START ML

KARPOV.COURSES

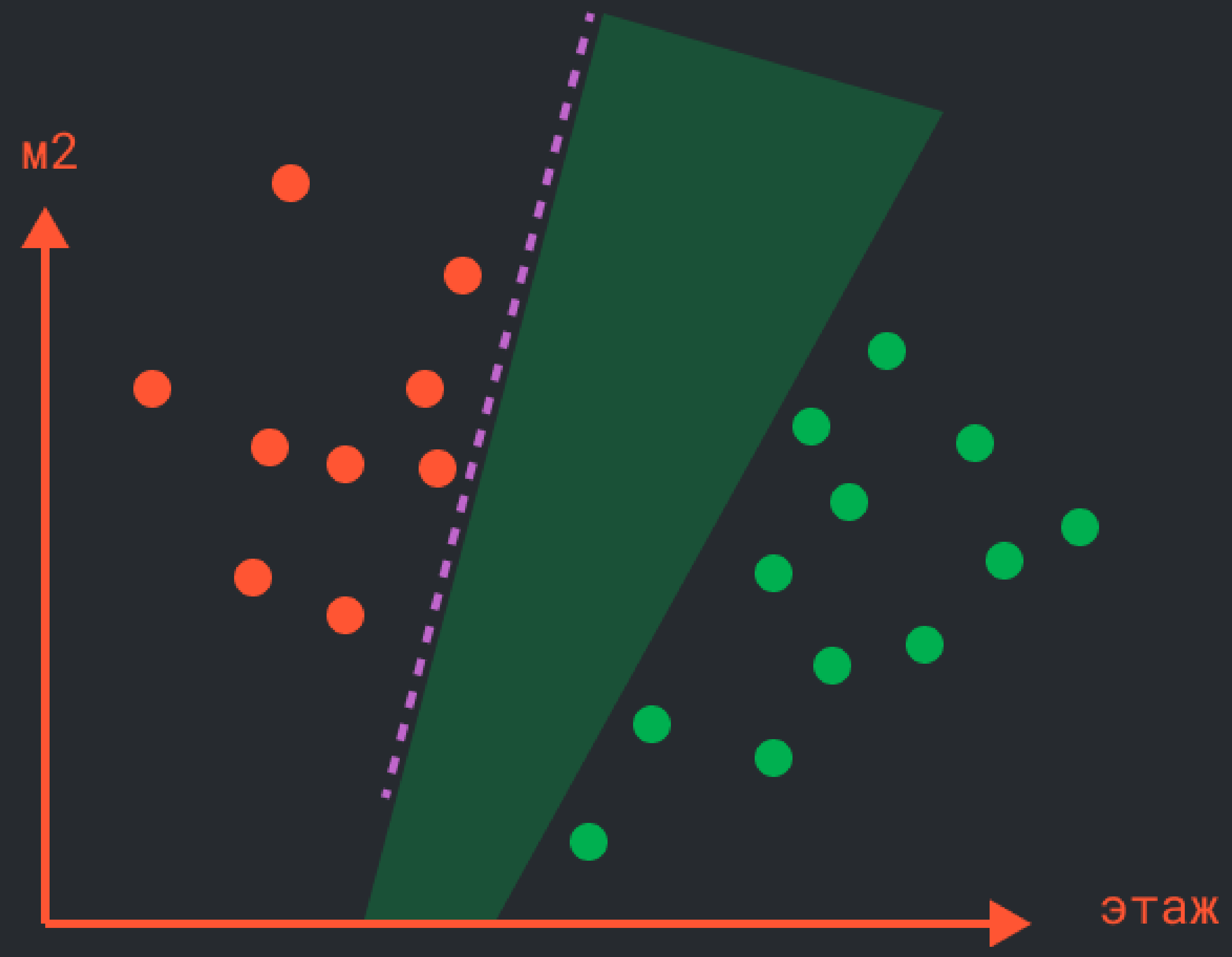
БИНАРНАЯ КЛАССИФИКАЦИЯ

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

— $\beta^* = argmin \sum_i^n L(M_i) = \sum_i^n \log(1 + e^{-y_i \cdot \langle \beta, x_i \rangle})$

— $P(y_i = +1 | x_i) = \frac{1}{1+e^{-\langle \beta, x_i \rangle}}$

— Получаем модель с корректной оценкой вероятности!

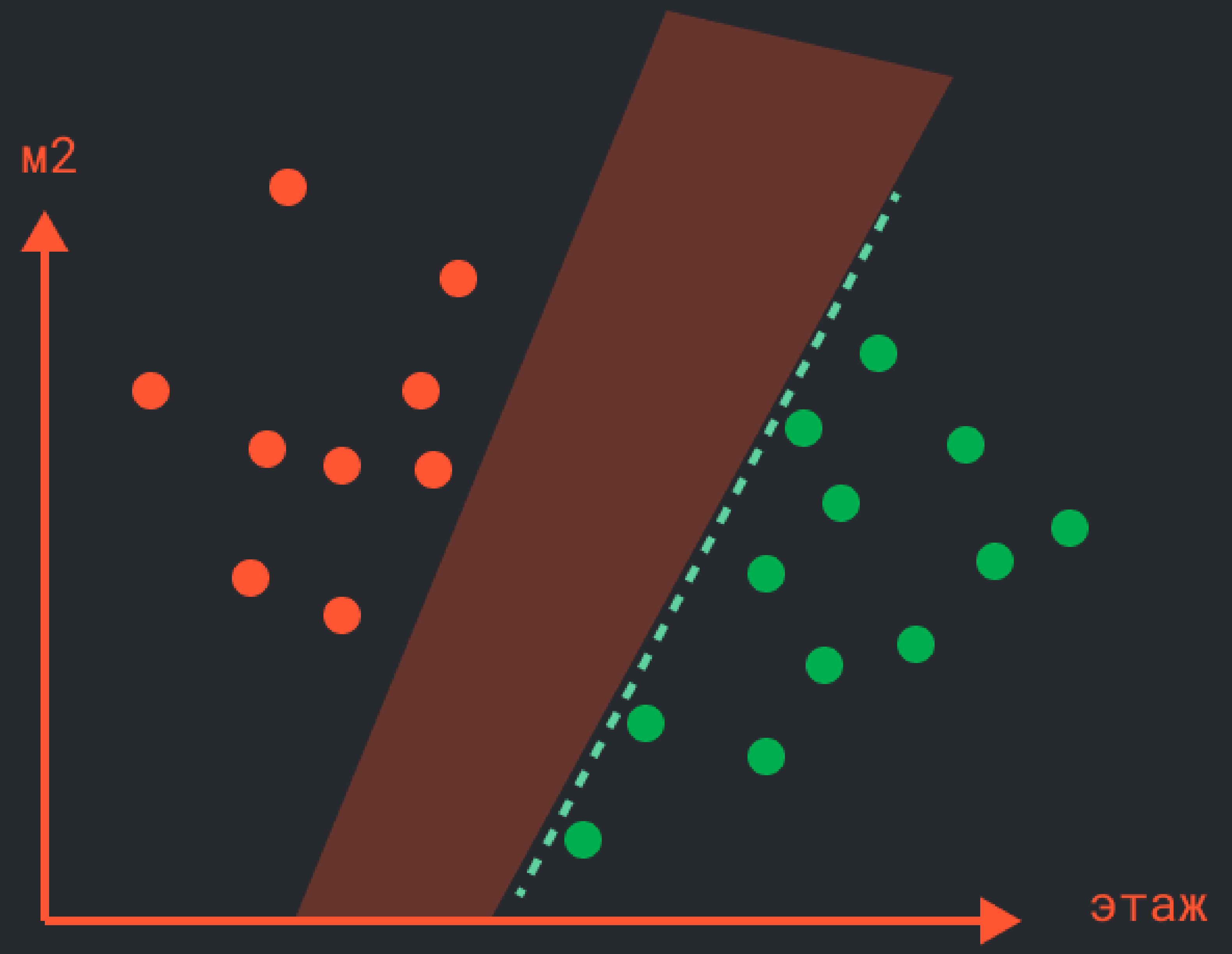


БИНАРНАЯ КЛАССИФИКАЦИЯ

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- $\beta^* = argmin \sum_i^n L(M_i) = \sum_i^n \log(1 + e^{-y_i \cdot \langle \beta, x_i \rangle})$
- $P(y_i = +1 | x_i) = \frac{1}{1+e^{-\langle \beta, x_i \rangle}}$

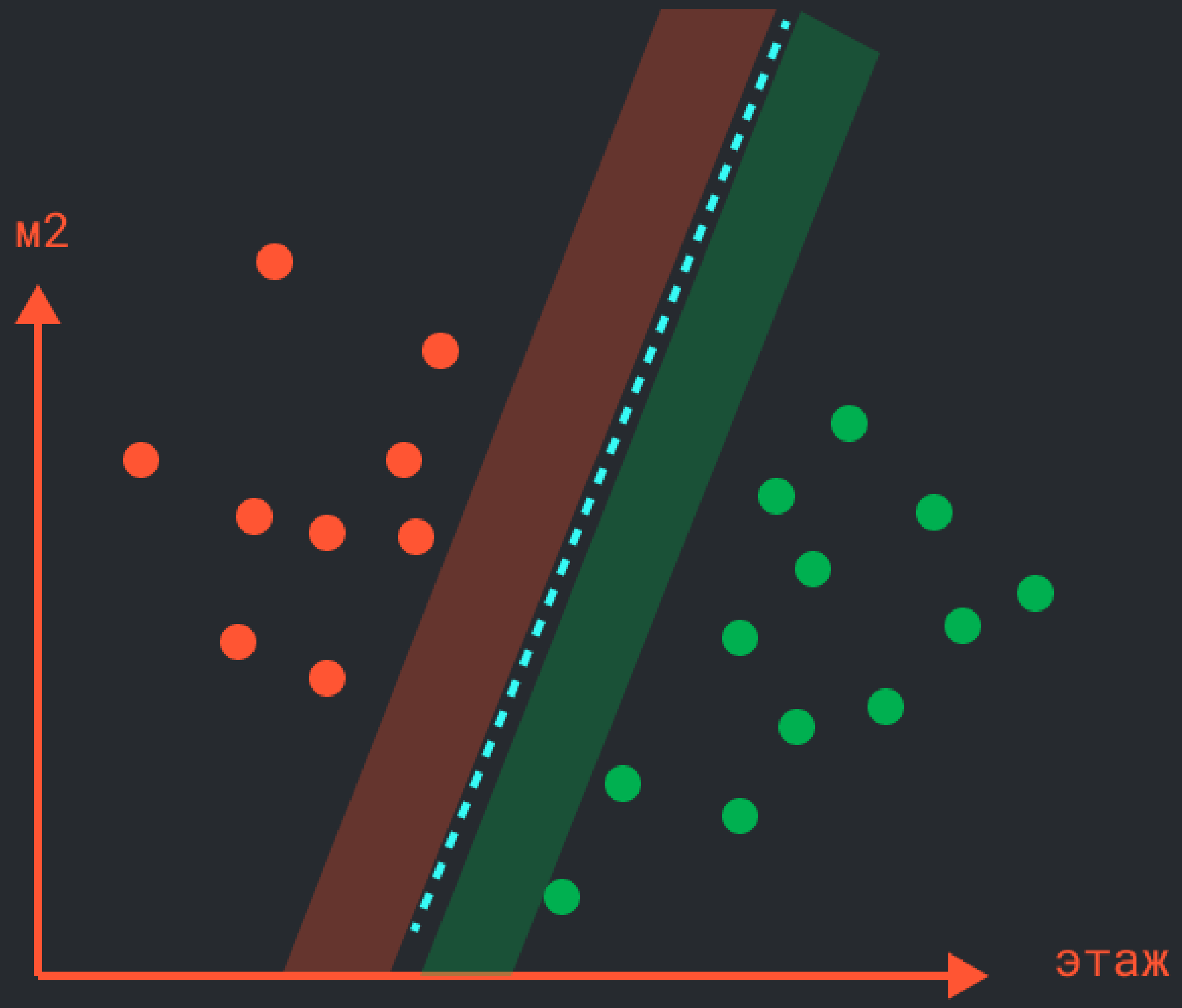
- Получаем модель с корректной оценкой вероятности!



БИНАРНАЯ КЛАССИФИКАЦИЯ

SVM

- *Support Vector Machine* – еще один из способов построить модель классификации
- Достаточно популярный механизм. В том числе, частенько используется в NLP
- Имеет больший геометрический смысл
- Можно показать, что является частным случаем того, что мы



БИНАРНАЯ КЛАССИФИКАЦИЯ

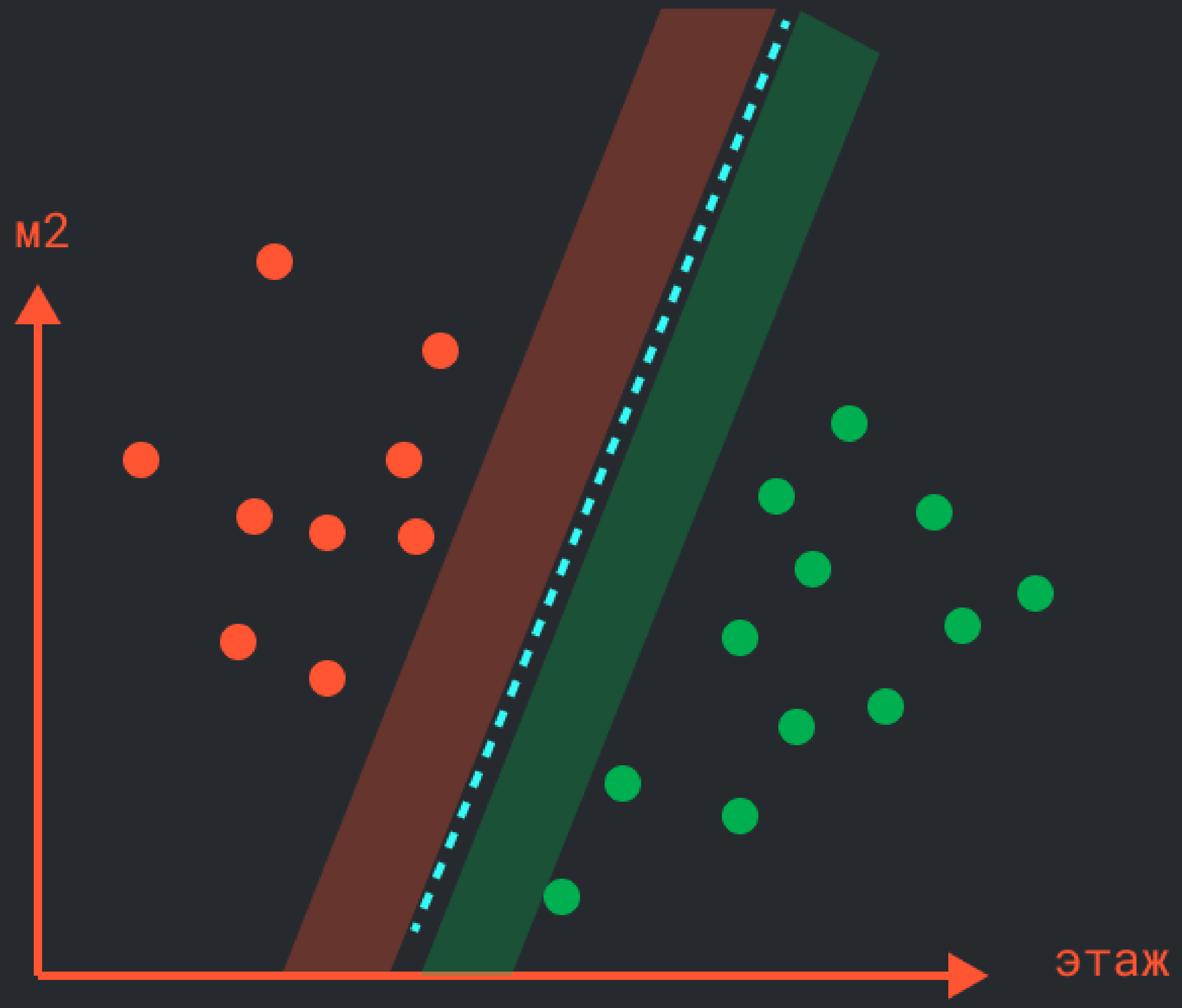
SVM

- Давайте максимизировать расстояние до ближайшего объекта!
- При этом учтем правильность разнесения разных классов по разным сторонам

$$\min_{x \in X} \; \rho(x_i, \beta) \to \max_{\beta}$$

$$s.t.$$

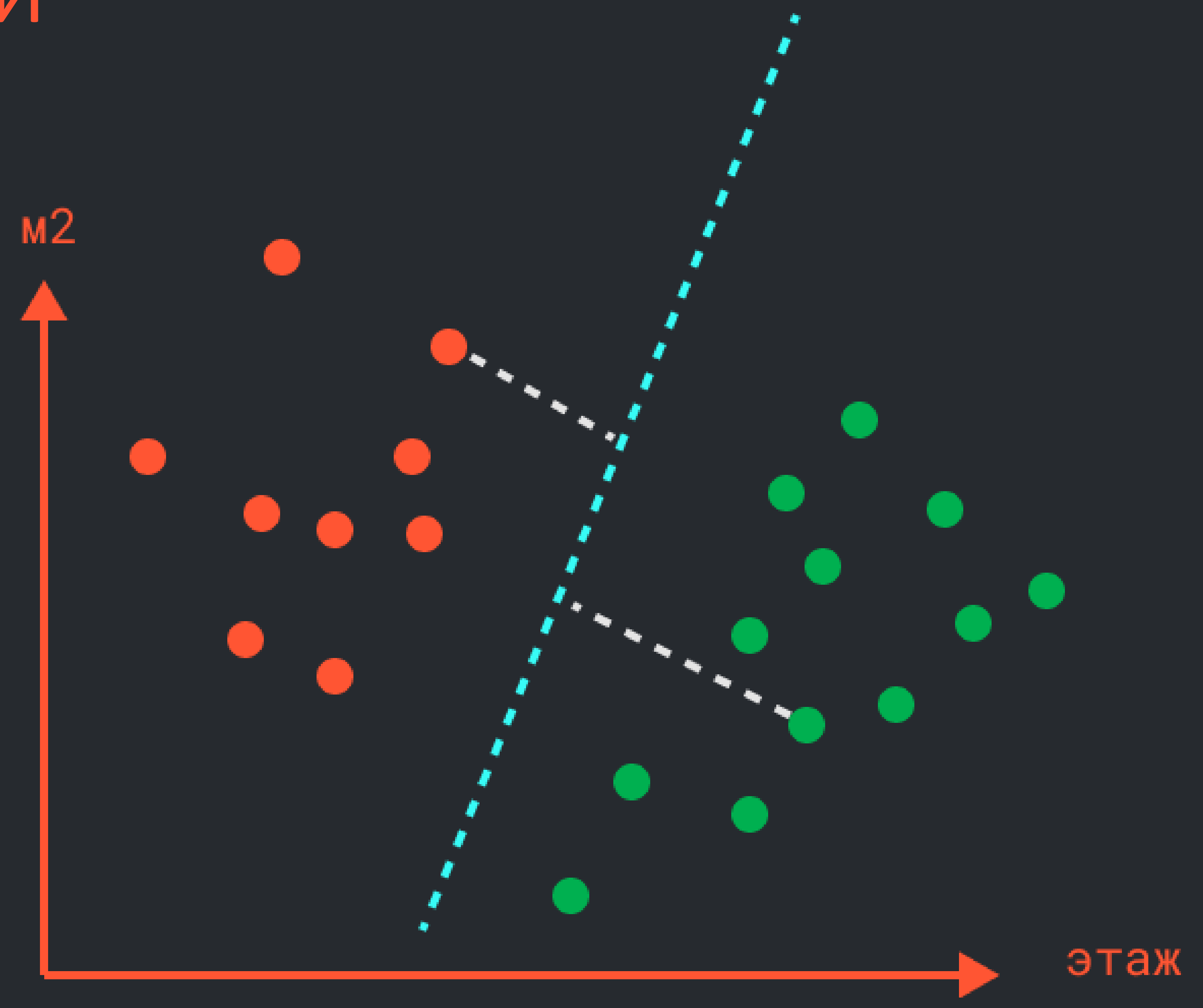
$$y_i \langle \beta, x_i \rangle \geq 0$$



ЛИКБЕЗ №1: РУБРИКА ЗАМЕЧАТЕЛЬНЫХ ФАКТОВ

РАССТОЯНИЕ ОТ ТОЧКИ ДО ПЛОСКОСТИ

$$\rho(x_i, \beta) = \frac{|\langle \beta, x_i \rangle|}{|\beta|} = \frac{|\beta_1 \cdot d_1 + \dots + \beta_n \cdot d_n + \beta_0|}{\sqrt{\beta_1^2 + \dots + \beta_n^2})}$$



ЛИКБЕЗ №1: РУБРИКА ЗАМЕЧАТЕЛЬНЫХ ФАКТОВ

РАССТОЯНИЕ ОТ ТОЧКИ ДО ПЛОСКОСТИ

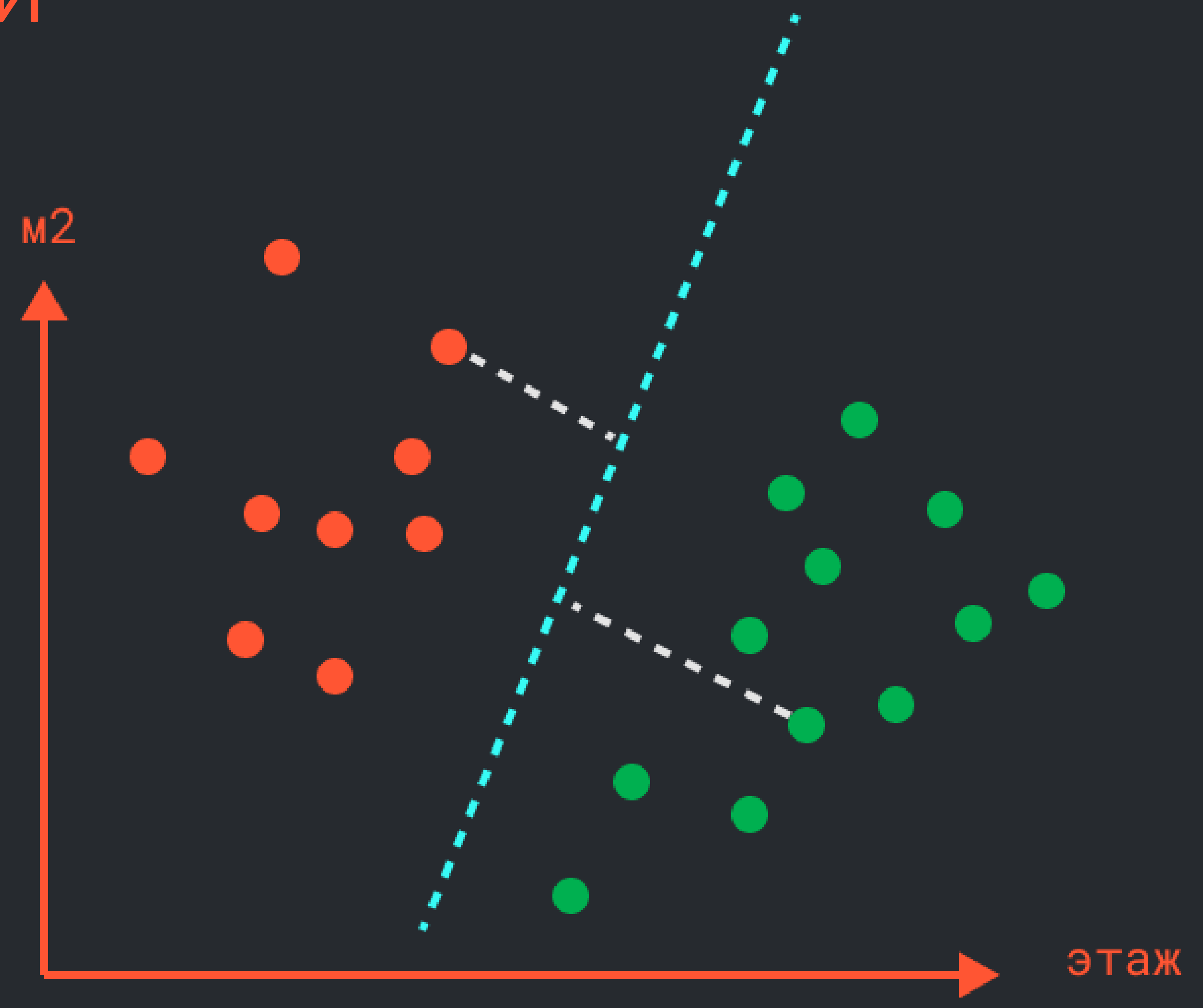
$$a(x) = sgn(1 \cdot \text{м}^2 + 1.5 \cdot \text{этаж} - 30)$$

$$x_1 = (10, \qquad 6)$$

$$x_2 = (19, \qquad 11)$$

$$\rho(x_1, \beta^*) = \frac{|1 \cdot 10 + 1.5 \cdot 6 - 30|}{\sqrt{1^2 + 1.5^2}} \approx 6.1$$

$$\rho(x_2, \beta^*) = \frac{|1 \cdot 19 + 1.5 \cdot 11 - 30|}{\sqrt{1^2 + 1.5^2}} \approx 3.05$$



ЛИКБЕЗ №1: РУБРИКА ЗАМЕЧАТЕЛЬНЫХ ФАКТОВ

ГИПЕРПЛОСКОСТЬ НЕЙТРАЛЬНА К УМНОЖЕНИЮ

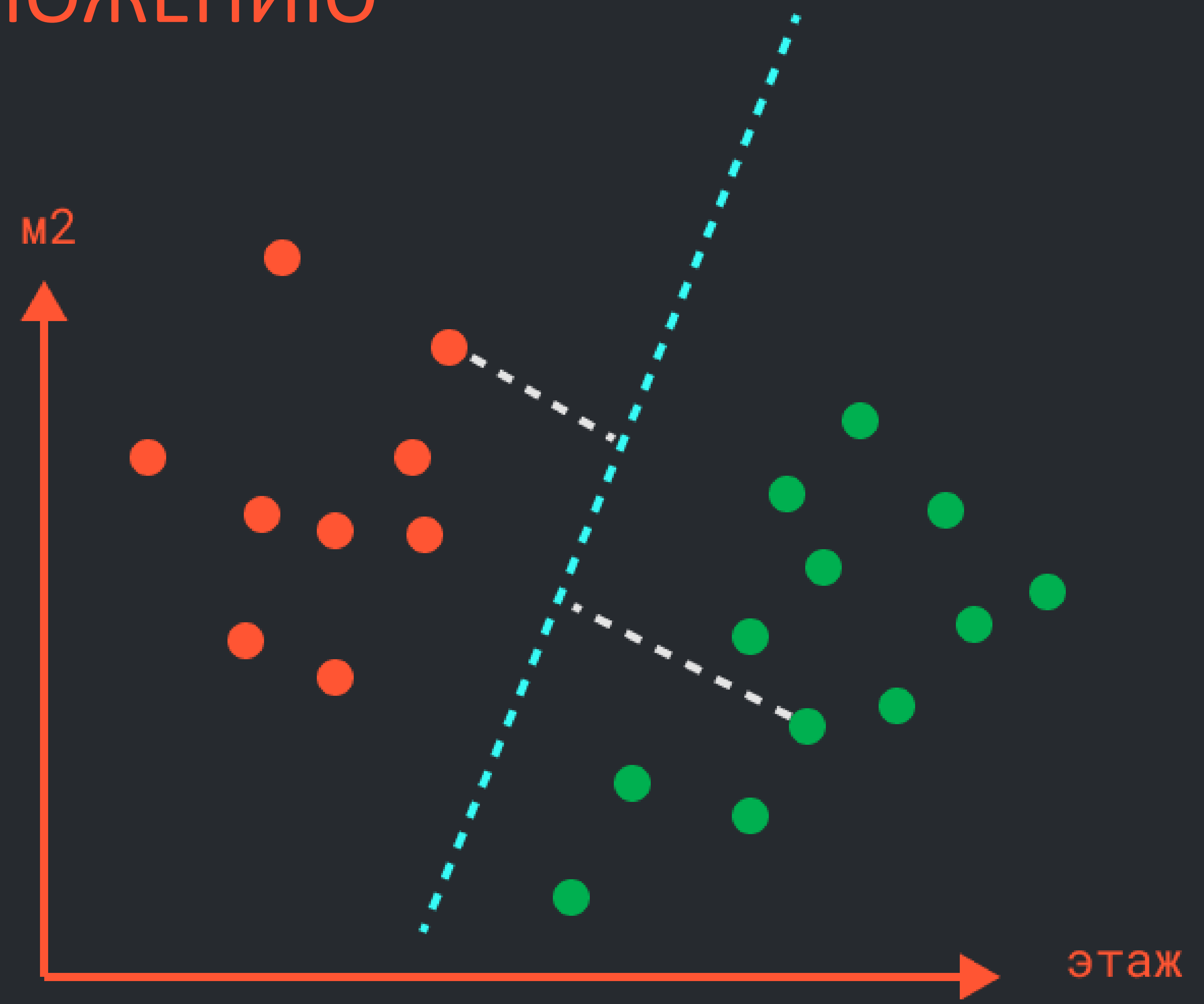
$$a(x) = sgn(1 \cdot \text{м}^2 + 1.5 \cdot \text{этаж} - 30)$$

Умножим все коэффициенты на одно число!

$$a^*(x) = sgn(2 \cdot \text{м}^2 + 3 \cdot \text{этаж} - 60)$$

$$\rho(x_1, \beta^*) = \frac{|2 \cdot 10 + 3 \cdot 6 - 60|}{\sqrt{2^2 + 3^2}} \approx 6.1$$

$$\rho(x_2, \beta^*) = \frac{|2 \cdot 19 + 3 \cdot 11 - 60|}{\sqrt{2^2 + 3^2}} \approx 3.05$$



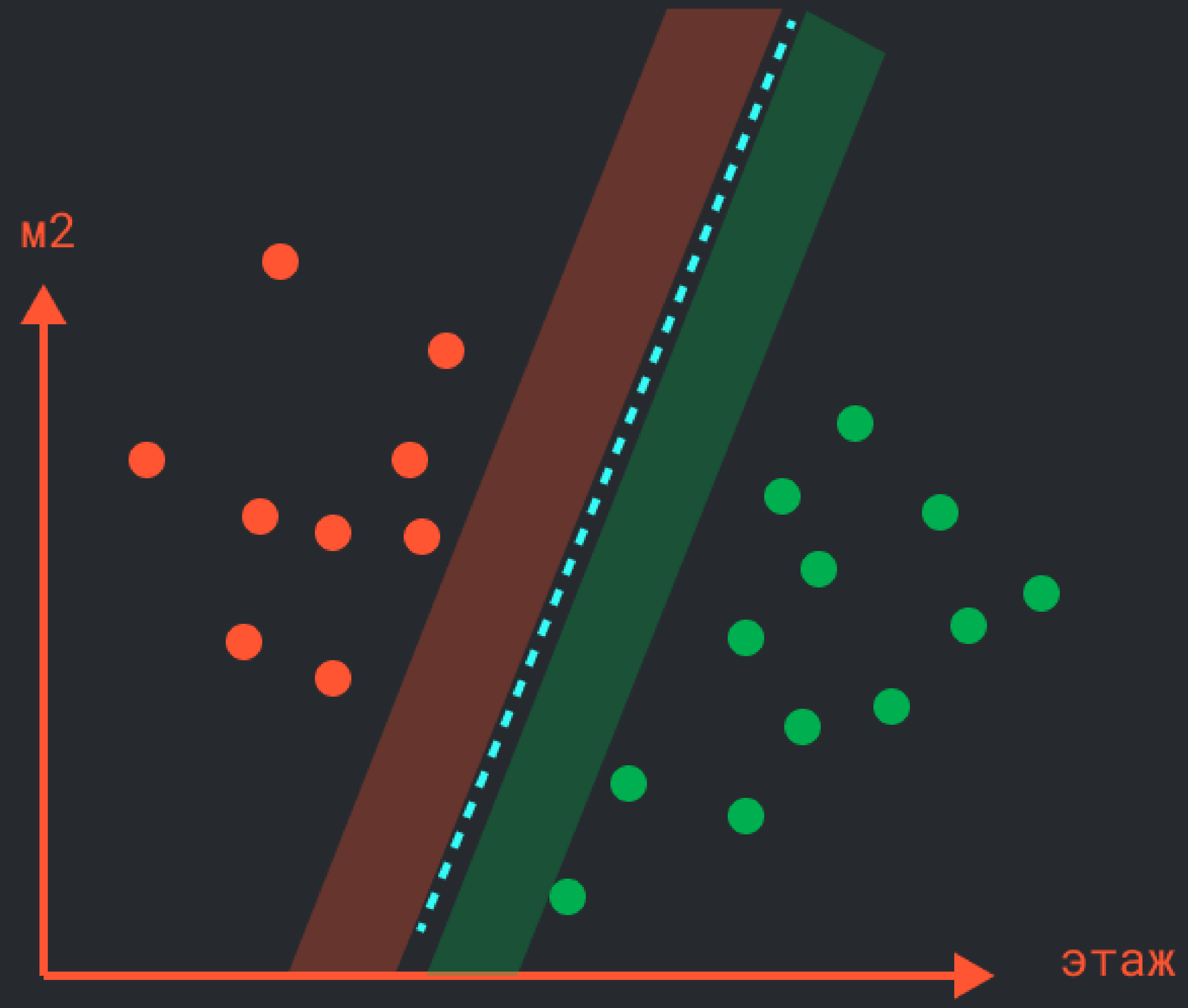
БИНАРНАЯ КЛАССИФИКАЦИЯ

SVM

$$\min_{x \in X} \quad \rho(x_i, \beta) \to \max_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 0$$



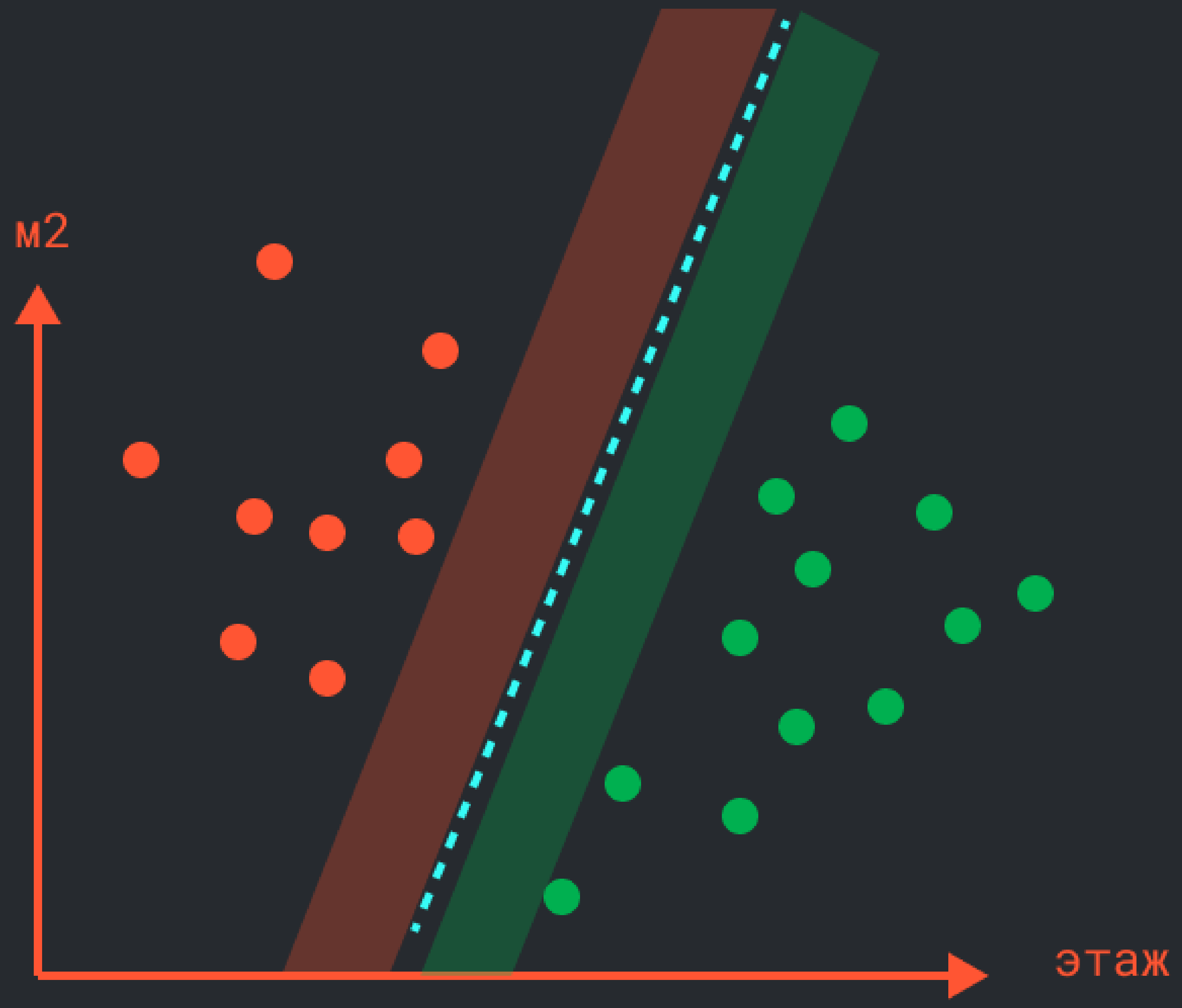
БИНАРНАЯ КЛАССИФИКАЦИЯ

SVM

$$\min_{x \in X} \frac{|\langle \beta, x \rangle|}{|\beta|} \to \max_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 0$$



Функция, которую оптимизируем, выглядит страшно! Что же, упростим ее! Договоримся:

$$\min_{x \in X} \langle \beta, X \rangle = 1$$

БИНАРНАЯ КЛАССИФИКАЦИЯ

SVM

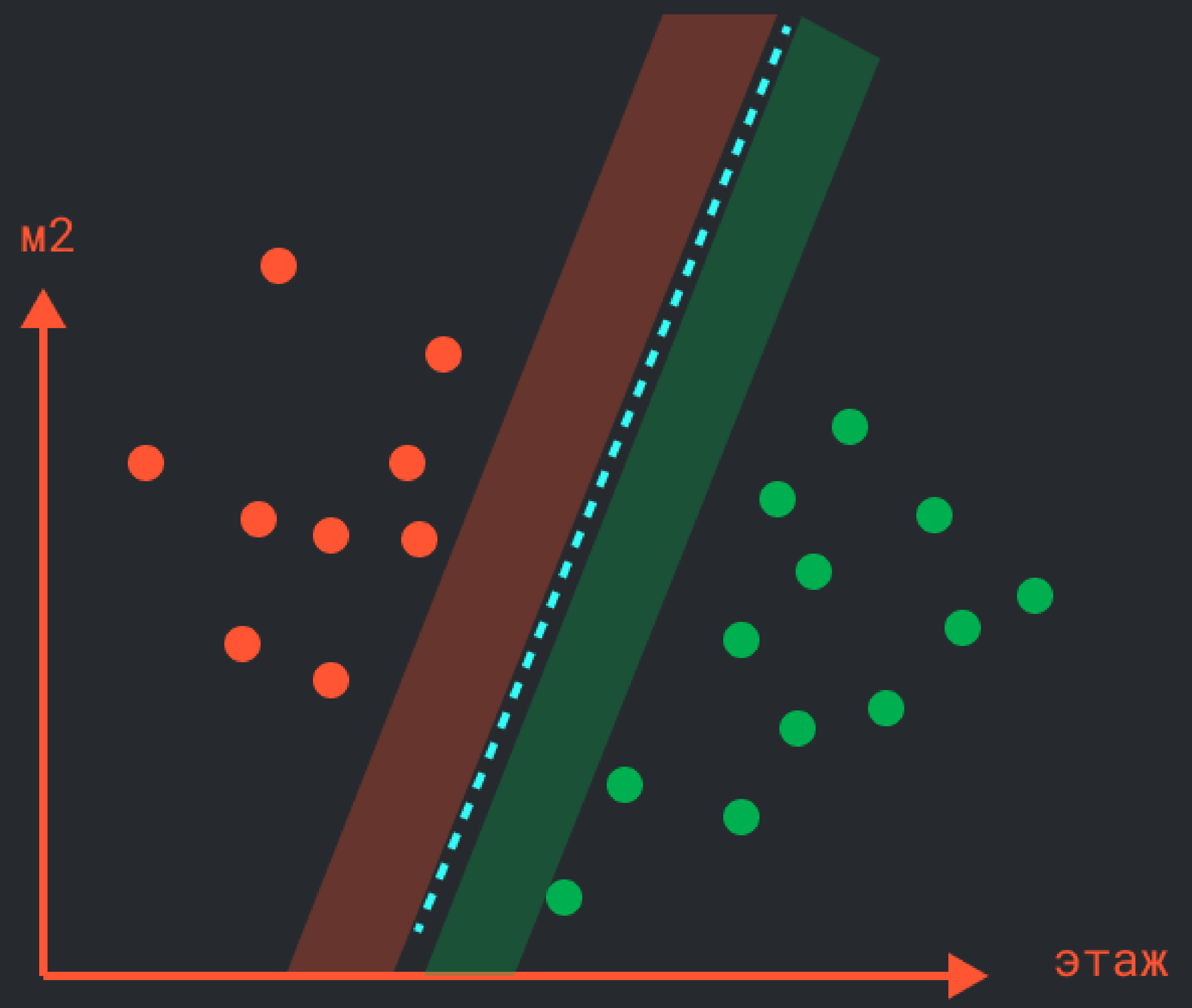
$$\min_{x \in X} \frac{|\langle \beta, x \rangle|}{|\beta|} \to \max_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 0$$

$$\min_{x \in X} \langle \beta, X \rangle = 1$$

Заметим, что $\min\limits_{x \in X} \frac{\langle \beta, x \rangle}{|\beta|} = \frac{\min_{x \in X} \langle \beta, x \rangle}{|\beta|} = \frac{1}{|\beta|}$



БИНАРНАЯ КЛАССИФИКАЦИЯ

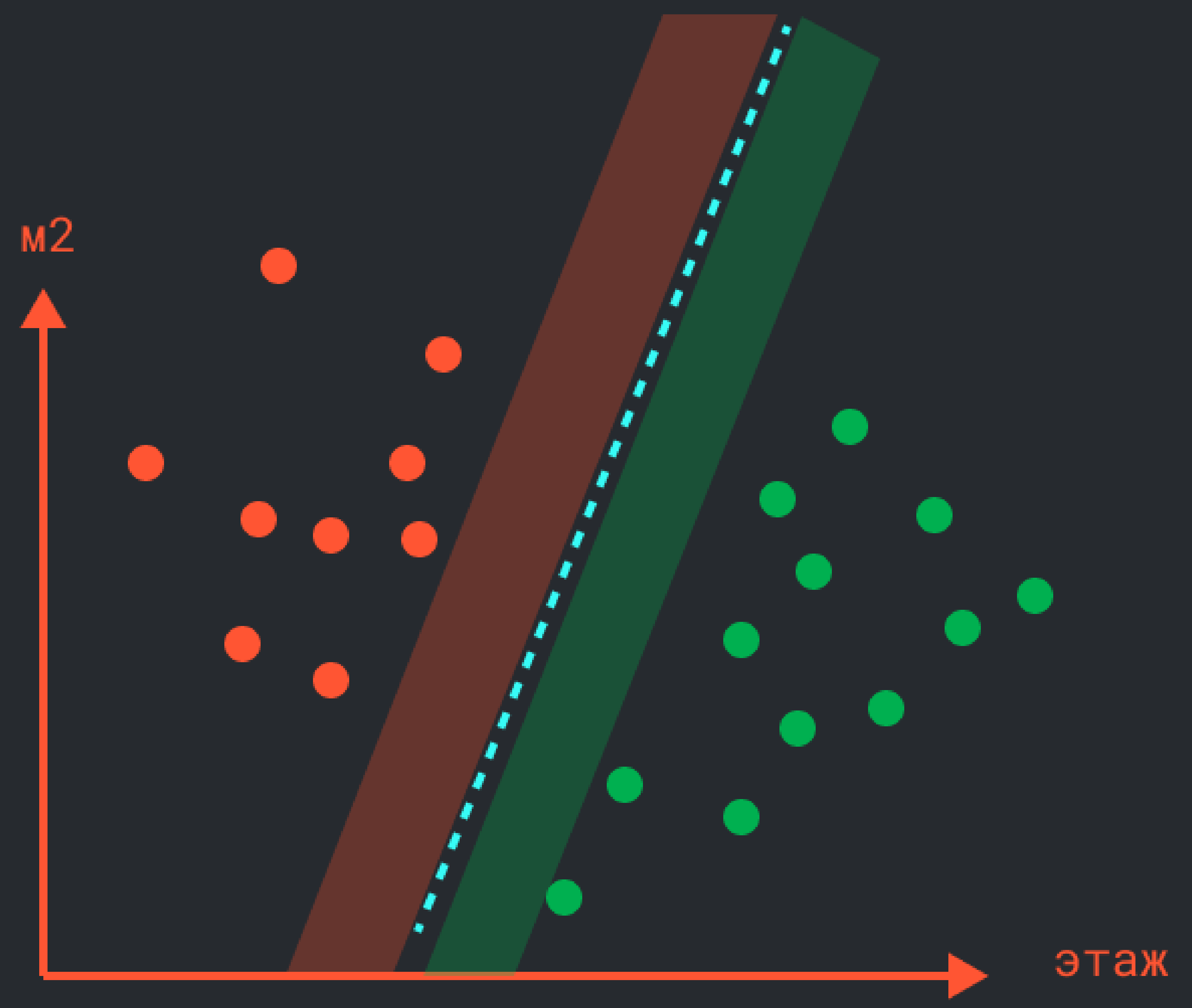
SVM

$$\frac{1}{|\beta|} \to \max_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 0$$

$$\min_{x \in X} \langle \beta, X \rangle = 1$$



Задачу оптимизации можно развернуть!

Чем больше $|\beta|$, тем меньше $\frac{1}{|\beta|}$

БИНАРНАЯ КЛАССИФИКАЦИЯ

SVM

$$|\beta| \to \min_{\beta}$$

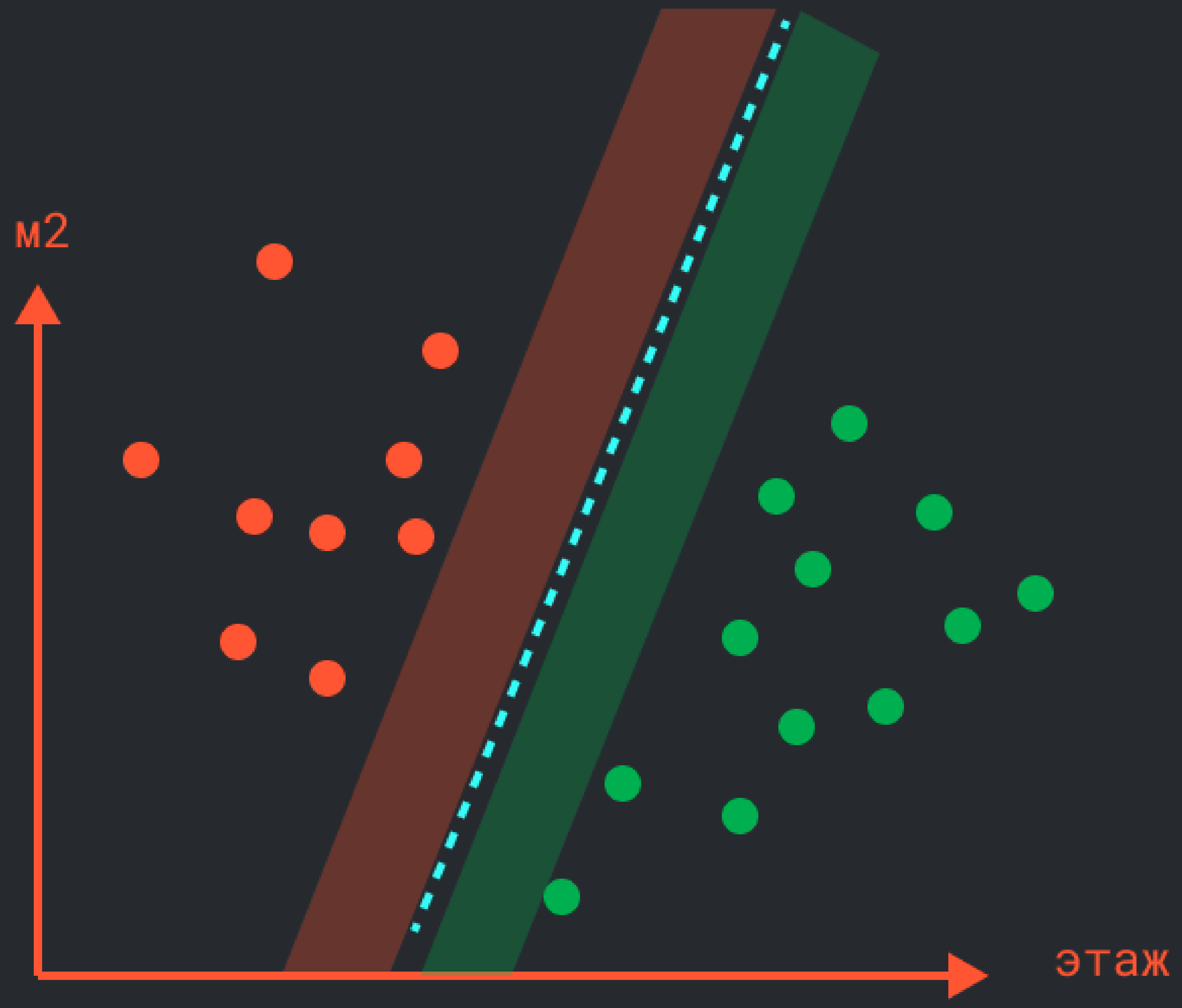
$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 0$$

$$\min_{x \in X} \langle \beta, X \rangle = 1$$

А еще брать производные от $|\beta|$ из-за корня:

$$|\beta| = \sqrt{\beta_1^2 + \cdots + \beta_n^2}$$



БИНАРНАЯ КЛАССИФИКАЦИЯ

SVM

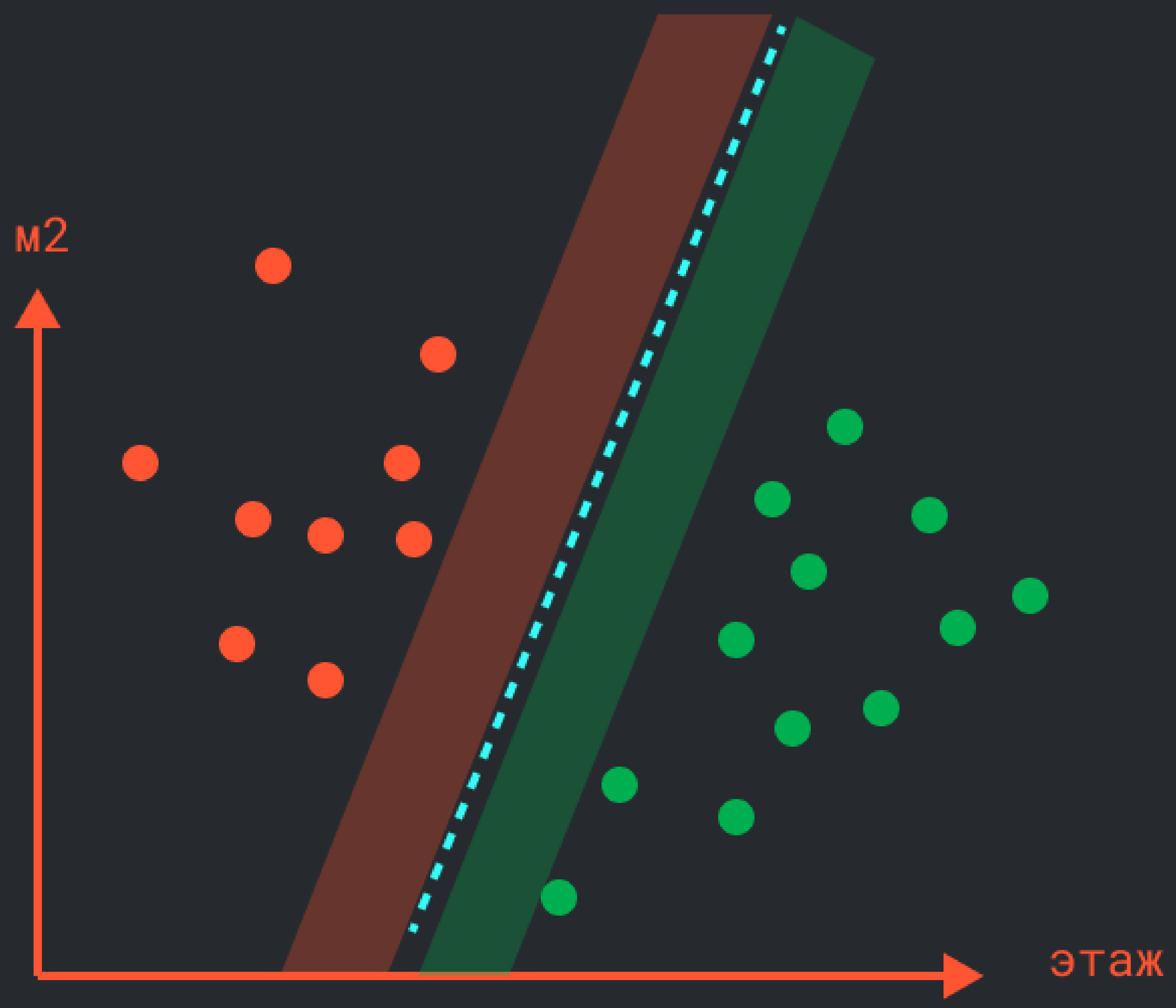
$$|\beta|^2 \to \min_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 0$$

$$\min_{x \in X} \langle \beta, X \rangle = 1$$

Заметим так же, что оба условия можно объединить в одно!



БИНАРНАЯ КЛАССИФИКАЦИЯ

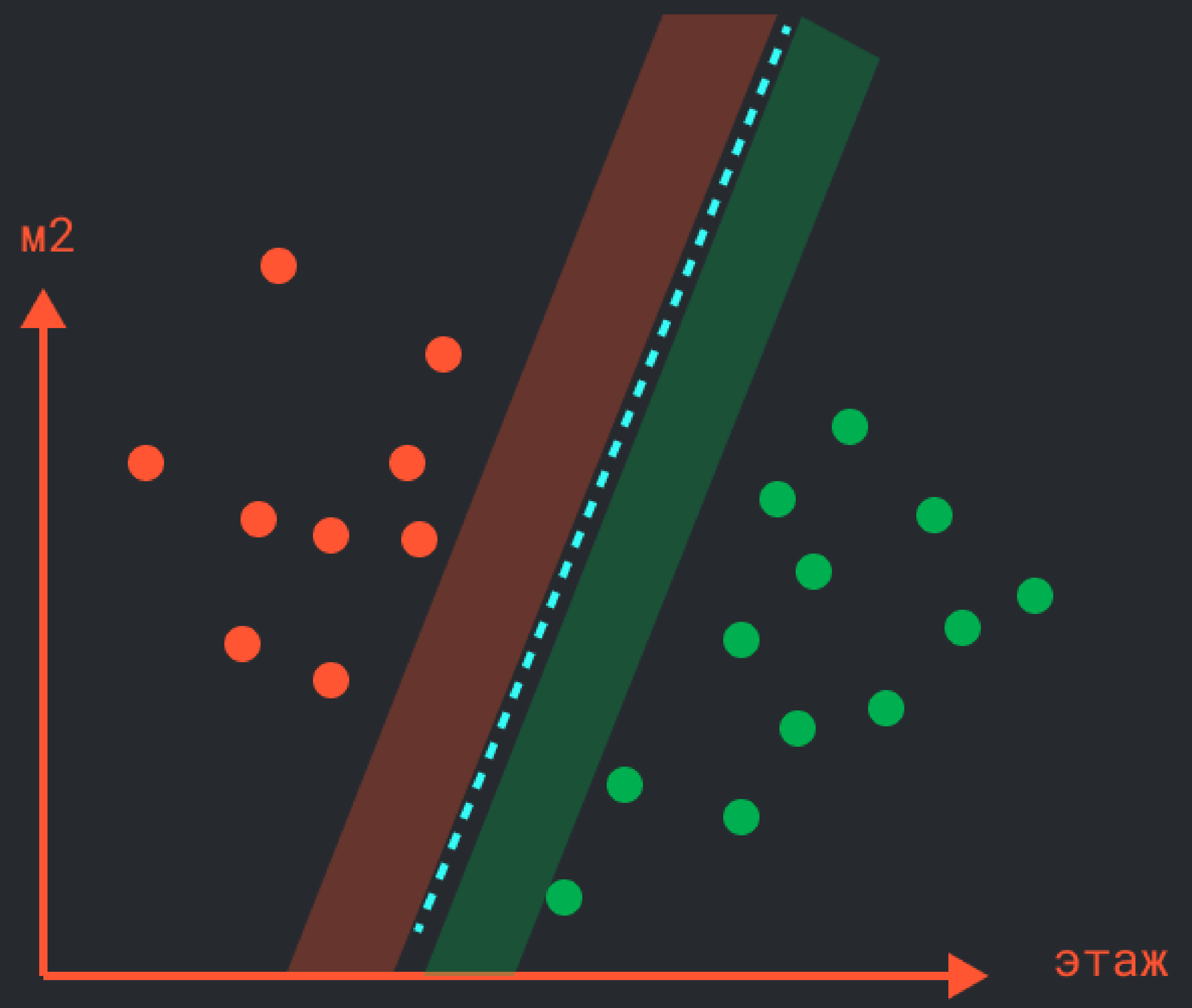
SVM

$$|\beta|^2 \to \min_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1$$

Таким образом мы и найдем лучшую разделяющую плоскость с самой большой разделяющей полосой!



БИНАРНАЯ КЛАССИФИКАЦИЯ

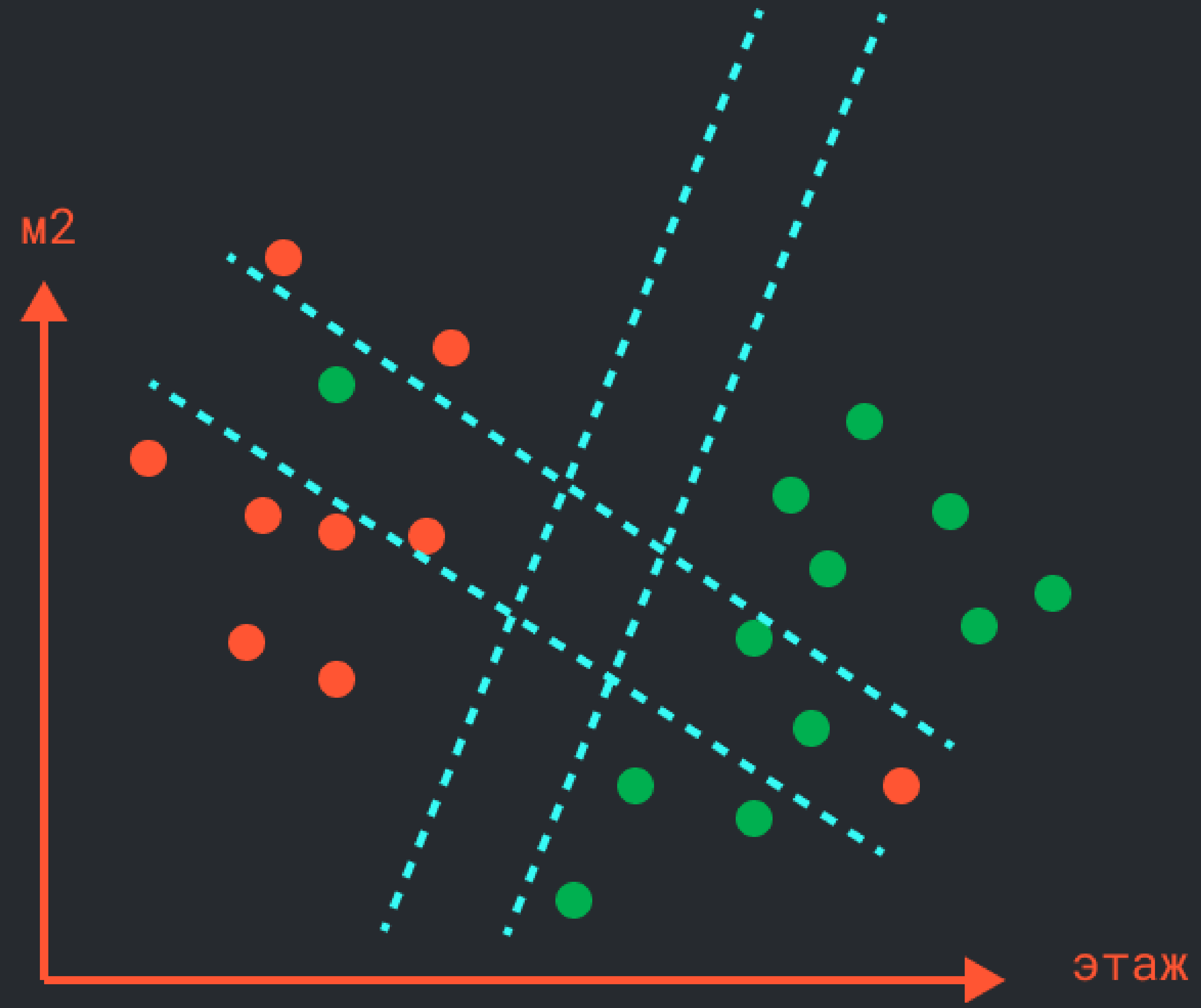
ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$|\beta|^2 \to \min_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1$$

Условие в задаче оптимизации ломается!



РЕЗЮМЕ

- Узнали, как работает метод классификации SVM
- Обсудили основную геометрическую интуицию метода
- Научились некоторым фишкам в области оптимизации

- Осталось научиться решать похожую задачу для линейно неразделимого случая!

БИНАРНАЯ КЛАССИФИКАЦИЯ

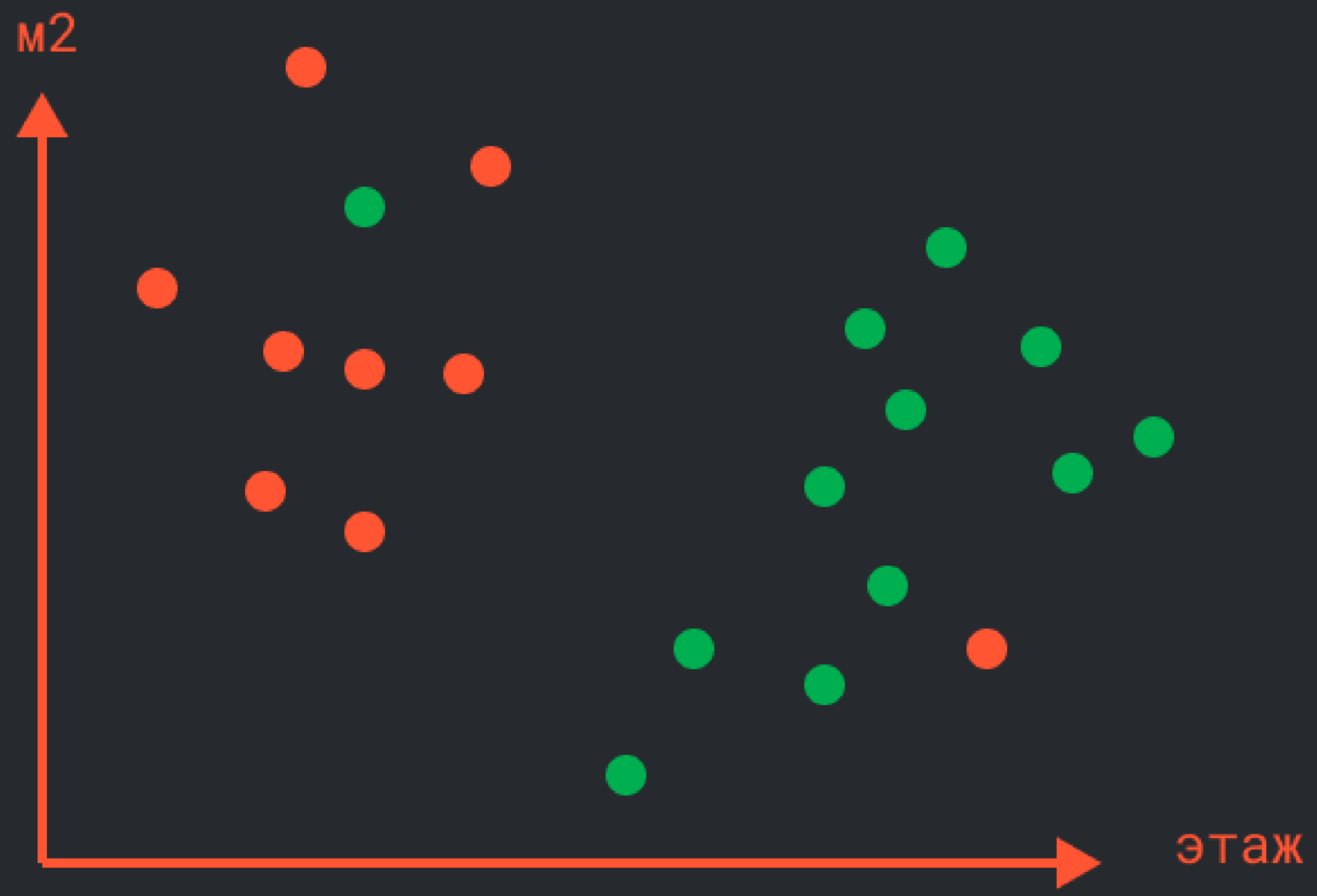
ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$|\beta|^2 \to \min_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1$$

Условие в задаче оптимизации ломается!



Может быть, его можно смягчить?

БИНАРНАЯ КЛАССИФИКАЦИЯ

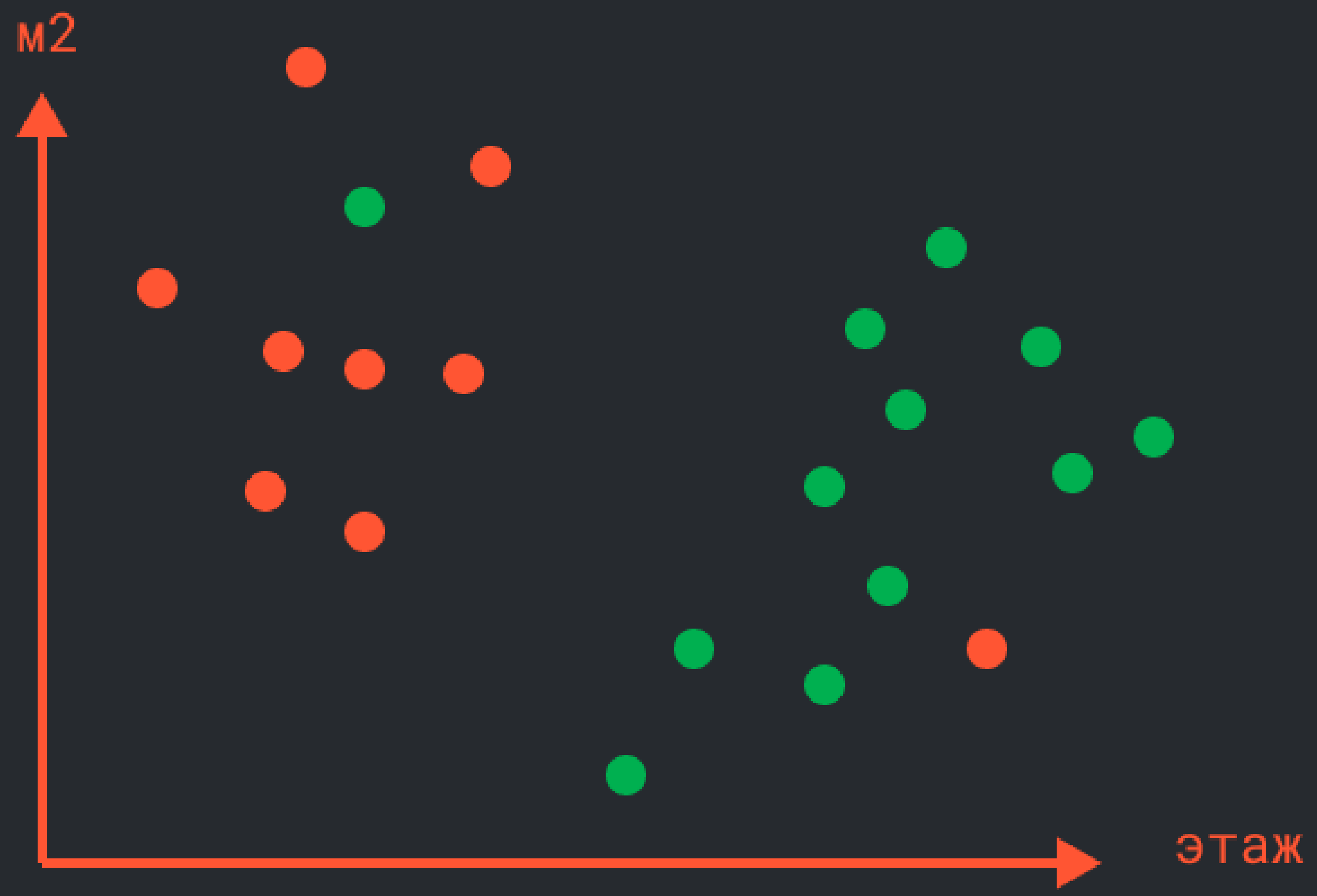
ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$|\beta|^2 \to \min_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$



Параметр ξ трансформирует условие жесткого разделения пространства по классам на более мягкое! Причем этот параметр можно подбирать

БИНАРНАЯ КЛАССИФИКАЦИЯ

ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

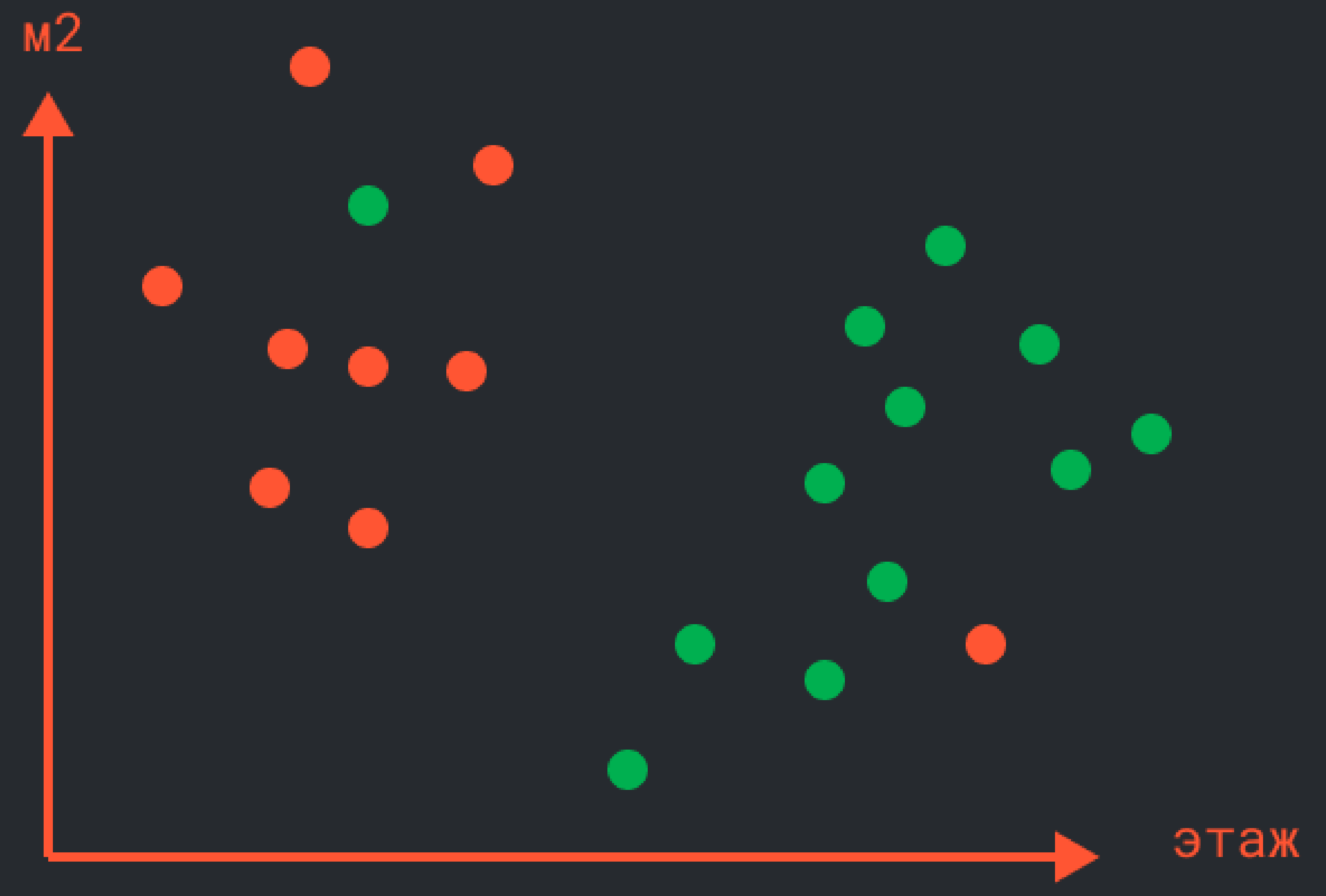
Например, если $\xi_i = 1.000.000$, то условие

$$y_i \cdot \langle \beta, x_i \rangle \geq 1 - \xi_i$$

трансформируется в

$$y_i \cdot \langle \beta, x_i \rangle \geq -999.999$$

Что эквивалентно тому, что мы разрешаем нашей гиперплоскости иметь большие отрицательные отступы на объектах – это очень плохо!



БИНАРНАЯ КЛАССИФИКАЦИЯ

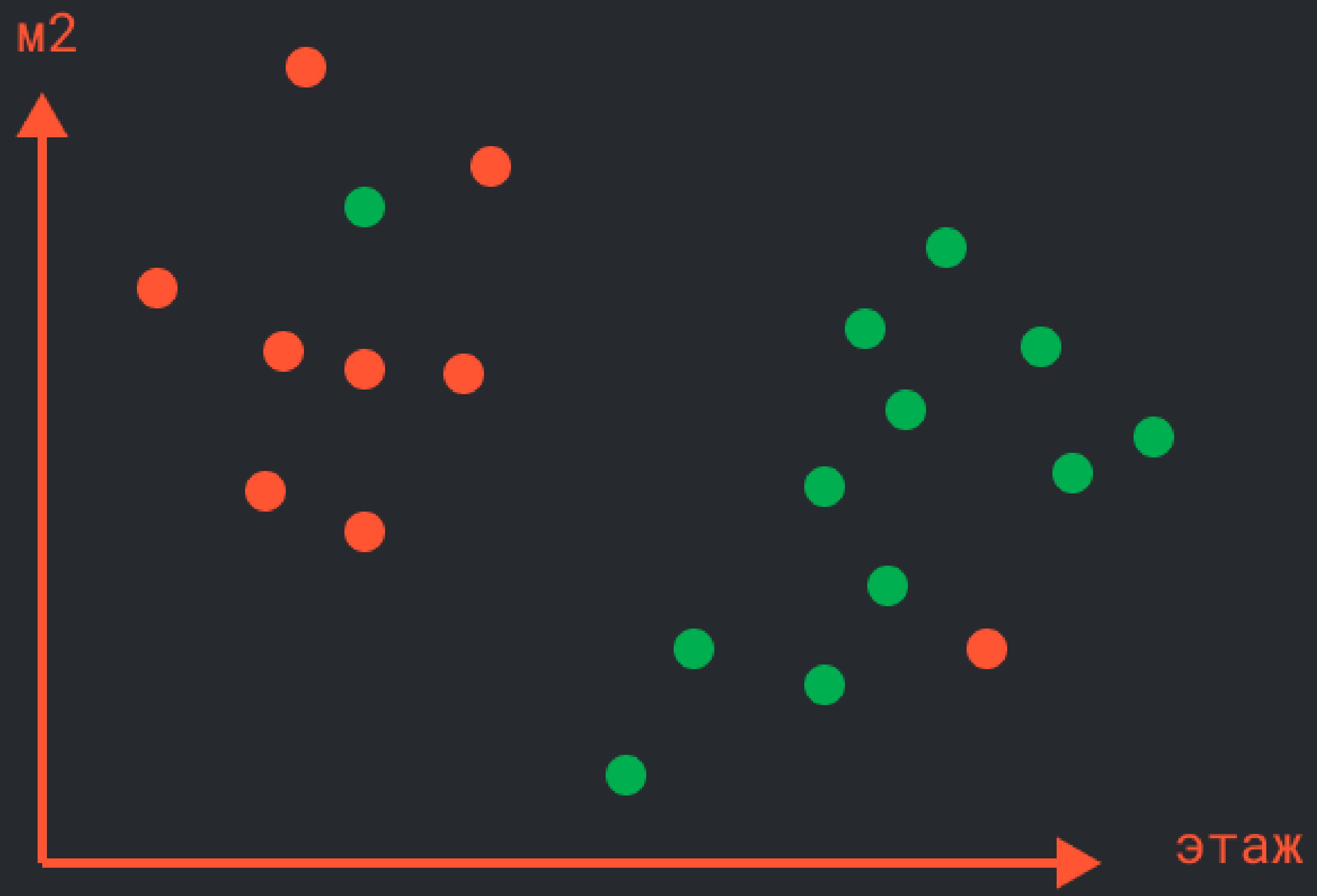
ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$|\beta|^2 \to \min_{\beta}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$



Поэтому ξ_i стоит выбирать аккуратно и не крайне лояльно. Делать это необходимо на каждом объекте. Может, заставить модель самой выбирать?

БИНАРНАЯ КЛАССИФИКАЦИЯ

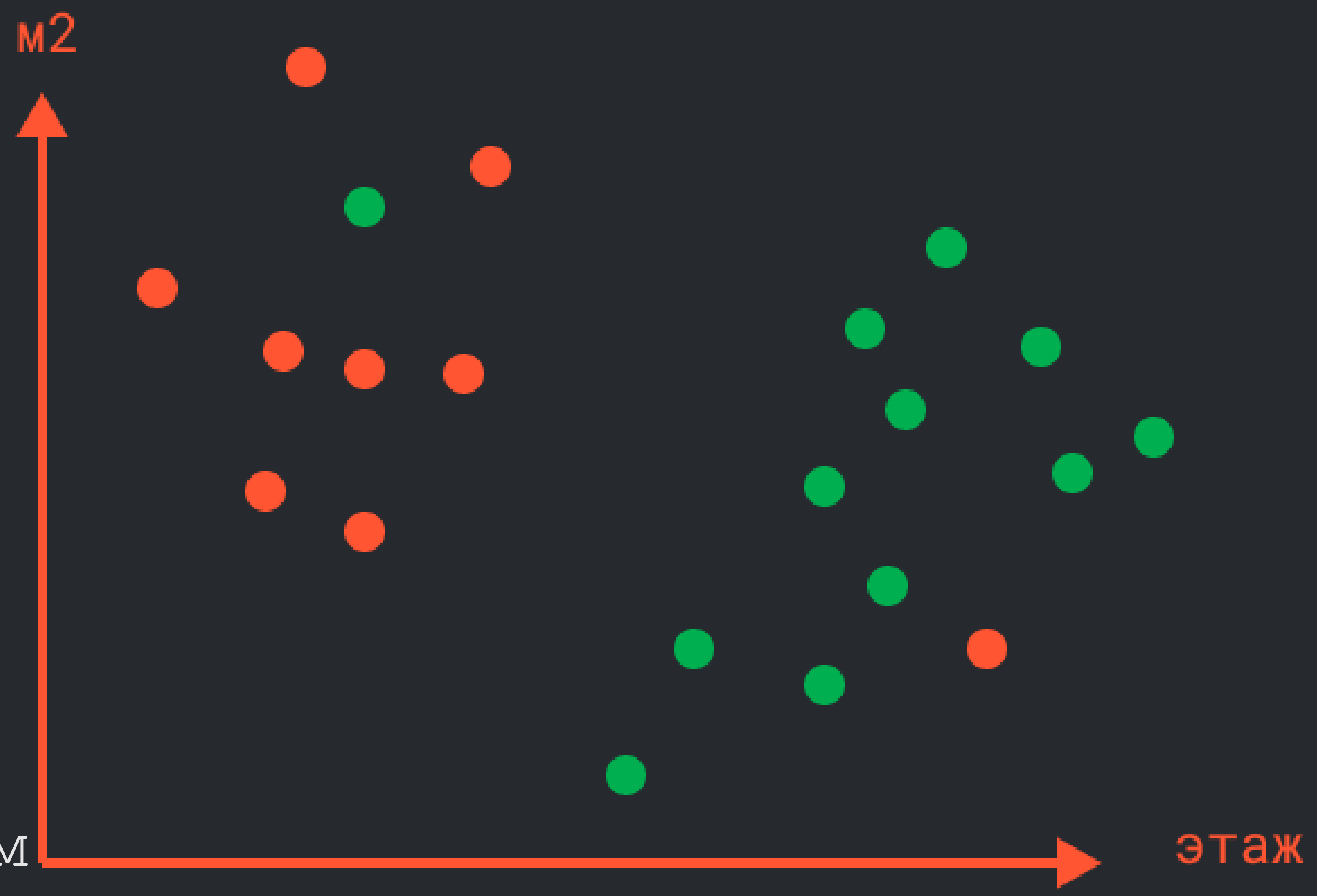
ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$|\beta|^2 + \lambda \cdot \sum_i^n \xi_i \quad \to \min_{\beta,\xi}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$



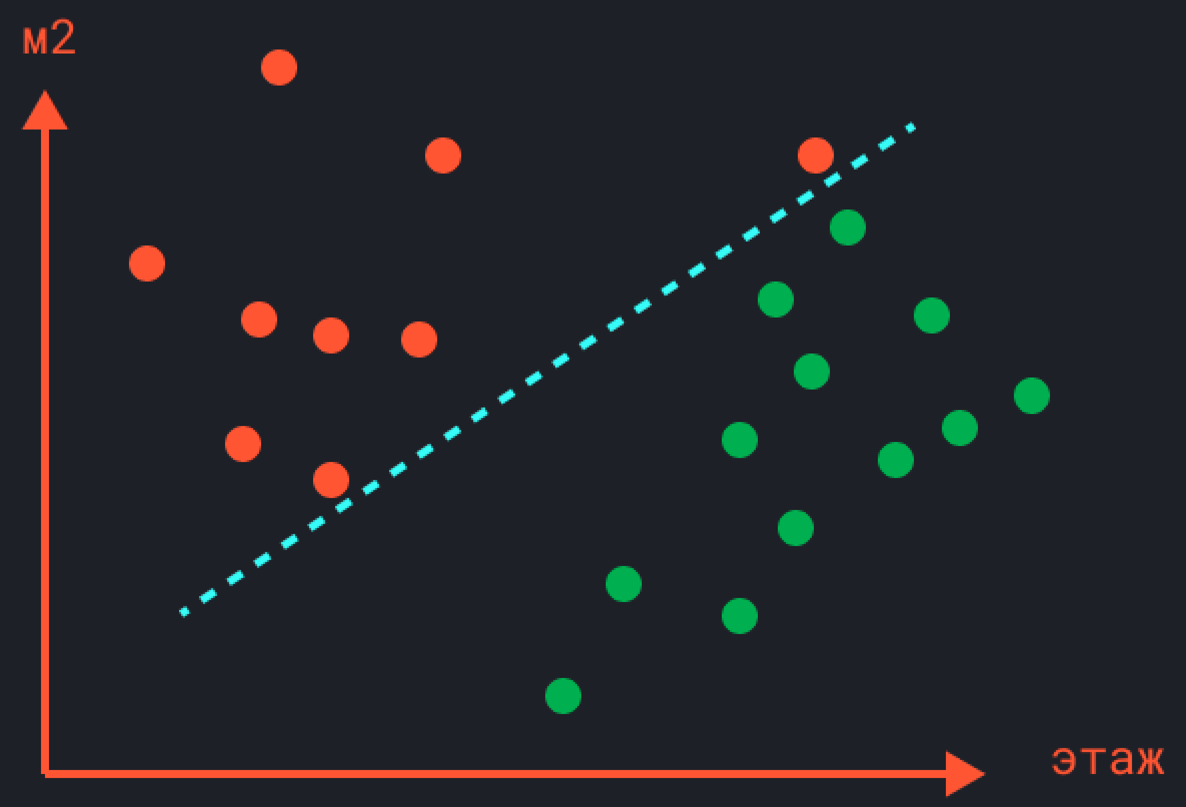
Добавим регуляризатор по всем элементам ξ_i!

Параметр λ, как и ранее, отвечает за силу

СРАВНЕНИЕ РЕГУЛЯРИЗАЦИИ РАЗНОЙ СИЛЫ

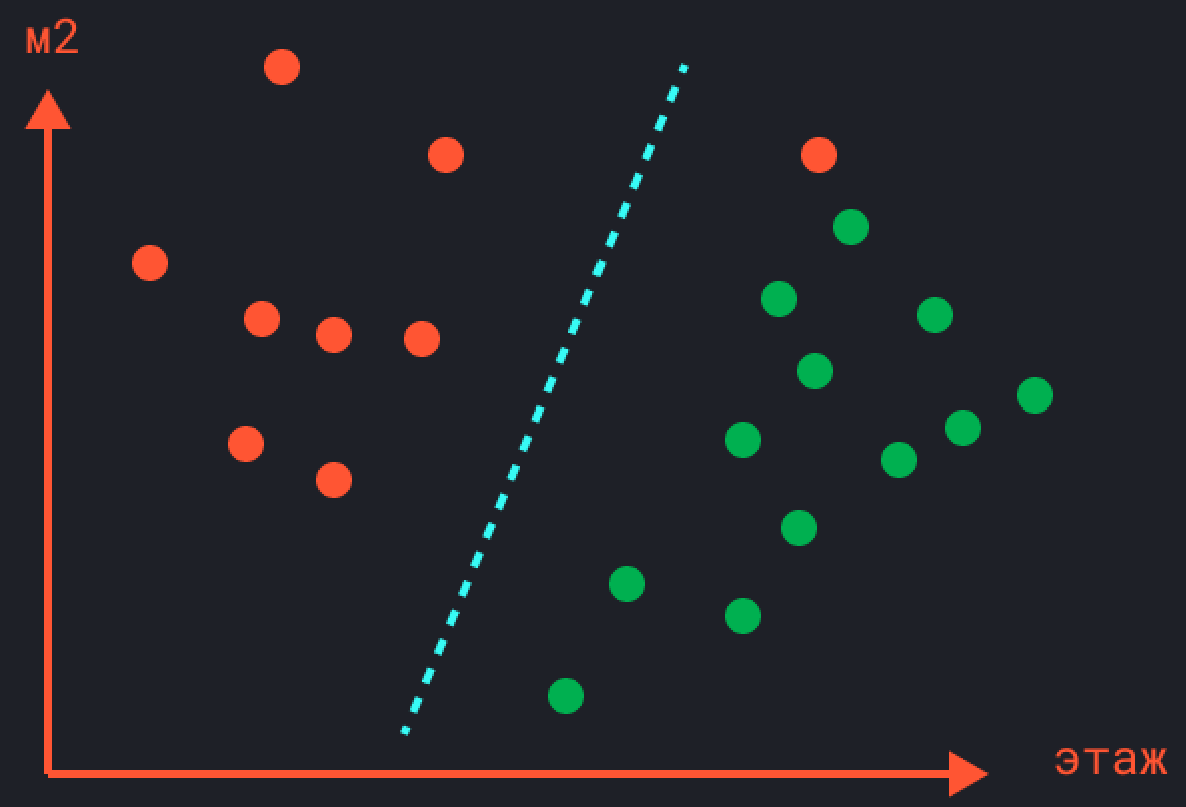
БОЛЬШОЕ $\lambda \to +\infty$

Боимся делать любые послабления



МАЛЕНЬКОЕ $\lambda \approx 0$

Акцентируемся на широкой полосе



БИНАРНАЯ КЛАССИФИКАЦИЯ

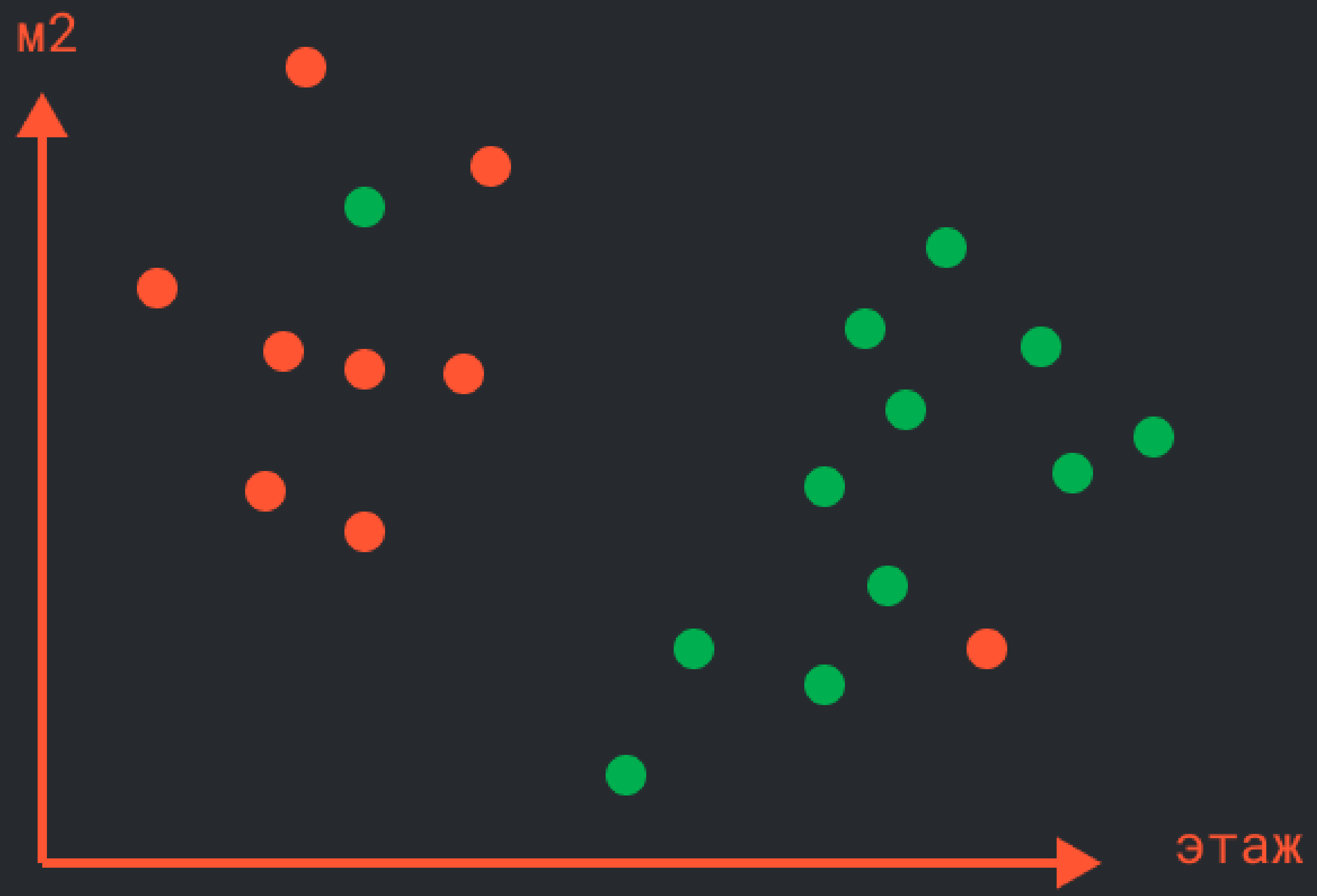
ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$|\beta|^2 + \lambda \cdot \sum_i^n \xi_i \quad \rightarrow \min_{\beta,\xi}$$

$$s.t.$$

$$y_i \cdot \langle \beta, x_i \rangle \geq 1 - \xi_i$$

$$\xi_i \ \geq 0$$



Вообще говоря, такая задача содержит много условий. Решать ее достаточно сложно. Может быть, можно уместить ее компактно? Например, безусловно

БИНАРНАЯ КЛАССИФИКАЦИЯ

ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

$$y_i \cdot \langle \beta, x_i \rangle \geq 1 - \xi_i$$

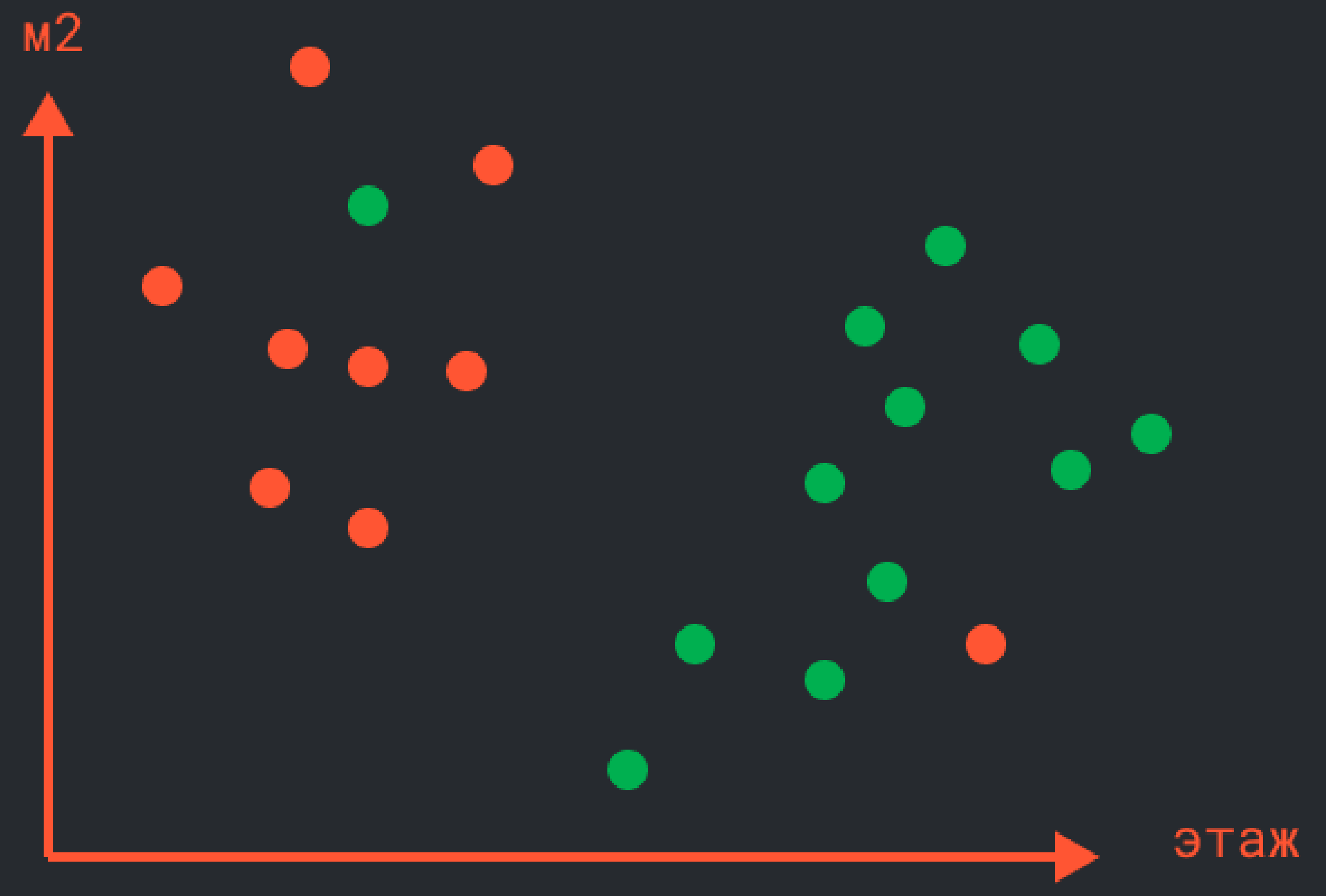
$$\xi_i \geq 0$$

Перепишем в следующем виде:

$$\xi_i \geq 1 - y_i \cdot \langle \beta, x_i \rangle$$

$$\xi_i \geq 0$$

$$\xi_i = \max(0, 1 - y_i \cdot \langle \beta, x_i \rangle)$$



БИНАРНАЯ КЛАССИФИКАЦИЯ

ЛИНЕЙНАЯ НЕРАЗДЕЛИМОСТЬ

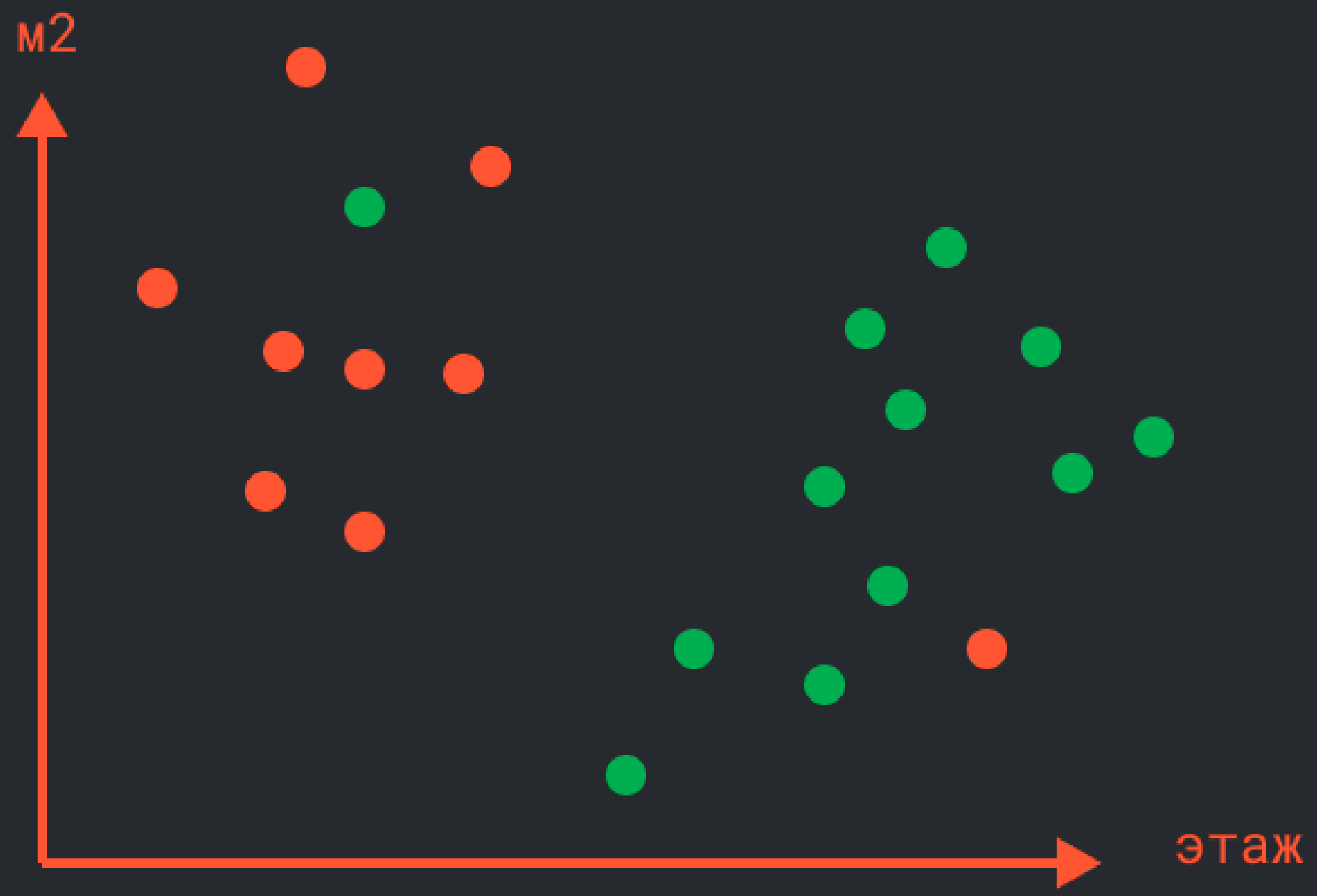
$$|\beta|^2 + \lambda \cdot \sum_i^n \xi_i \to \min_{\beta,\xi}$$

$$s.t.$$

$$\xi_i = \max(0, 1 - y_i \cdot \langle \beta, x_i \rangle)$$

Или, в одну строчку:

$$|\beta|^2 + \lambda \cdot \sum_i^n \max(0, 1 - y_i \cdot \langle \beta, x_i \rangle) \to \min_{\beta,\xi}$$



РЕЗЮМЕ

- Познакомились с обобщенным SVM
- Узнали про механизм регуляризации
- А так же посмотрели, как задачу построения SVM-гиперплоскости можно свести к безусловному экстремуму!

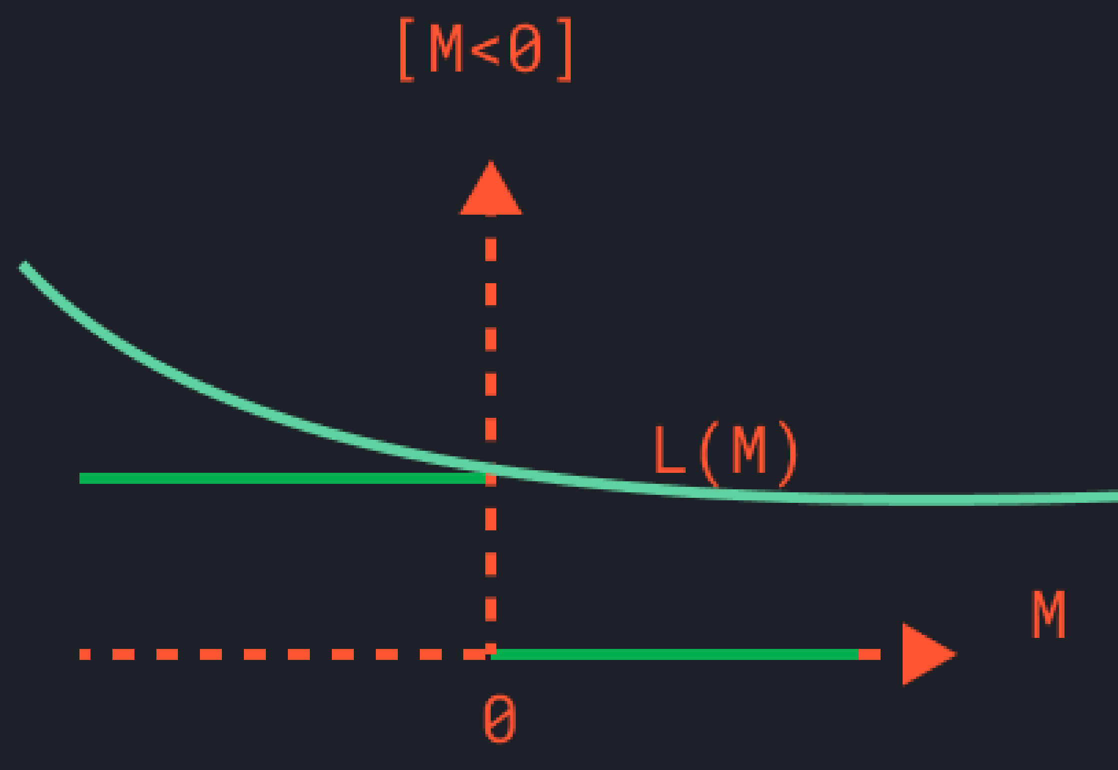
- Пора сравнить со старой доброй логистической регрессией!

СРАВНЕНИЕ ПОДХОДОВ

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

$$\sum_i^n [M < 0] \le$$

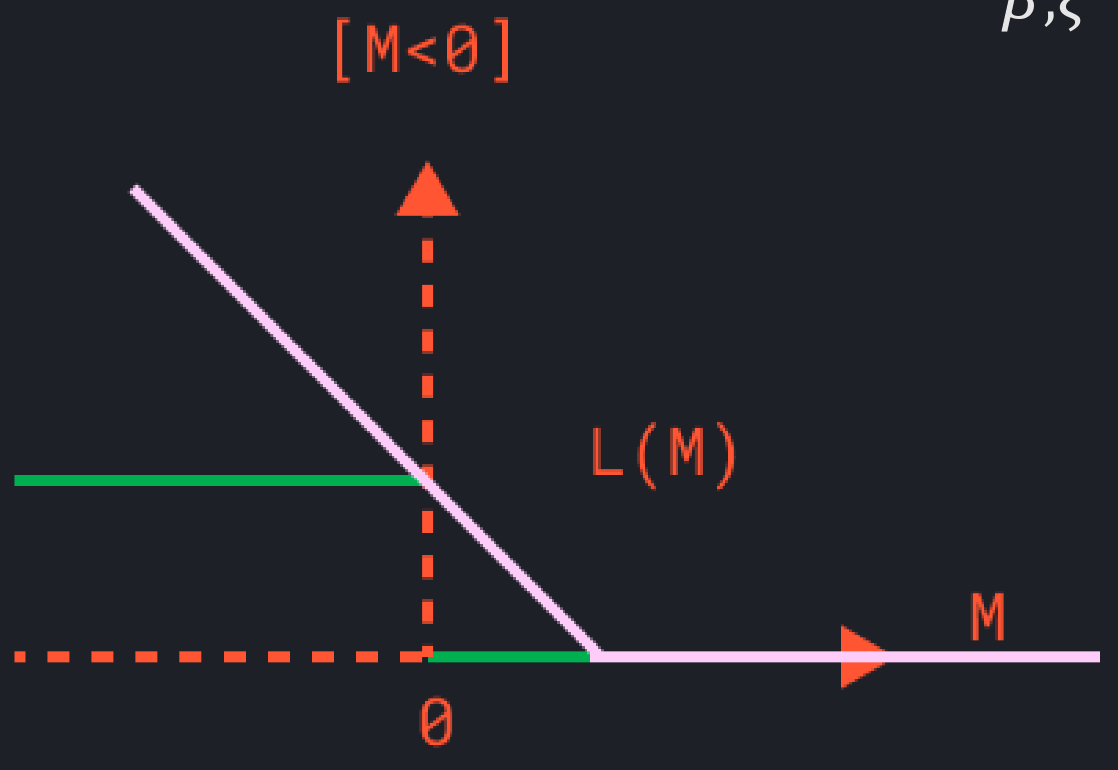
$$\sum_i^n \log(1 + e^{-M}) \quad \to min$$



SVM

$$|\beta|^2 + \lambda \cdot \sum_i^n \max(0, 1 - y_i \cdot \langle \beta, x_i \rangle) \to \min_{\beta,\xi}$$

$$|\beta|^2 + \lambda \cdot \sum_i^n \max(0, 1 - M) \to \min_{\beta,\xi}$$



РЕЗЮМЕ

- Оказалось, что SVM – это частный случай обычной минимизации индикаторов отступов, просто со специфичной верхней оценкой и модификацией в виде регуляризации
- В отличие от логистической регрессии, SVM не пытается корректно оценить вероятности и максимизировать уверенности своих прогнозов
- Метод опорных векторов скорее про построение разделяющей гиперплоскости
- Такой, у которой оказывается достаточно широкая разделяющая полоса
- P.S. Но даже если мы строим SVM, все еще можем оценивать вероятности
- Калибровка нам в помощь!

СПАСИБО

ТАБАКАЕВ НИКИТА