START ML

KARPOV.COURSES

ЭМПИРИЧЕСКИЙ ФАКТ

В ПЕРЕОБУЧЕННЫХ МОДЕЛЯХ ОКАЗЫВАЮТСЯ БОЛЬШИЕ ВЕСА

Пусть предсказываем цену квартиры и получаем следующую модель:

$$\text{цена} = 3.000.000 \cdot \text{м}^2 - 5.000.000 \cdot \text{расстояние до метро}$$

Интерпретация: коэффициенты в ЛР показывают, как изменится прогноз при изменении соответствующего признака на единицу.

Даже если качество на обучающей выборке хорошее, видно, что некоторые коэффициенты оказываются неадекватно большими.

КАК БОРОТЬСЯ С ПЕРЕОБУЧЕНИЕМ?

КАК ПРИ ЭТОМ ОСТАТЬСЯ В КЛАССЕ ЛИНЕЙНЫХ МОДЕЛЕЙ?

Обычно при переобучении получаются достаточно большие веса модели.

Можно начать штрафовать нашу модель за каждую единицу в норме весов.

$$Q_{regul} = Q(a(x, \beta), X) + \lambda \cdot R(\beta) \to \min_{\beta}$$

КАК БОРОТЬСЯ С ПЕРЕОБУЧЕНИЕМ?

$$\text{цена} = 3.000.000 \cdot \text{м}^2 - 5.000.000 \cdot \text{расстояние до метро}$$

$$Q_{train}(\beta_1; \beta_2) \to min$$

$$Q^*_{train} = Q_{train}(\beta^*_1; \beta^*_2) = Q_{train}(3.000.000; -5.000.000) = X$$

Теперь давайте решать новую задачу. Будем минимизировать не просто нашу метрику, а сумму метрики и некоторого регуляризатора – функции, штрафующей нас за прирост в β

$$Q_{regul} = Q_{train}(\beta_1; \beta_2) + \lambda \cdot R(\beta)$$

$$Q_{regul} = Q_{train}(\beta_1; \beta_2) + 10 \cdot (\beta_1^2 + \beta_2^2) \to min$$

КАК БОРОТЬСЯ С ПЕРЕОБУЧЕНИЕМ?

$$Q_{train} = Q_{train}(\beta_1; \beta_2\,) + 10 \cdot (\beta_1^2 + \beta_2^2) \to min$$

Пусть мы нашли какие-то оптимальные коэффициентики в новой задаче! Они гарантированно будут меньше, чем раньше. Из-за регуляризатора. Например,

$$(\beta_1^{**}; \beta_2^{**}\,) = (20.000; -10.000)$$

$$(\beta_1^{*}; \beta_2^{*}\,) = (3.000.000; -5.000.000)$$

$$\text{цена}_{new} = 20.000 \cdot \text{м}^2 - 10.000 \cdot \text{расстояние до метро}$$

КАК БОРОТЬСЯ С ПЕРЕОБУЧЕНИЕМ?

$$Q_{regul} = Q_{train}(\beta_1; \beta_2) + 10 \cdot (\beta_1^2 + \beta_2^2) \to min$$

$$(\beta_1^{**}; \beta_2^{**}) = (20.000; -10.000)$$

$$Q_{regul}^{*} = \mathrm{Q}_{\mathrm{train}}(\beta_1^{**}; \beta_2^{**}) + 10 \cdot (\beta_1^{**2} + \beta_2^{**2})$$

При этом, очевидно, качество на трейне ухудшится. Ведь если бы оно могло не ухудшаться, то и изначально без регуляризатора мы получили бы $(\beta_1^{**}; \beta_2^{**})$.

Например, теперь $\mathrm{Q}_{\mathrm{train}}(\beta_1^{**}; \beta_2^{**}) = 1.1 \cdot X$

РАНЬШЕ:

$$\text{цена} = 3.000.000 \cdot \text{м}^2 - 5.000.000 \cdot \text{расстояние до метро}$$

$$Q^*_{train} = Q_{train}(3.000.000; -5.000.000) = X$$

С РЕГУЛЯРИЗАЦИЕЙ:

$$\text{цена}_{new} = 20.000 \cdot \text{м}^2 - 10.000 \cdot \text{расстояние до метро}$$

$$Q^*_{train} = Q_{train}(\beta_1^*; \beta_2^*) = Q_{train}(20.000; -10.000) = 1.1 \cdot X$$

Модель стала адекватнее!

Немного потеряли в качестве на тренировочной выборке

Зато есть надежда на то, что мы побороли переобучение, и на тестовых данных будет все супер!

КОМПРОМИСС

КАК МОЖНО БАЛАНСИРОВАТЬ МЕЖДУ КАЧЕСТВОМ И ПЕРЕОБУЧЕНИЕМ?

$$Q_{regul} = Q_{train}(\beta_1; \beta_2) + \lambda \cdot R(\beta)$$

$$\lambda = 10:$$

$$\text{цена}_{new} = 20.000 \cdot \text{м}^2 - 10.000 \cdot \text{расстояние до метро}$$

$$Q^*_{train} = Q_{train}(\beta_1^*; \beta_2^*) = Q_{train}(20.000; -10.000) = 1.1 \cdot X$$

$$\lambda = 3:$$

$$\text{цена}_{new} = 100.000 \cdot \text{м}^2 - 50.000 \cdot \text{расстояние до метро}$$

$$Q^*_{train} = Q_{train}(\beta_1^*; \beta_2^*) = Q_{train}(100.000; -50.000) = 1.05 \cdot X$$

L-2 И L-1 РЕГУЛЯРИЗАЦИЯ

КАКИМ БЫВАЕТ $R(\beta)$?

$$Ridge\ Regularization:\ R(\beta) = ||\beta||_2 = \sum \beta_i^2 = \beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots$$

$$Lasso\ Regularization:\ R(\beta) = ||\beta||_1 = \sum |\beta_i| = |\beta_1| + |\beta_2| + |\beta_3| + \cdots$$

$$Q_{L_2} = Q + \lambda \cdot (\beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots)$$

$$Q_{L_1} = Q + \lambda \cdot (|\beta_1| + |\beta_2| + |\beta_3| + \cdots)$$

L-2 И L-1 РЕГУЛЯРИЗАЦИЯ

КАК ЭТО ВЛИЯЕТ НА ГРАДИЕНТНЫЙ СПУСК?

$$Q_{L_2} = Q + \lambda \cdot (\beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots)$$
$$Q_{L_1} = Q + \lambda \cdot (|\beta_1| + |\beta_2| + |\beta_3| + \cdots)$$

Так, при градиентном спуске чем больше норма весов в текущей точке, тем радикальнее будет следующий шаг (в смысле длины шага), так как в градиенте появляется дополнительное слагаемое $\lambda \cdot \nabla R(\beta)$.

$$\nabla Q_{L_1} = \nabla Q + \lambda \cdot (\mathrm{sgn}(\beta_1) \quad \mathrm{sgn}(\beta_2) \quad \ldots)$$
$$\nabla Q_{L_2} = \nabla Q + \lambda \cdot (2\beta_1 \quad 2\beta_2 \quad \ldots)$$

РЕЗЮМЕ

- Эмпирически можно показать, что при переобучении появляются гигантские коэф-ты
- Узнали, как бороться с большими весами модели
- Поняли, что можем балансировать между переобучением и качеством на трейне
- Лучший λ стоит находить - экспериментировать.
- Познакомились с двумя видами регуляризации: L1, L2
- А что, если коэффициенты у некоторых признаков и вправду должны быть большими?

МАСШТАБИРОВАНИЕ

ВСЕГДА ЛИ СПРАВЕДЛИВО ШТРАФОВАТЬ ВЕСА?

Иногда весам нужно быть справедливо большими

$Y \sim 10^6$

$\text{цена квартиры} = \beta_1 \cdot \text{м}^2 + \beta_2 \cdot \text{расстояние до метро}$

$d_1 \sim 10^1$

$d_2 \sim 10^3$

$\beta_1 \sim 10^5$

$\beta_2 \sim 10^3$

МАСШТАБИРОВАНИЕ

ВСЕГДА ЛИ СПРАВЕДЛИВО ШТРАФОВАТЬ ВЕСА?

$$\text{цена квартиры} = \beta_1 \cdot \text{м}^2 + \beta_2 \cdot \text{расстояние до метро}$$

Пусть построили модель, у которой в итоге всего 1 коэффициент оказался неадекватным

$$\text{цена квартиры} = 200.000 \cdot \text{м}^2 - 500.000 \cdot \text{расстояние до метро}$$

После регуляризации наши изначально хорошие коэффициенты могут пострадать:

МАСШТАБИРОВАНИЕ

ХОЧЕТСЯ, ЧТОБЫ ПРИЗНАКИ БЫЛИ ПРИМЕРНО ОДНОГО ПОРЯДКА

Например, не так: $\text{цена квартиры} = \beta_1 \cdot \text{м}^2 + \beta_2 \cdot \text{расстояние до метро}$

А так: $\text{цена квартиры} = \beta_1 \cdot \text{дц}^2 + \beta_2 \cdot \text{расстояние до метро}$

Ведь тогда

$$d_1 \sim 10^3 \rightarrow \beta_1 \sim 10^3$$

$$d_2 \sim 10^3 \rightarrow \beta_2 \sim 10^3$$

МАСШТАБИРОВАНИЕ

StandardScaler

$$d_j = \frac{d_j - \mu}{\sigma}$$

$$\mu = \frac{1}{n} \cdot \sum d_j$$

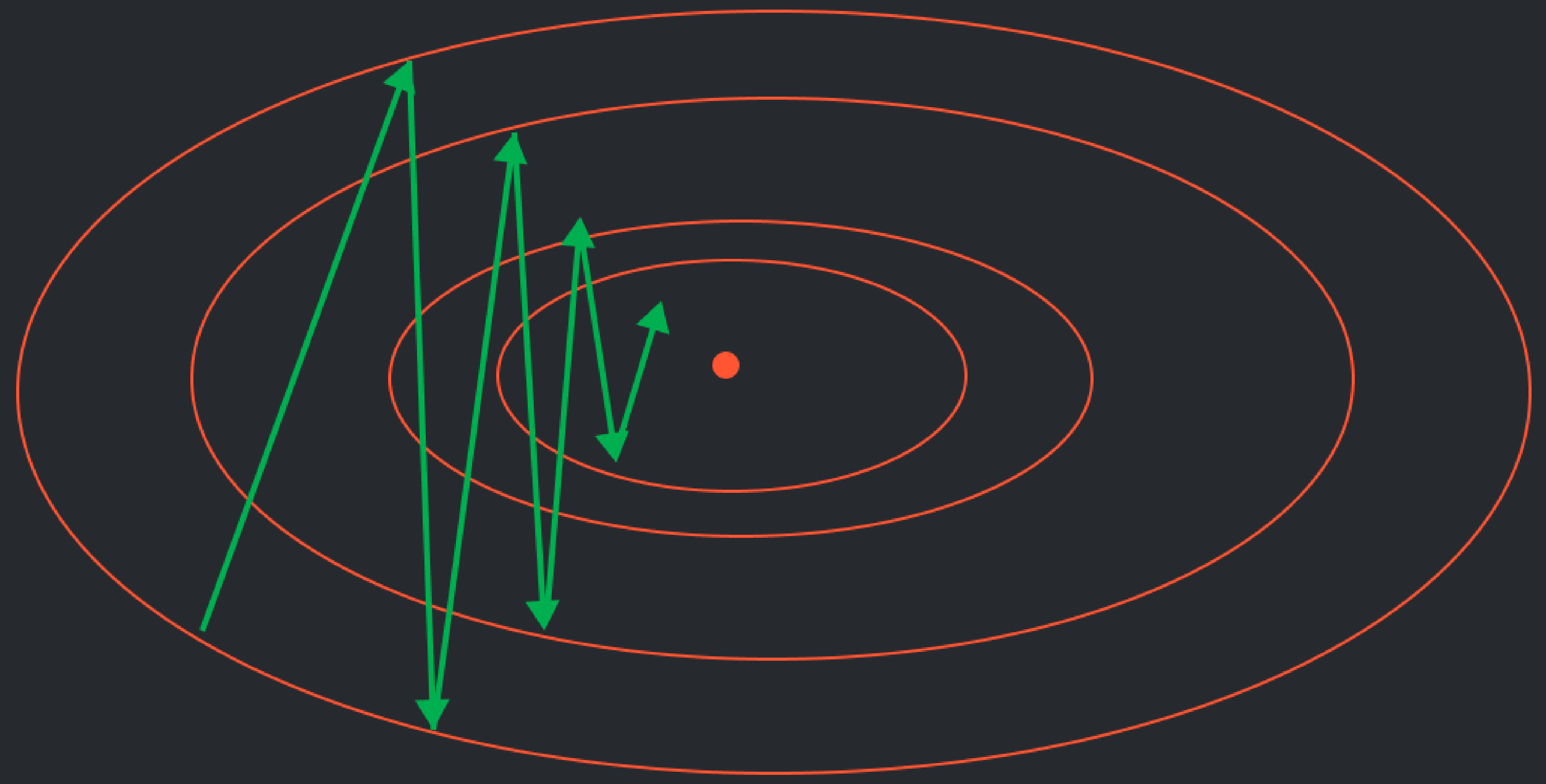
$$\sigma^2 = \frac{1}{n} \cdot \sum (d_j - \mu)^2$$

MinMaxScaler

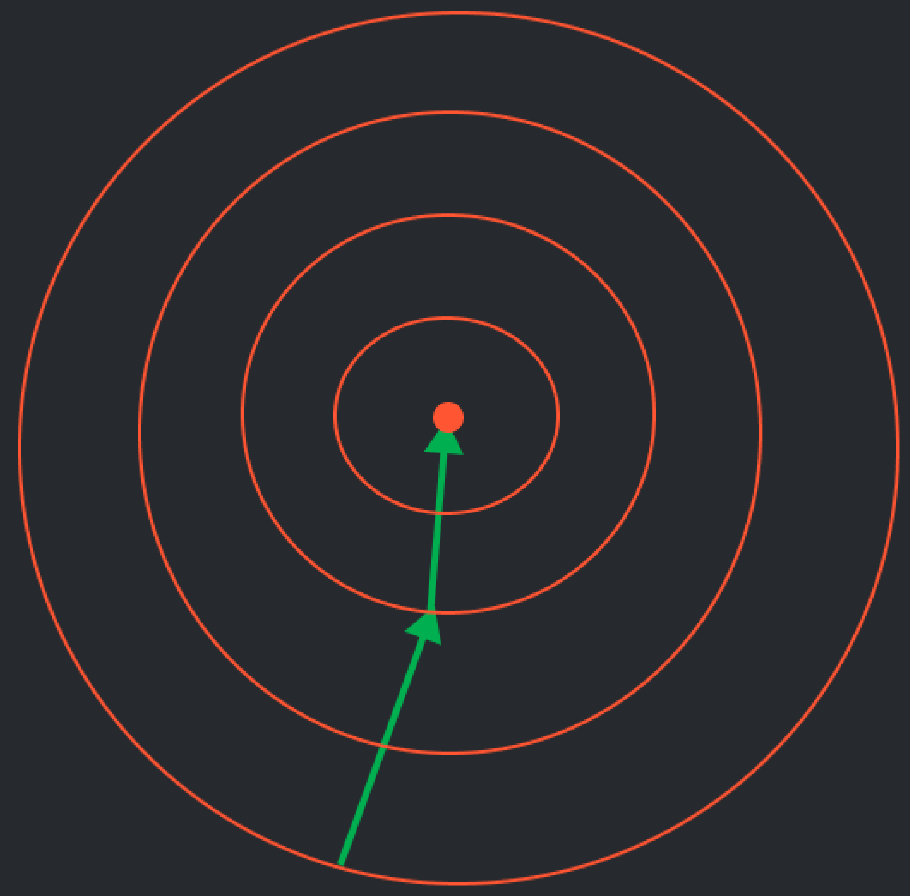
$$d_j = \frac{d_j - \min d}{\max d - \min d}$$

МАСШТАБИРОВАНИЕ: БЫСТРЕЕ СХОДИТСЯ ГРАДИЕНТ

$Q(w_1, w_2) = w_1^2 + 3w_2^2$



$Q(w_1, w_2) = w_1^2 + w_2^2$



МАСШТАБИРОВАНИЕ

ПРЕДСКАЗЫВАЕМ ЦЕНУ КВАРТИРЫ

d_1 – м^2

d_2 – сотен метров до метро

d_3 – средняя стоимость в соседних домах

$$a^*(x) = 50000 \cdot d_1 - 35000 \cdot d_2 + 0.5 \cdot d_3$$

Неужели квадратура самый важный признак?

Как, вообще говоря, их сравнивать между собой?

Пока можем только интерпретировать знак.

РЕЗЮМЕ

- Поняли, что признаки нужно масштабировать
- Например, с помощью min-max или standard scaler'ов
- Это позволяет привести признаки примерно к одному порядку
- И свободно использовать регуляризаторы
- Не бояться, что в одних параметрах мы улучшимся, а в других - наоборот
- Хотя параметр λ все еще придется выбирать и экспериментировать с ним
- Легче понимать, какой признак важнее!

РЕЗЮМЕ

- Теперь, когда мы наблюдаем переобучение, можно побороться с ним следующим образом:
- Промасштабировать признаки (привести к одному порядку)
- Регуляризировать модель
- Определиться с удовлетворительным параметром λ
- Сравнивить значения коэффициентов между собой и сделать выводы
- Наслаждаться!

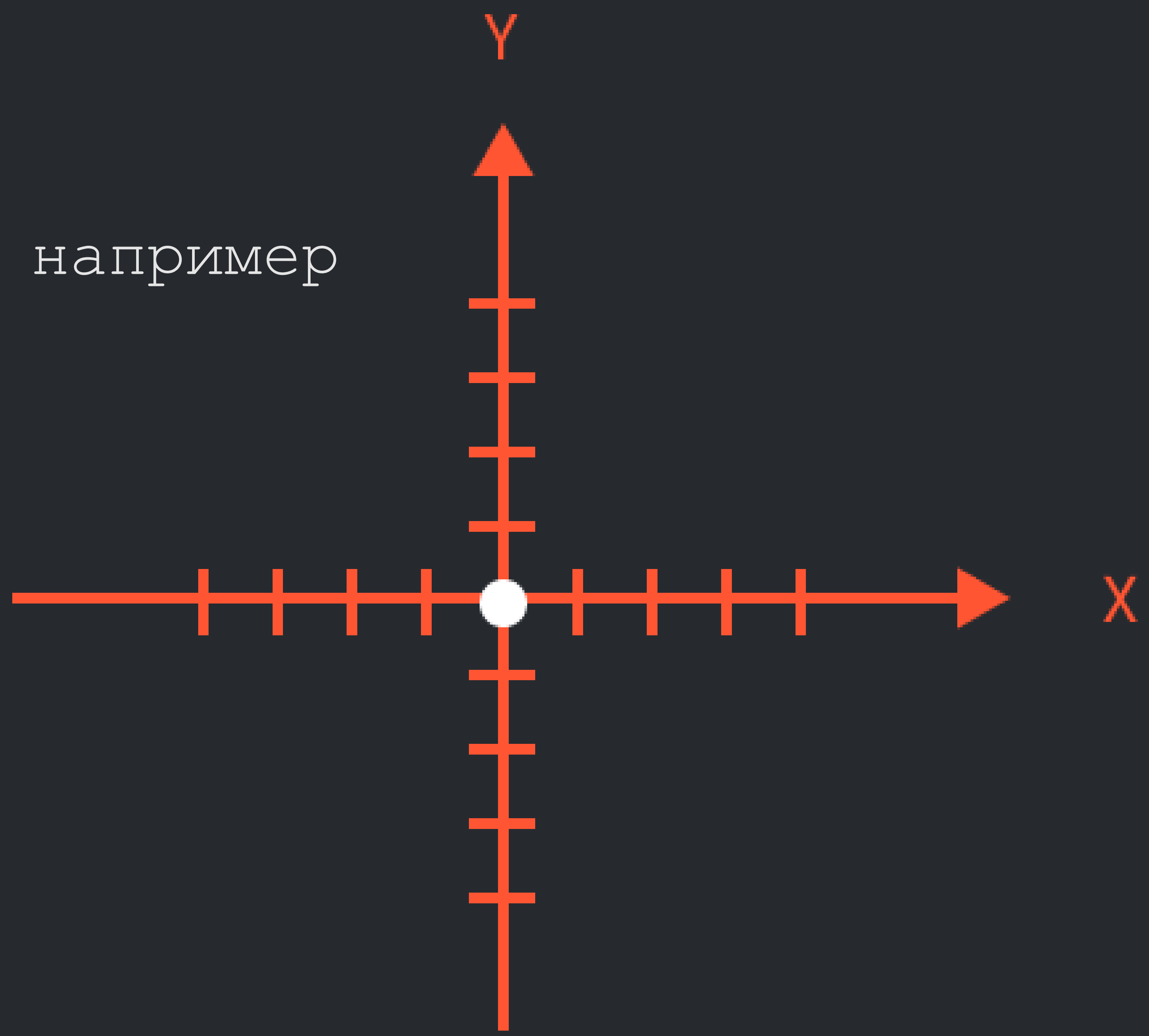
- Бонус: а давайте посмотрим, как регуляризация выглядит графически?

ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Пусть имеем $z(x, y) = x^2 + y^2 \to min$

— Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$

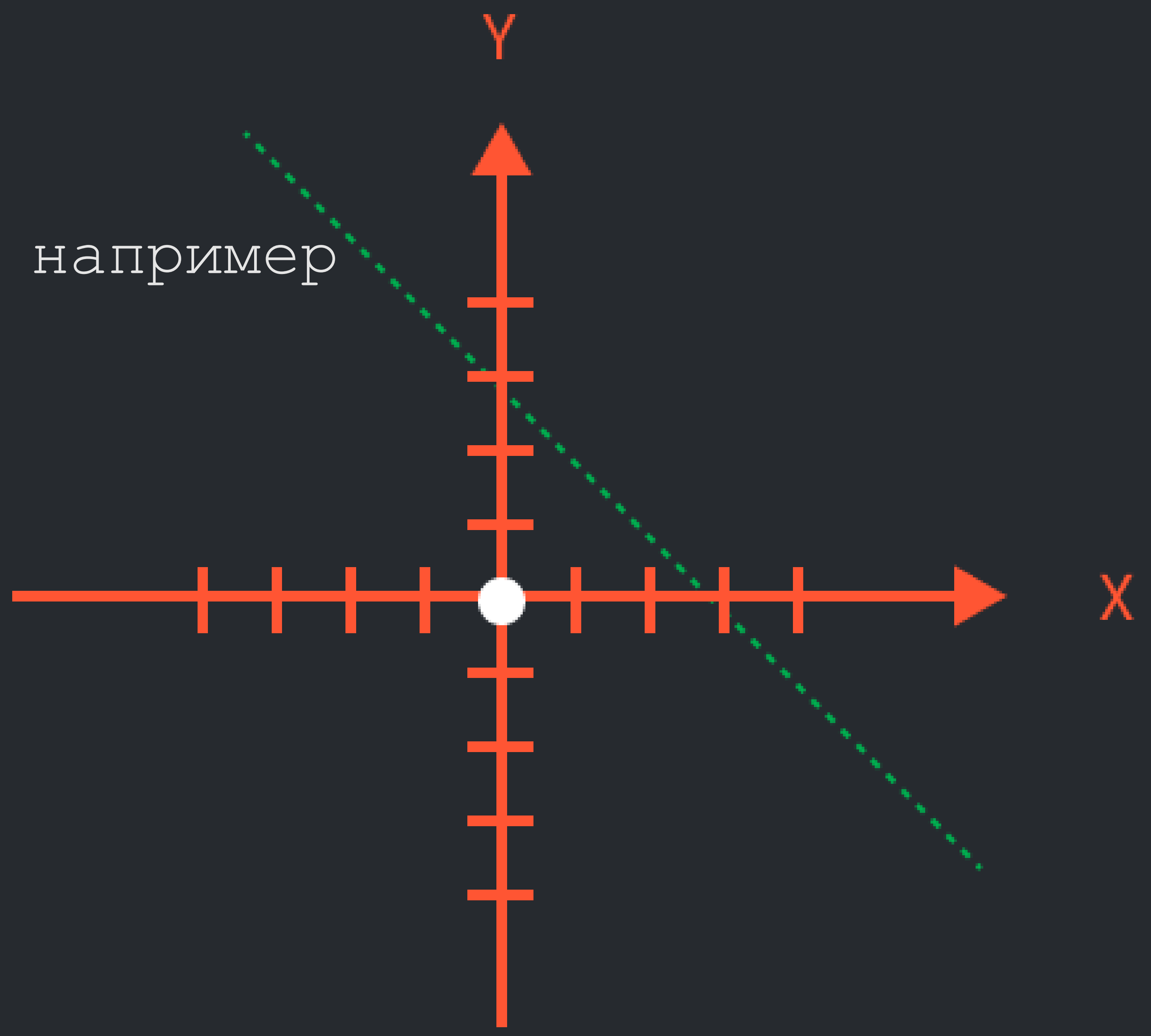


ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Пусть имеем $z(x, y) = x^2 + y^2 \to min$

— Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$



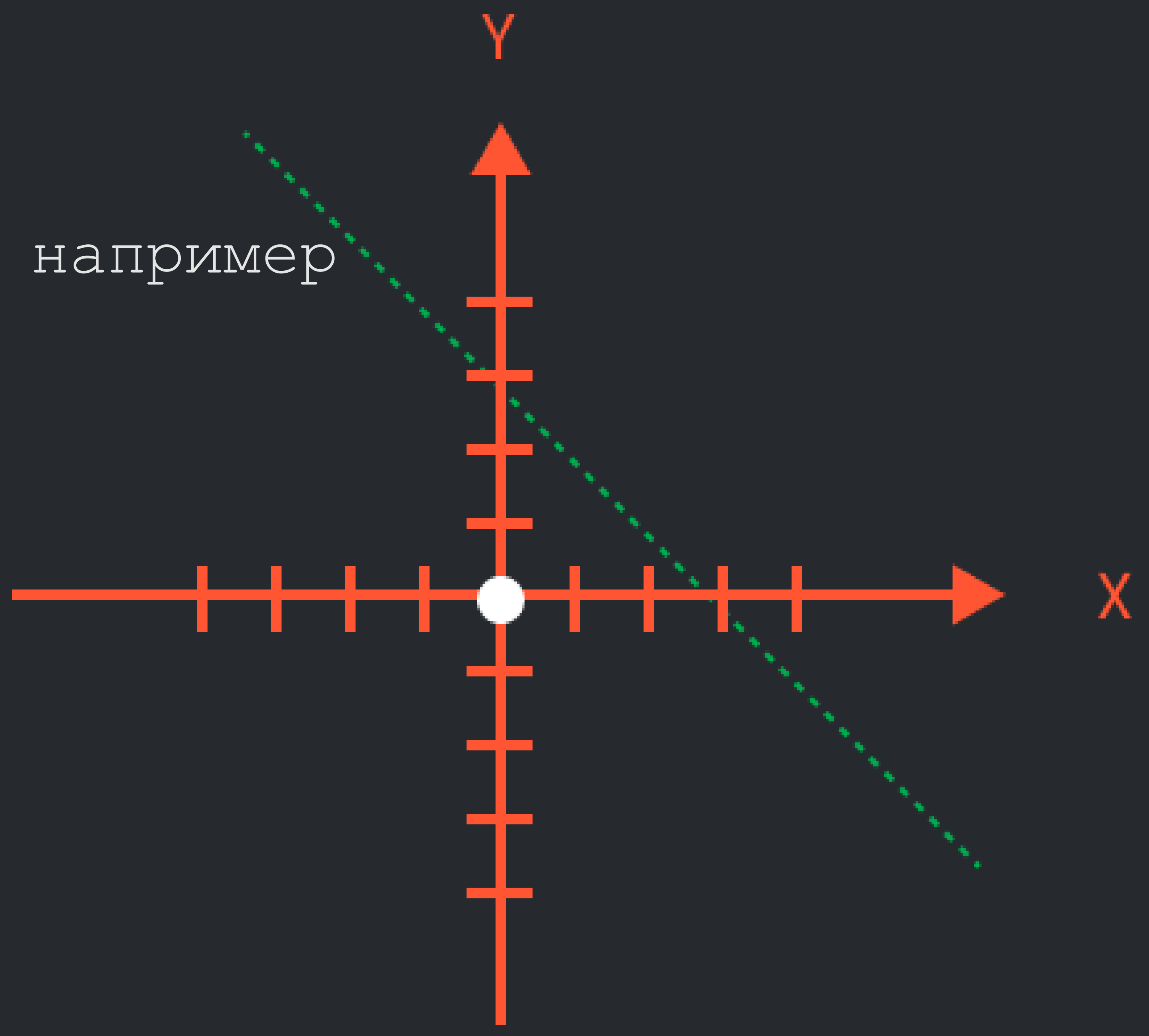
ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Пусть имеем $z(x, y) = x^2 + y^2 \rightarrow min$

— Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$

— $C = 25$: $x^2 + y^2 = 25$

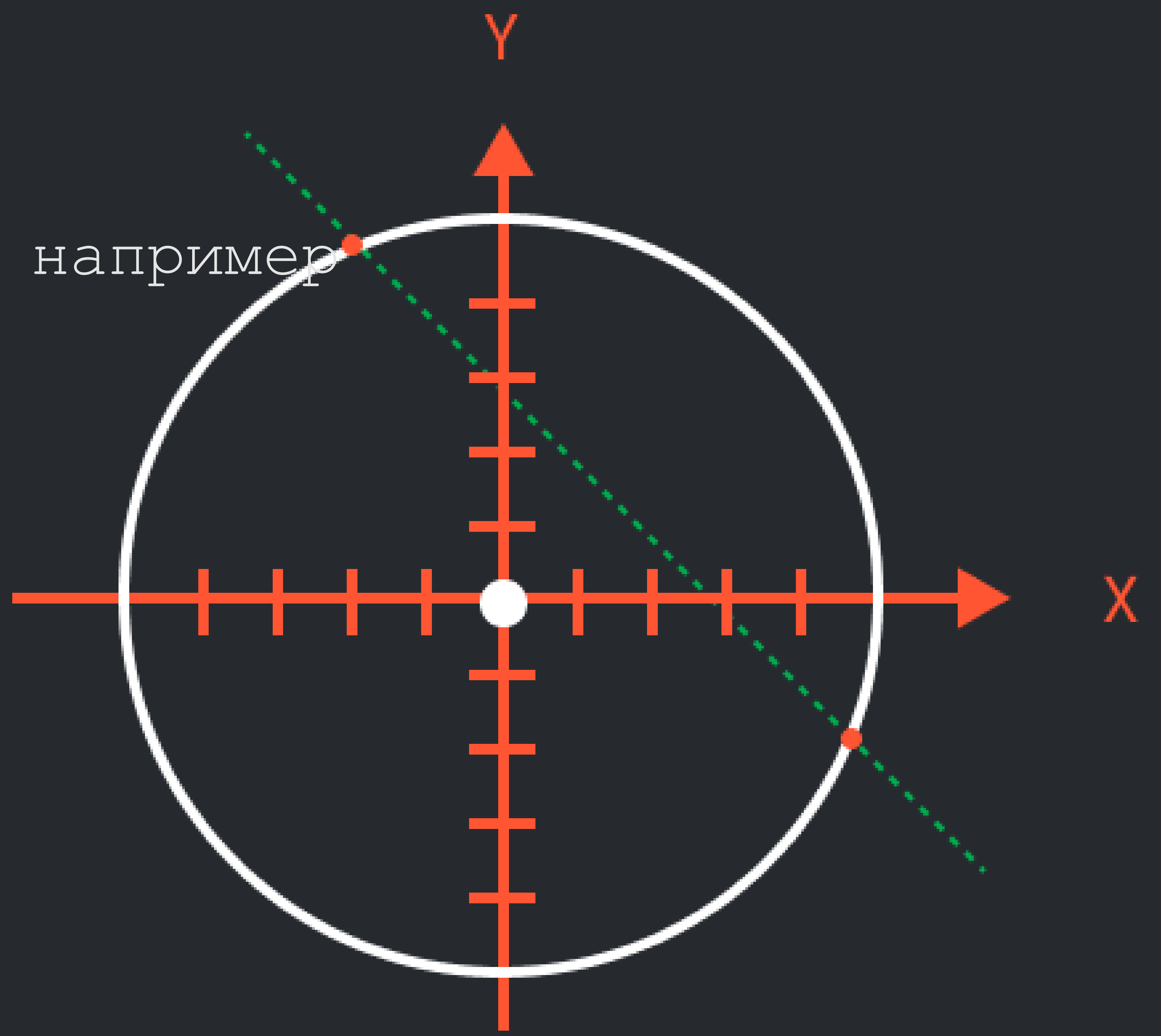


ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

- Пусть имеем $z(x, y) = x^2 + y^2 \to min$
- Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$

- $C = 25$: $x^2 + y^2 = 25$
- $C = 16$: $x^2 + y^2 = 16$



ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Пусть имеем $z(x,y) = x^2 + y^2 \to min$

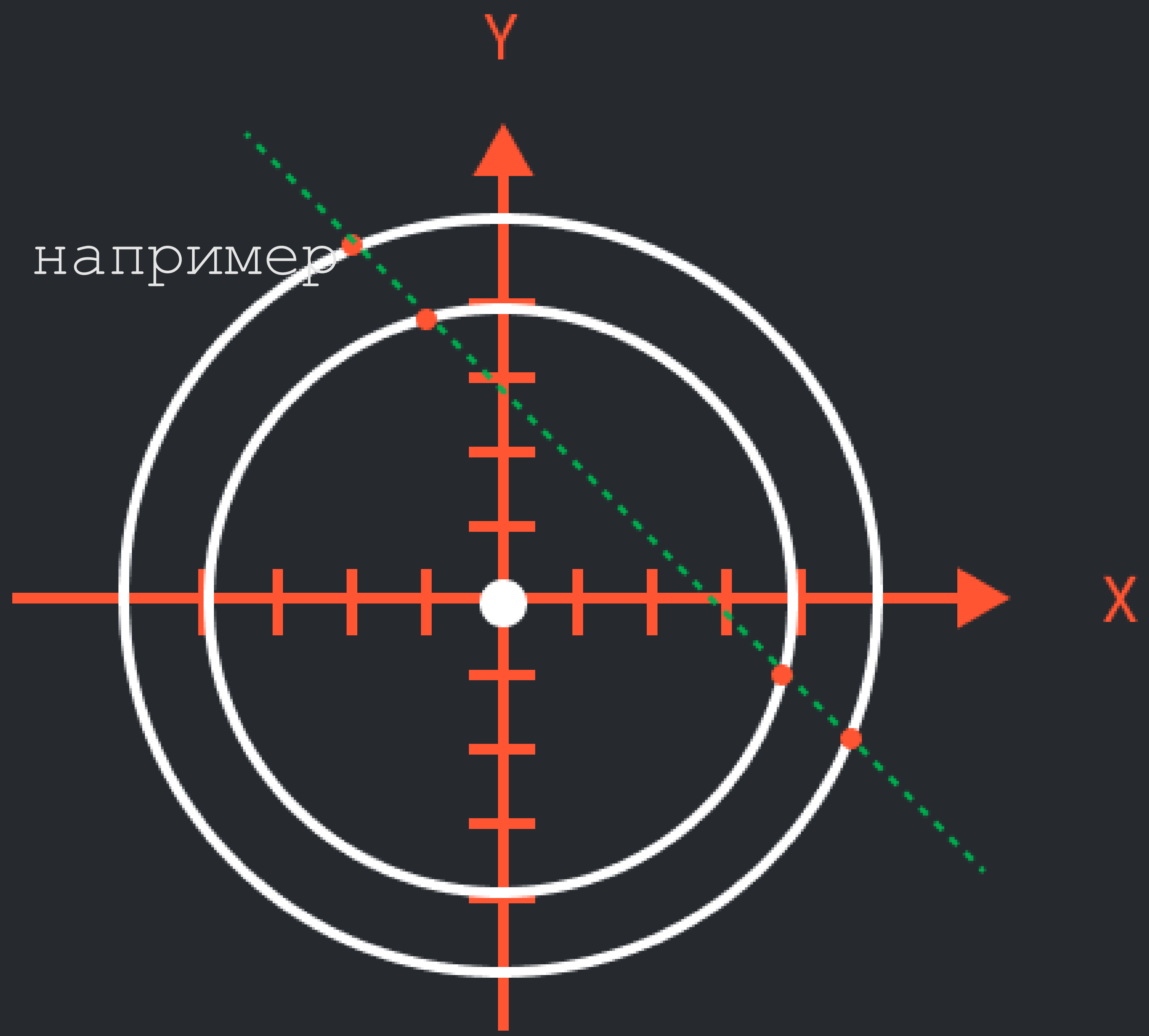
— Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$

— $C = 25$: $x^2 + y^2 = 25$

— $C = 16$: $x^2 + y^2 = 16$

— $C = 4$: $x^2 + y^2 = 4$



ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Пусть имеем $z(x, y) = x^2 + y^2 \to min$

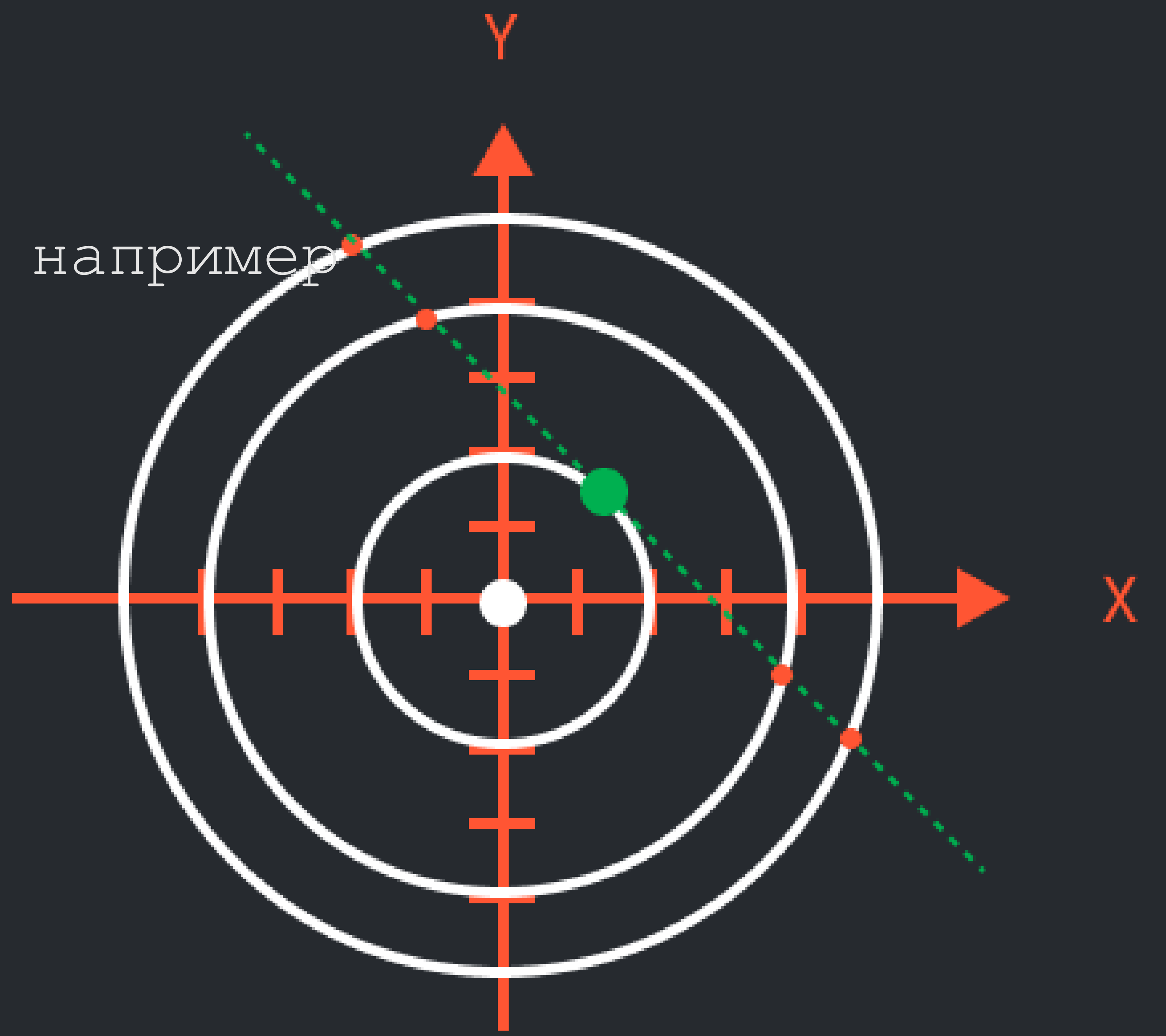
— Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$

— $C = 25$: $x^2 + y^2 = 25$

— $C = 16$: $x^2 + y^2 = 16$

— $C = 4$: $x^2 + y^2 = 4$



ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Пусть имеем $z(x, y) = x^2 + y^2 \to min$

— Пусть так же есть некоторое ограничение, например

$$x + y - 2 \cdot \sqrt{2} = 0$$

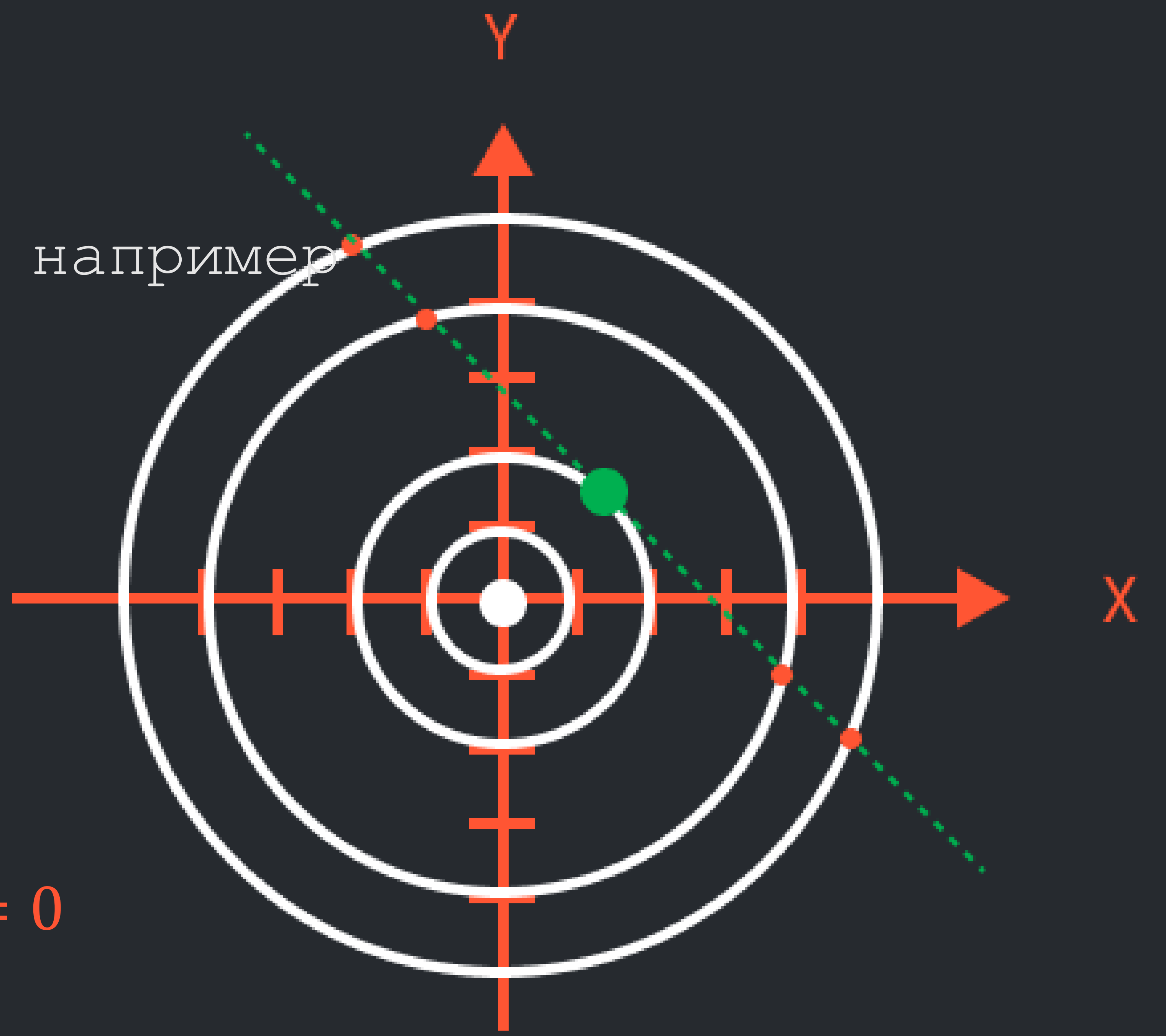
— $C = 25$: $x^2 + y^2 = 25$

— $C = 16$: $x^2 + y^2 = 16$

— $C = 4$: $x^2 + y^2 = 4$

— Ранее имели безусловный экстремум $z(0; 0) = 0$

— Теперь условный экстремум $z\left(\sqrt{2}, \sqrt{2}\right) = 4$



ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

— Задачу вида

$$Q(a(x,\beta),X) + \lambda \cdot R(\beta) \to \min_{\beta}$$

— Часто можно свести к

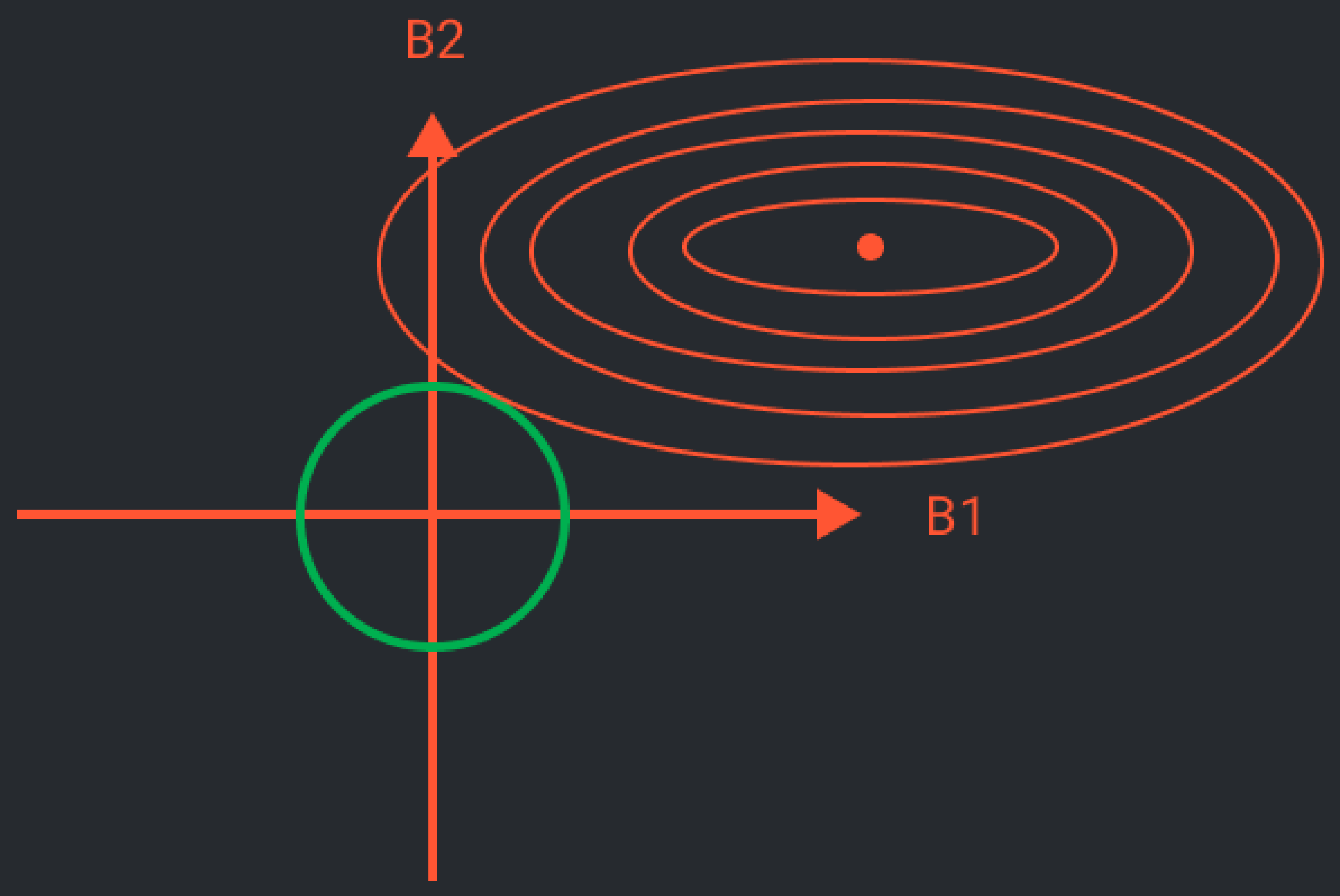
$$Q(a(x,\beta),X) \to \min_{\beta}$$

$$s.t.$$

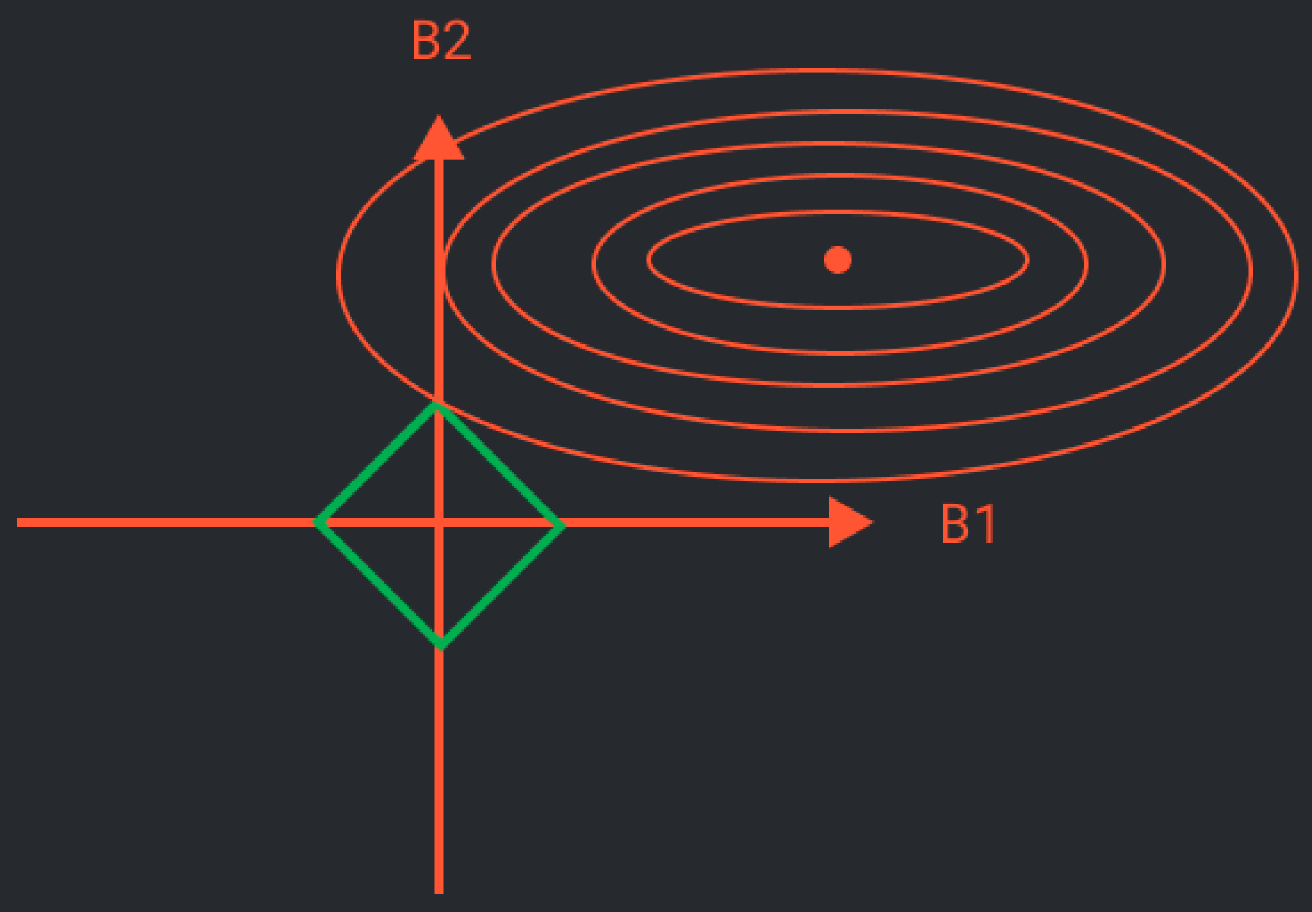
$$R(\beta) \leq \lambda$$

ЛИКБЕЗ №1: УСЛОВНЫЙ ЭКСТРЕМУМ

$$Q(a(x,\beta),X) \to \min_{\beta}$$
$$s.t.$$
$$\beta_1^2 + \beta_2^2 \le \lambda$$



$$Q(a(x,\beta),X) \to \min_{\beta}$$
$$s.t.$$
$$|\beta_1| + |\beta_2| \le \lambda$$



РЕЗЮМЕ

- Узнали, что регуляризированный случай можно свести к задаче условного экстремума
- Поняли, в чем состоит ключевое отличие Lasso и Ridge
- Кажется, теперь мы гуру по борьбе с переобучением!

МУЛЬТИКОЛЛИНЕАРНОСТЬ

ЧТО ЭТО?

Это наличие линейной зависимости между объясняющими переменными.

Приводит к построению достаточно неустойчивой модели, есть шанс переобучиться.

В случае с OLS-регрессией гарантирует отсутствие единственного минимума.

Невозможно найти обратную матрицу в формуле $\beta^* = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$

ЛИКБЕЗ №2: ЛИНЕЙНАЯ ЗАВИСИМОСТЬ

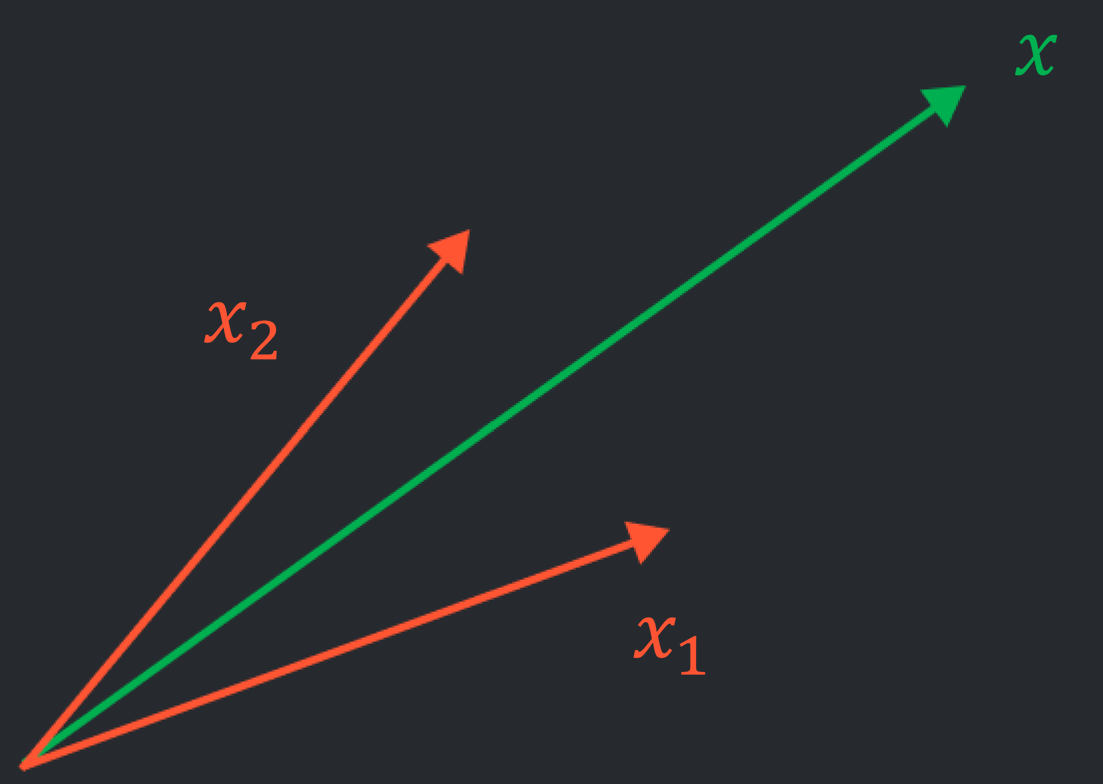
Система векторов $\{x_1, x_2, x_3 \dots\}$ называется ЛЗ, если хотя бы 1 вектор можно выразить как линейную комбинацию остальных

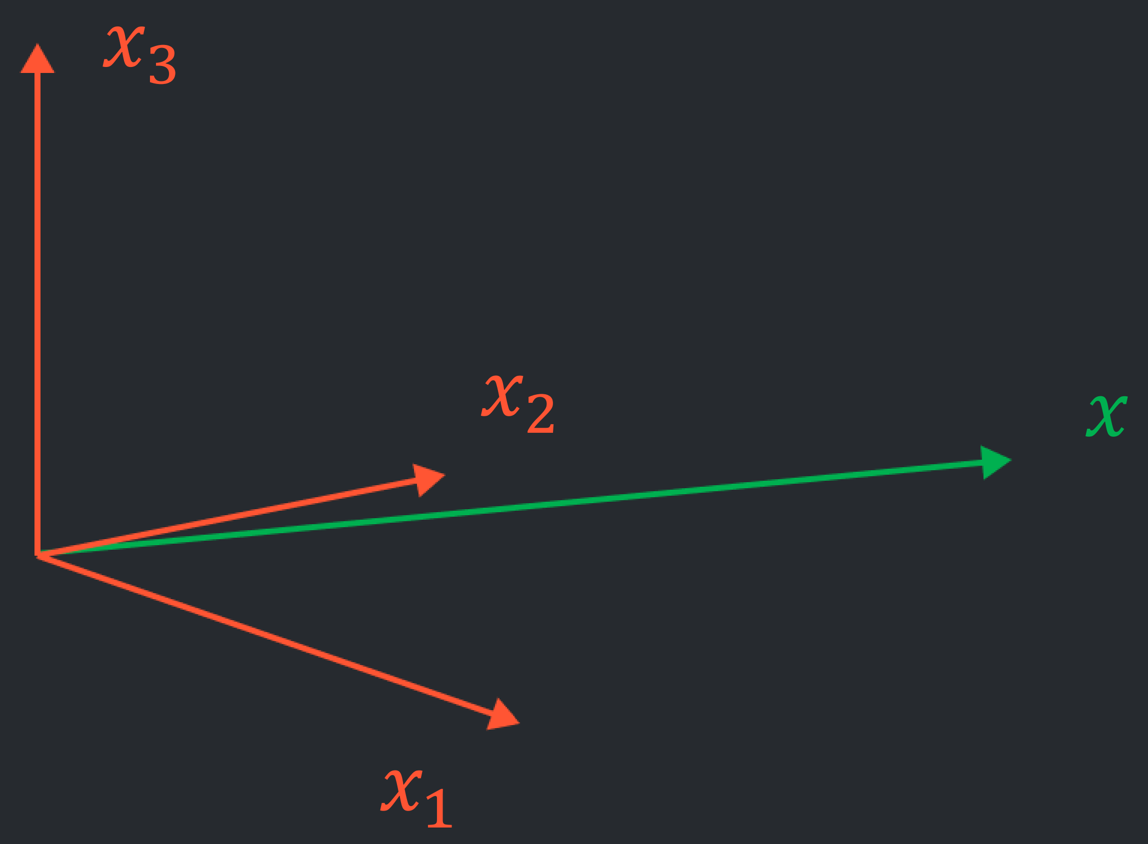
$x_1 = (3 \;\; 1), \quad x_2 = (2 \;\; 2)$

$x = 1 \cdot x_1 + 1 \cdot x_2 = (5 \;\; 3)$

$x_1 = (1 \;\; 0 \;\; 0), \quad x_2 = (0 \;\; 1 \;\; 0), \quad x_3 = (0 \;\; 0 \;\; 1)$

$x = 3 \cdot x_1 + 0.5 \cdot x_2 + 0 \cdot x_3 = (3 \;\; 0.5 \;\; 0)$





МУЛЬТИКОЛЛИНЕАРНОСТЬ

ПРИМЕР:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} 23 & 46 \\ 35 & 70 \\ 18 & 36 \end{pmatrix} \end{array}$$

Видно, что вектора из признаков Линейно Зависимы:

$$d_2 = 2 \cdot d_1$$

Проблема мультиколлинеарности!

Линейная регрессия либо сломается (формула с матрицами)

Либо попадем в абы какую точку градиентными методами

МУЛЬТИКОЛЛИНЕАРНОСТЬ

ПРИМЕР:

$$
\begin{array}{c c}
 & \begin{array}{cc} d_1 & d_2 \end{array} \\
\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} & \begin{pmatrix} 23 & 46 \\ 35 & 70 \\ 18 & 36 \end{pmatrix}
\end{array}
$$

По факту, это значит, что в наших данных лишняя информация

Если знаем d_1, то и d_2 легко восстановить сможем

А тогда d_2 лучше просто убрать.

Поэтому мы на первом занятии чистили колонки после One-Hot Encoding метода!

МУЛЬТИКОЛЛИНЕАРНОСТЬ

ПРИМЕР:

$$
\begin{array}{c}
 \\ x_1 \\ x_2 \\ x_3
\end{array}
\begin{array}{c}
\text{Валюта} \\
\begin{pmatrix} \text{USD} \\ \text{EUR} \\ \text{RUB} \end{pmatrix}
\end{array}
\rightarrow
\begin{array}{c}
\begin{array}{ccc} USD? & EUR? & RUB? \end{array} \\
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}
\end{array}
\rightarrow
\begin{array}{c}
\begin{array}{cc} USD? & EUR? \end{array} \\
\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}
\end{array}
$$

МУЛЬТИКОЛЛИНЕАРНОСТЬ

РЕГУЛЯРИЗАЦИЯ ПОМОГАЕТ БОРОТЬСЯ С МУЛЬТИКОЛЛИНЕАРНОСТЬЮ

Регуляризация склонна обнулять веса у плохих признаков в модели.

Тогда и у какого-то линейно зависимого от остальных признака такой коэффициент тоже с большой вероятностью обнулится!

$$a(x) = 100 \cdot d_1 + 300 \cdot d_2 + 0 \cdot d_3$$

РЕЗЮМЕ

- Узнали, что такое мультиколлинеарность
- Это достаточно сложная проблема!
- Она ломает как формулу с матрицами
- Так и усложняет градиентный спуск

- Нужно либо чистить данные (подробнее узнаем позже!)
- Либо использовать регуляризацию
- Пора к практике!

СПАСИБО

ТАБАКАЕВ НИКИТА