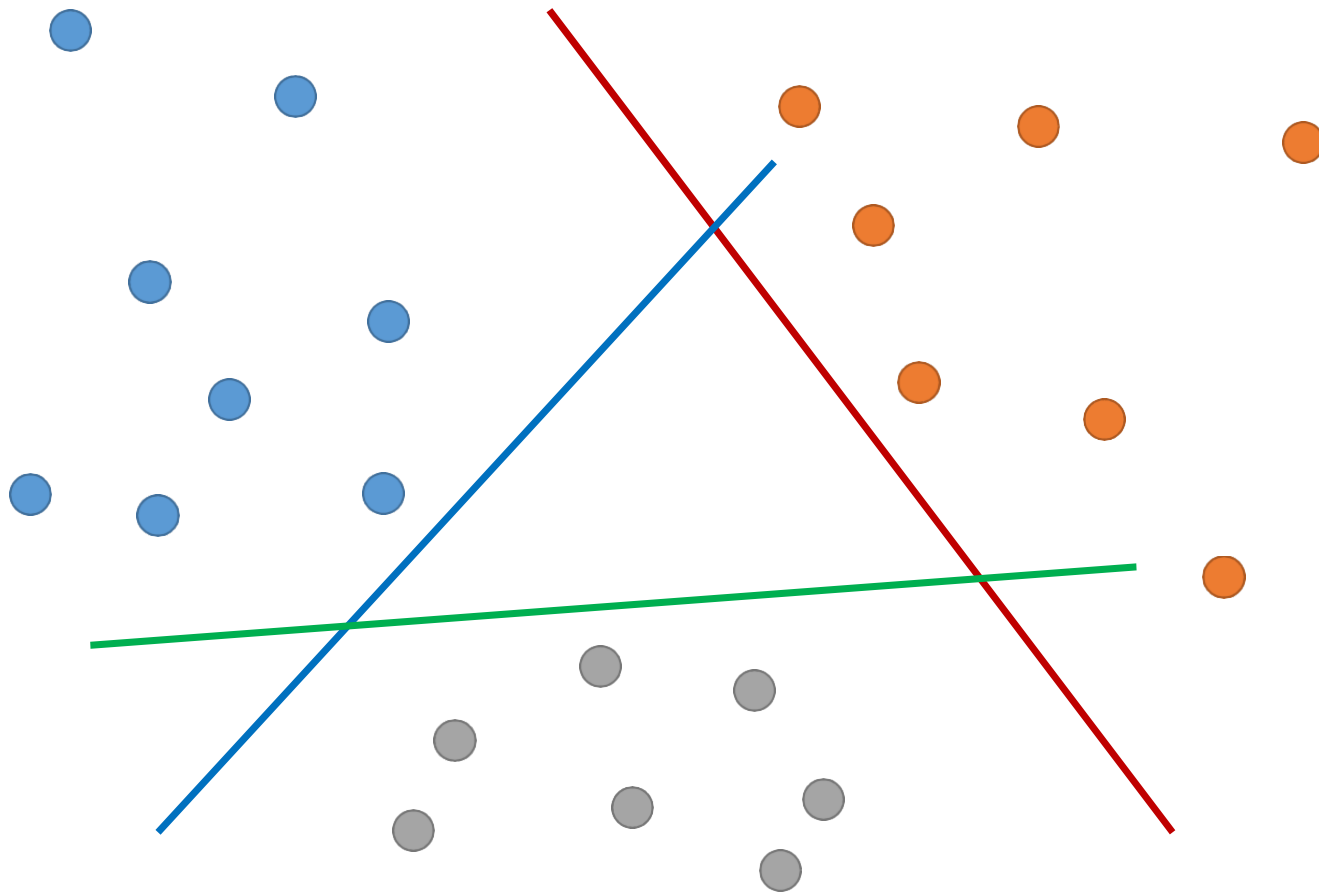
Бинарная классификация



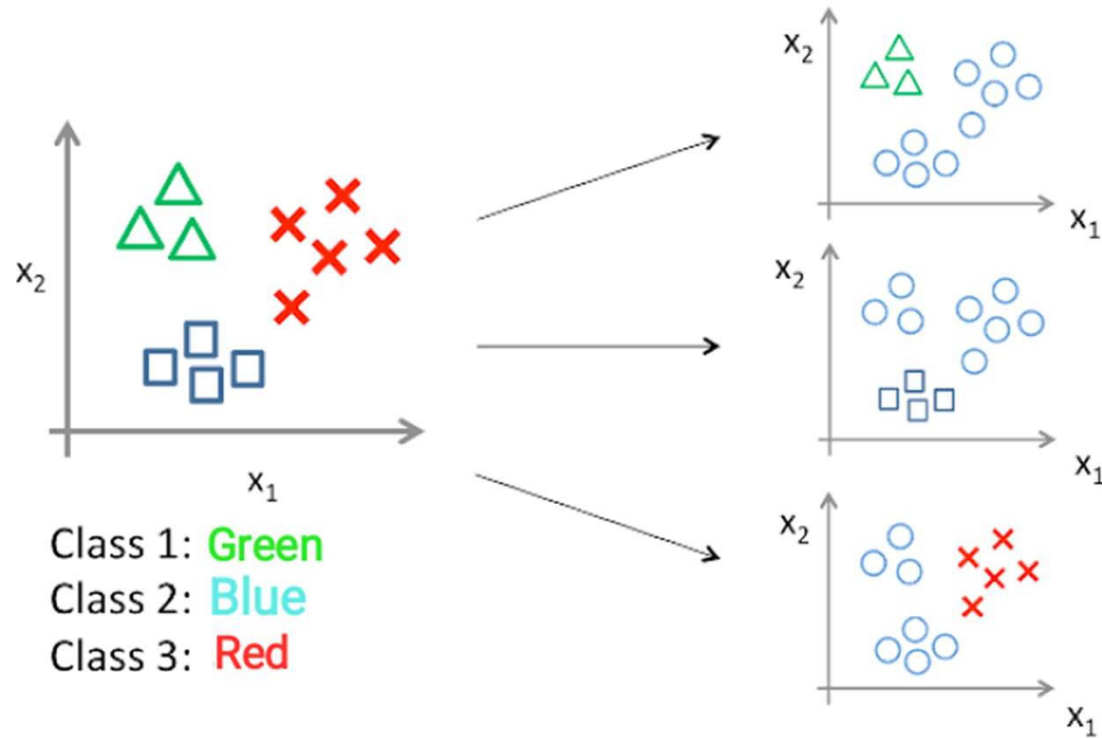
Многоклассовая классификация



Многоклассовая классификация



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



One-vs-All

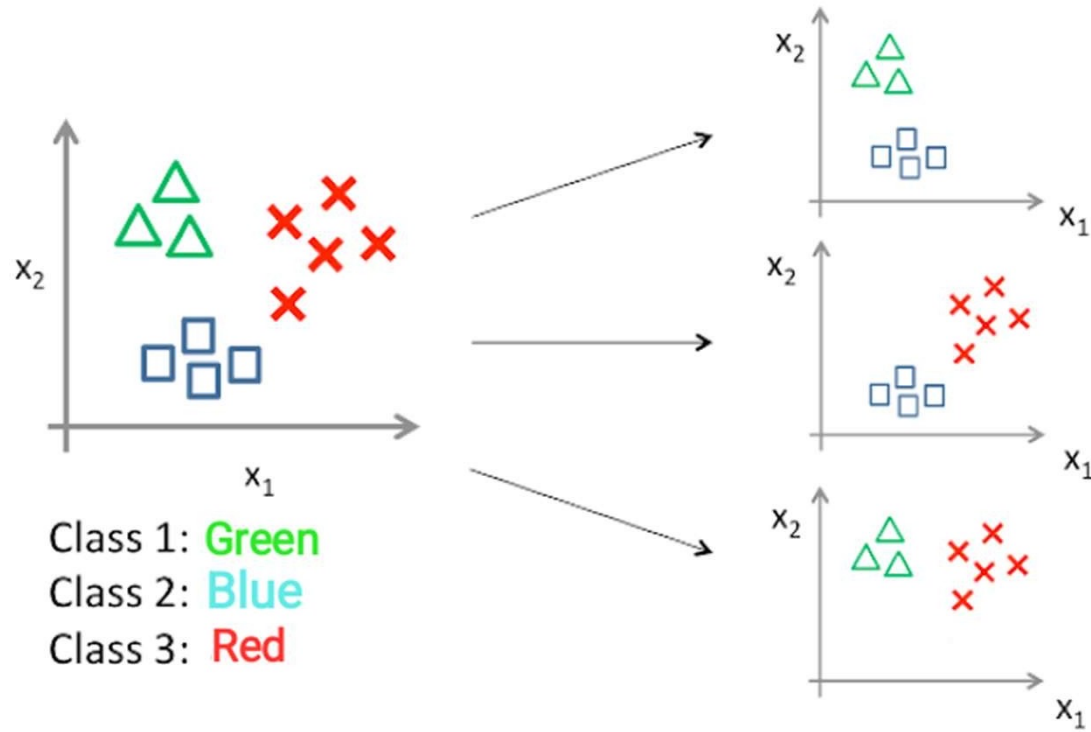
- K классов: $\mathbb{Y} = \{1, \dots, K\}$
- $X_k = (x_i, [y_i = k])_{i=1}^{\ell}$
- Обучаем $a_k(x)$ на X_k , $k = 1, \dots, K$
- $a_k(x)$ должен выдавать оценки принадлежности классу (например, $\langle w, x \rangle$ или $\sigma(\langle w, x \rangle)$)
- Итоговая модель:

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

One-vs-All

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

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



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:

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

All-vs-Al

- 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!

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/...

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

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 больший вклад

Макро-усреднение

Игнорирует размеры классов

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	ЦАО
Район	ЮАО
Район	ЦАО
Район	САО
Район	ЮАО



Район	0
Район	1
Район	0
Район	2
Район	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

Mean encoding

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:

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

Mean encoding

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

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

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

Борьба с переобучением в счётчиках

- Решение 1: добавление шума

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)

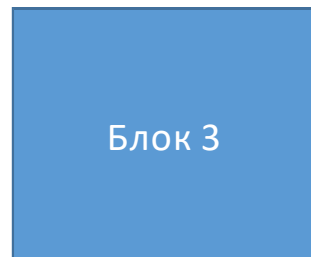
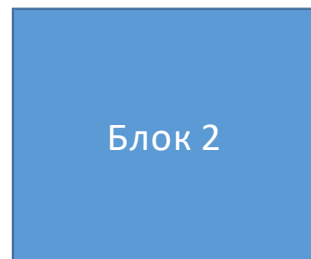
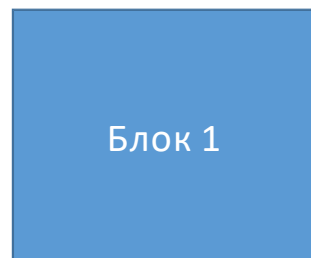
$$\widetilde{x}_{ij} = \frac{\text{target}(j, x_{ij}) + a}{\text{count}(j, x_{ij}) + b}$$

- Например:

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

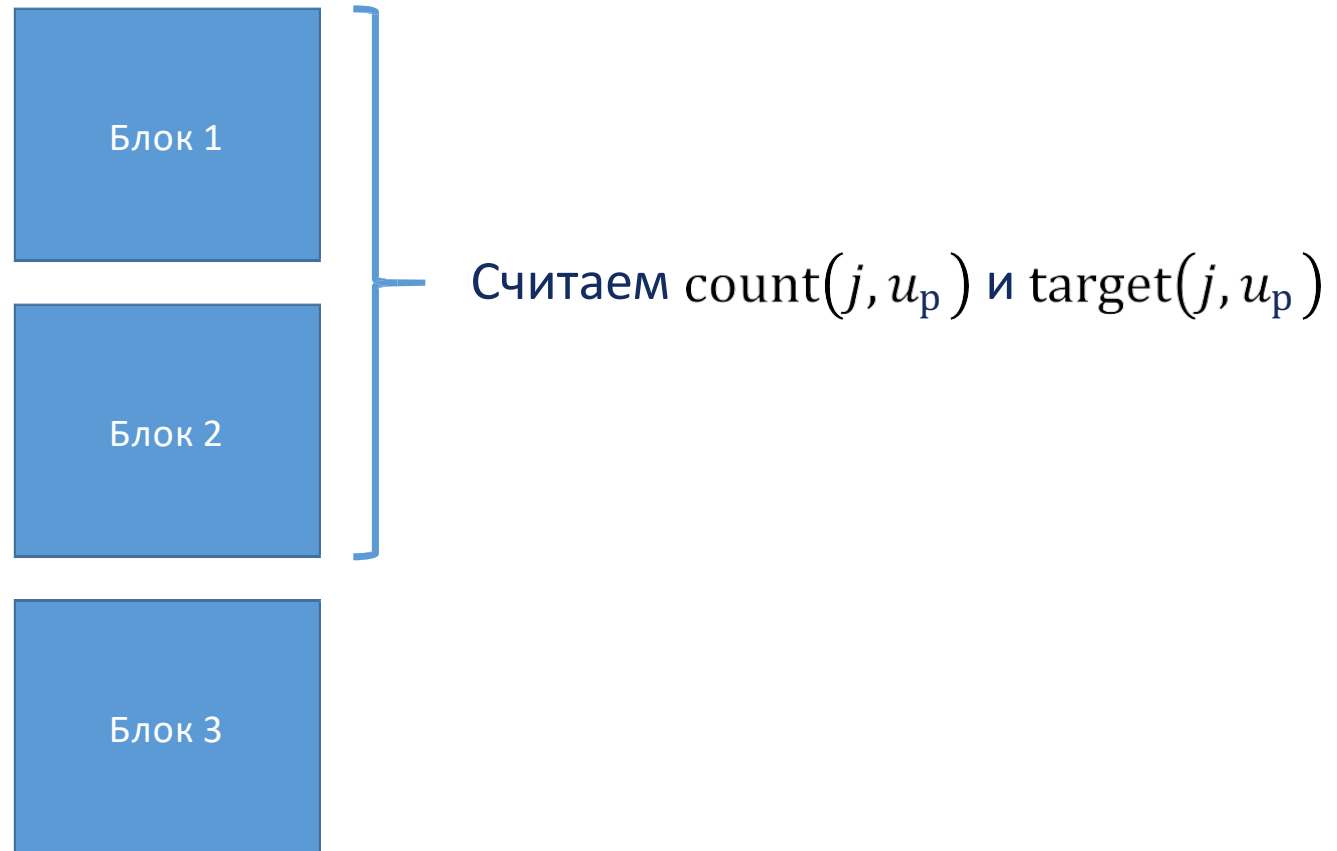
Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



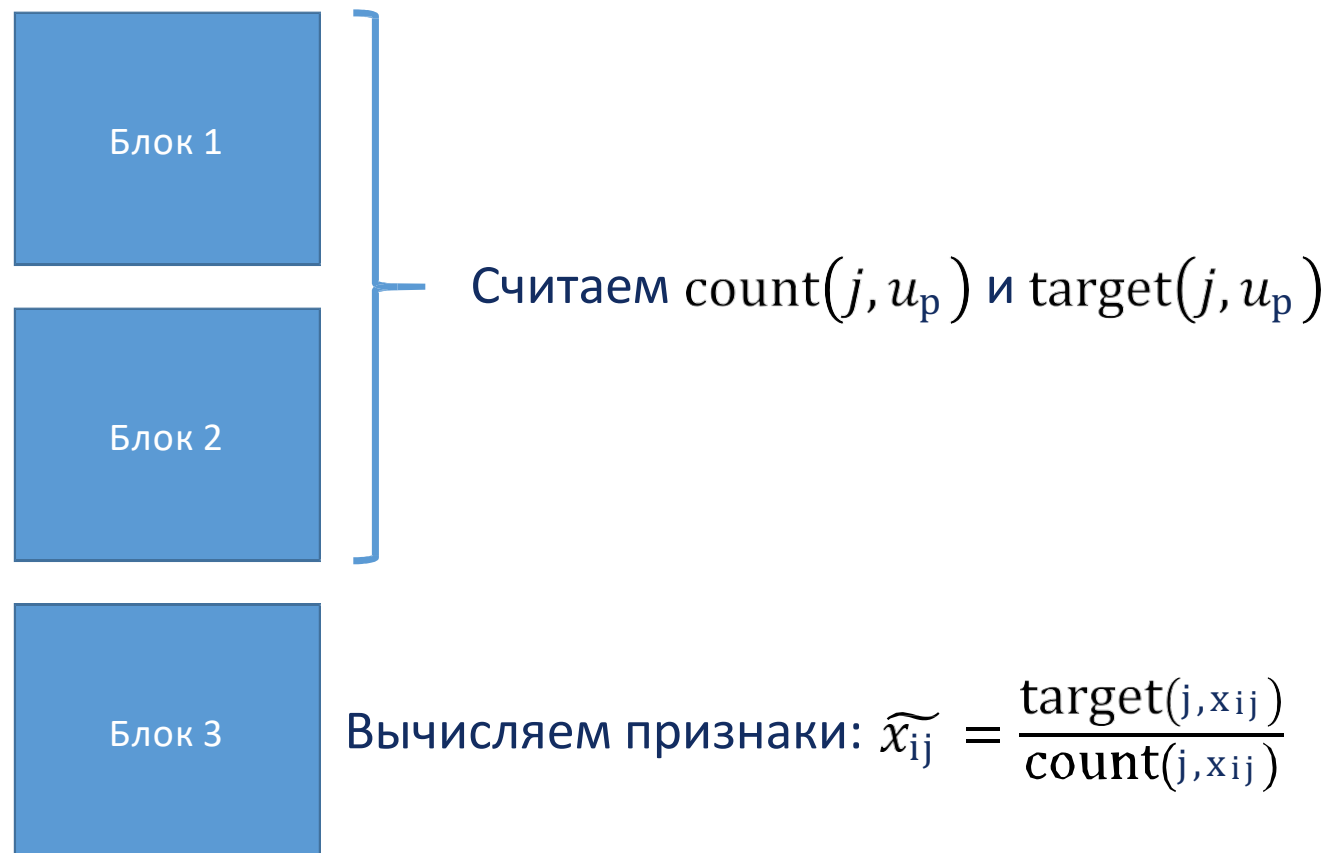
Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



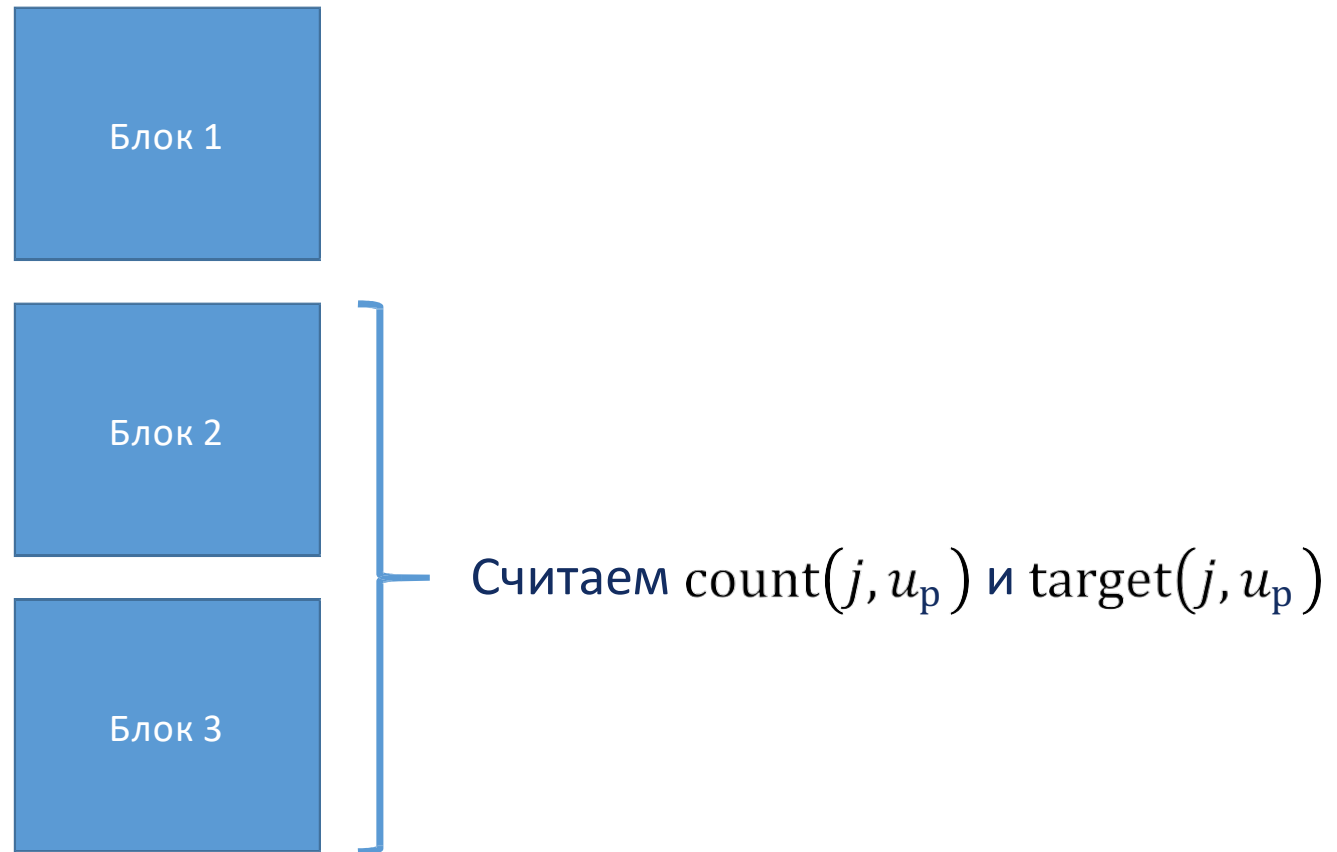
Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Борьба с переобучением в счётчиках

- Решение 3: кросс-валидация счётчиков



Борьба с переобучением в счётчиках

- 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

Спасибо за внимание!



Ildar Safilo

@Ildar_Saf

irsafilo@gmail.com

<https://www.linkedin.com/in/isafilo/>