Линейная классификация

Занятие 4

Линейная модель
Повторим

x - ? ($x_1, x_2, \ldots, x_n$ - ?)

y - ?

Линейная модель

Повторим

x - объект (x_1, x_2 ,..., x_n - признаки)

y - ответ (целевая переменная)

Линейная модель

Повторим

x - объект (x_1, x_2 ,..., x_n - признаки)

y - ответ (целевая переменная)

Модель линейной регрессии: ответ складывается из ...?

$$a(\boldsymbol{x}, w) = \sum_{i=1}^{l} w_j x_j$$

Линейная модель
Повторим

x - объект (x_1, x_2 ,..., x_n - признаки)

y - ответ (целевая переменная)

Модель линейной регрессии:

ответ складывается из суммы «признаки» * «веса»

$$a(\boldsymbol{x}, w) = \sum_{i=1}^{l} w_j x_j$$

Линейная модель
Повторим

Метод обучения (минимизируем разницу между предсказанием и правильным ответом) - с помощью чего?

Линейная модель

Повторим

Метод обучения (минимизируем разницу между предсказанием и правильным ответом) - MSE

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

Линейная модель

Повторим

x - объект (x_1, x_2 ,..., x_n - признаки)

y - ответ (целевая переменная)

Модель линейной регрессии: ответ складывается из суммы «признаки» * «веса»

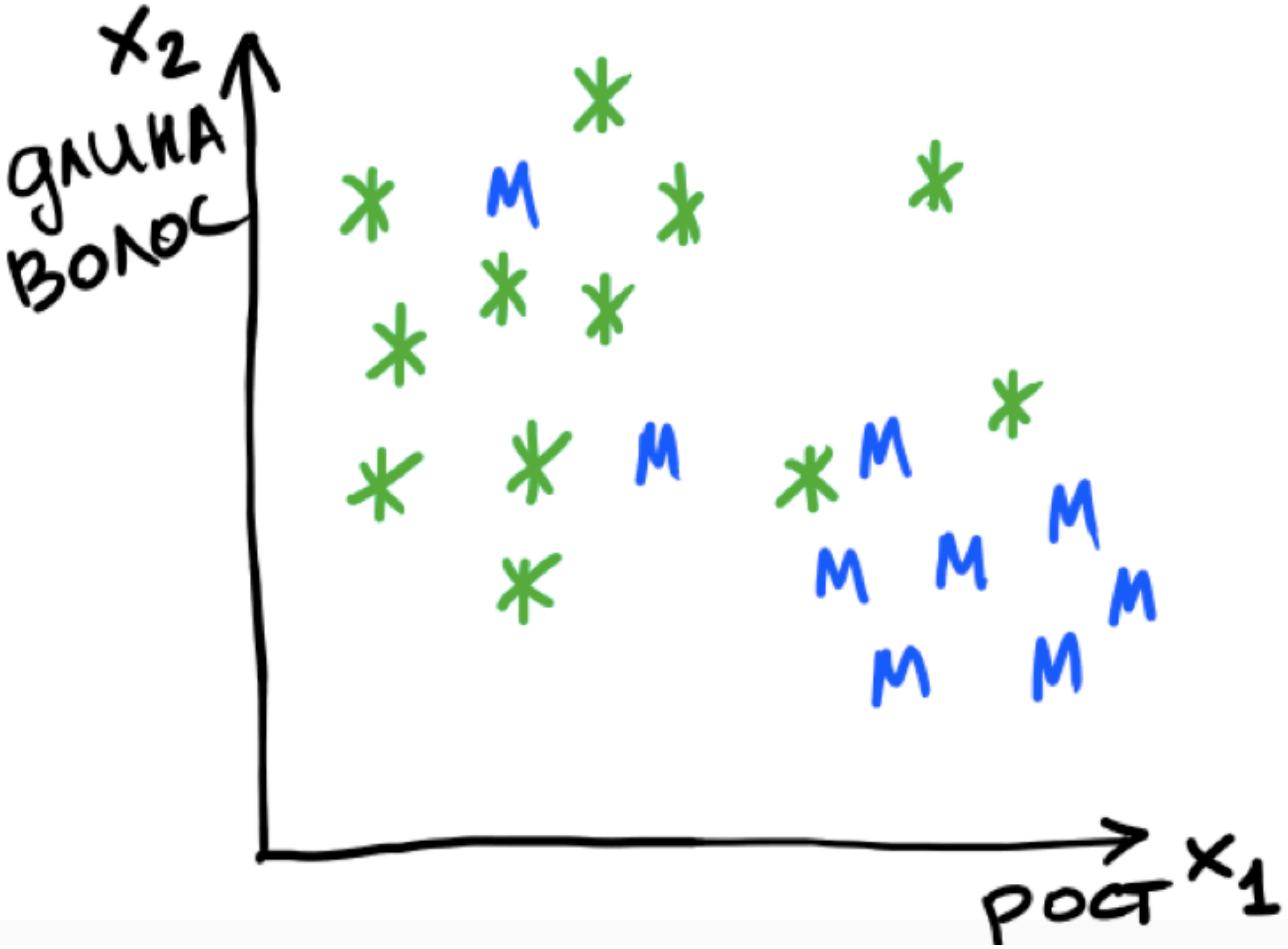
$$a(\boldsymbol{x}, w) = \sum_{i=1}^{l} w_j x_j$$

Метод обучения (минимизируем разницу между предсказанием и правильным ответом) с помощью MSE

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

Бинарная классификация

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?

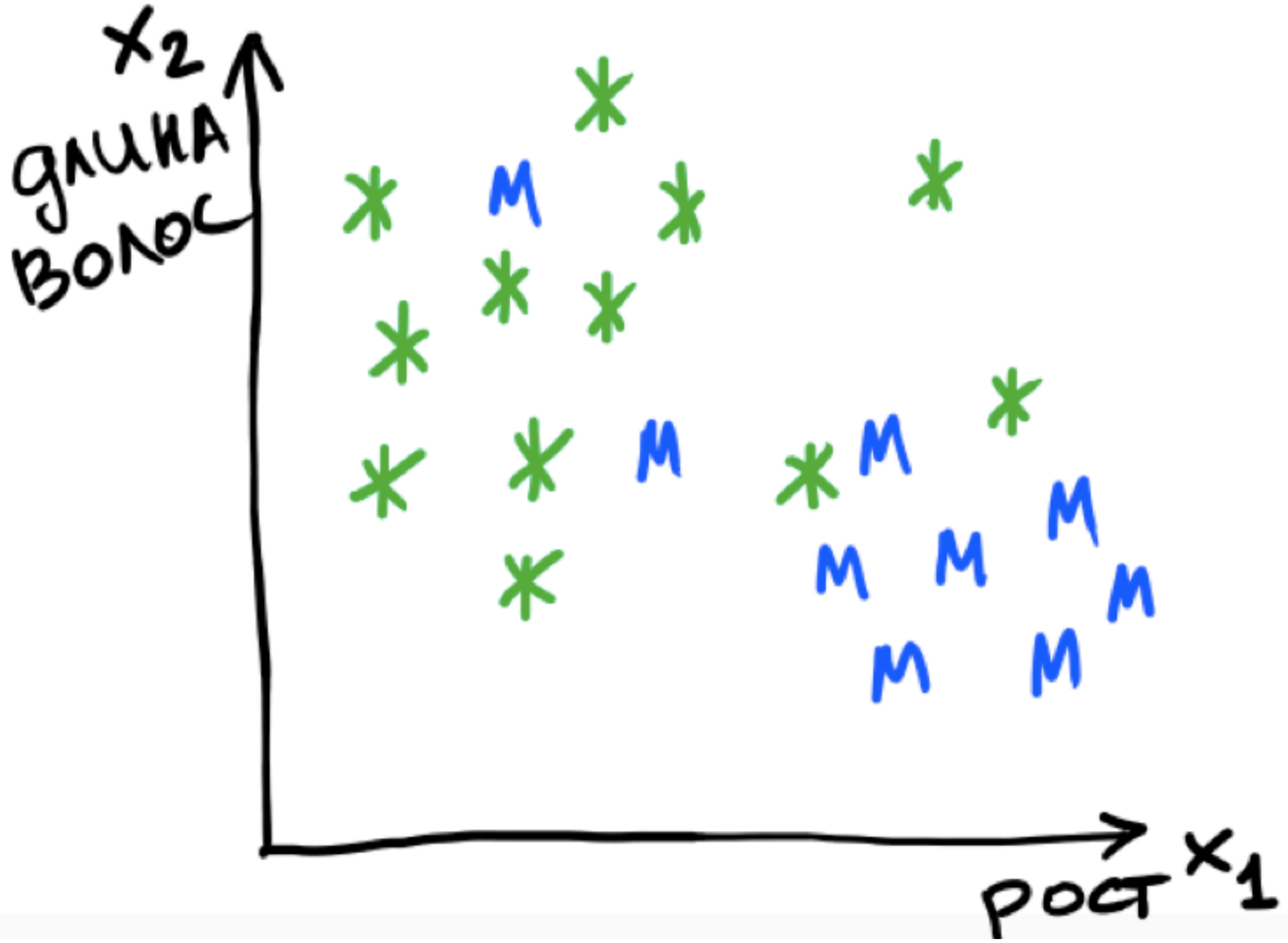


$y_1, y_2, \ldots, y_n$ - ответы (только +1 или -1)

У нас на рисунке Ж будет соответствовать +1, а М -1

Бинарная классификация

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?

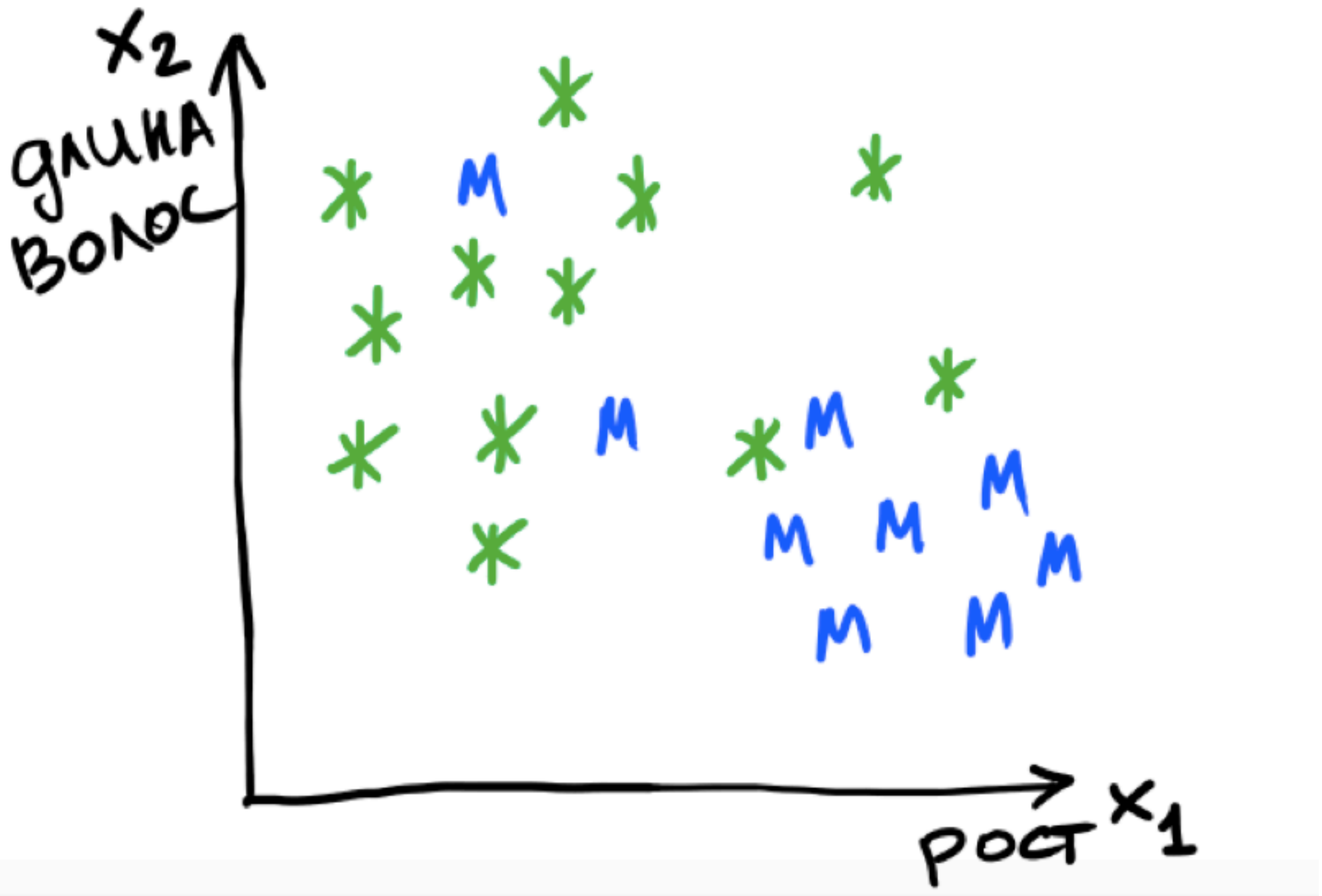


У нас на рисунке Ж будет соответствовать +1, а М -1

Можем ли мы использовать ту же формулу, что и в линейной регрессии, чтобы найти *y*?

$$a(\boldsymbol{x}, w) = \sum_{i=1}^{l} w_j x_j$$

Бинарная классификация



Можем ли мы использовать ту же формулу, что и в линейной регрессии, чтобы найти *y*?

$$a(\boldsymbol{x}, w) = \sum_{i=1}^{l} w_j x_j$$

Нет, она выдает произвольные числа. Нам не важно значение *y*, **нужен только знак: - или +**

Модель линейного классификатора

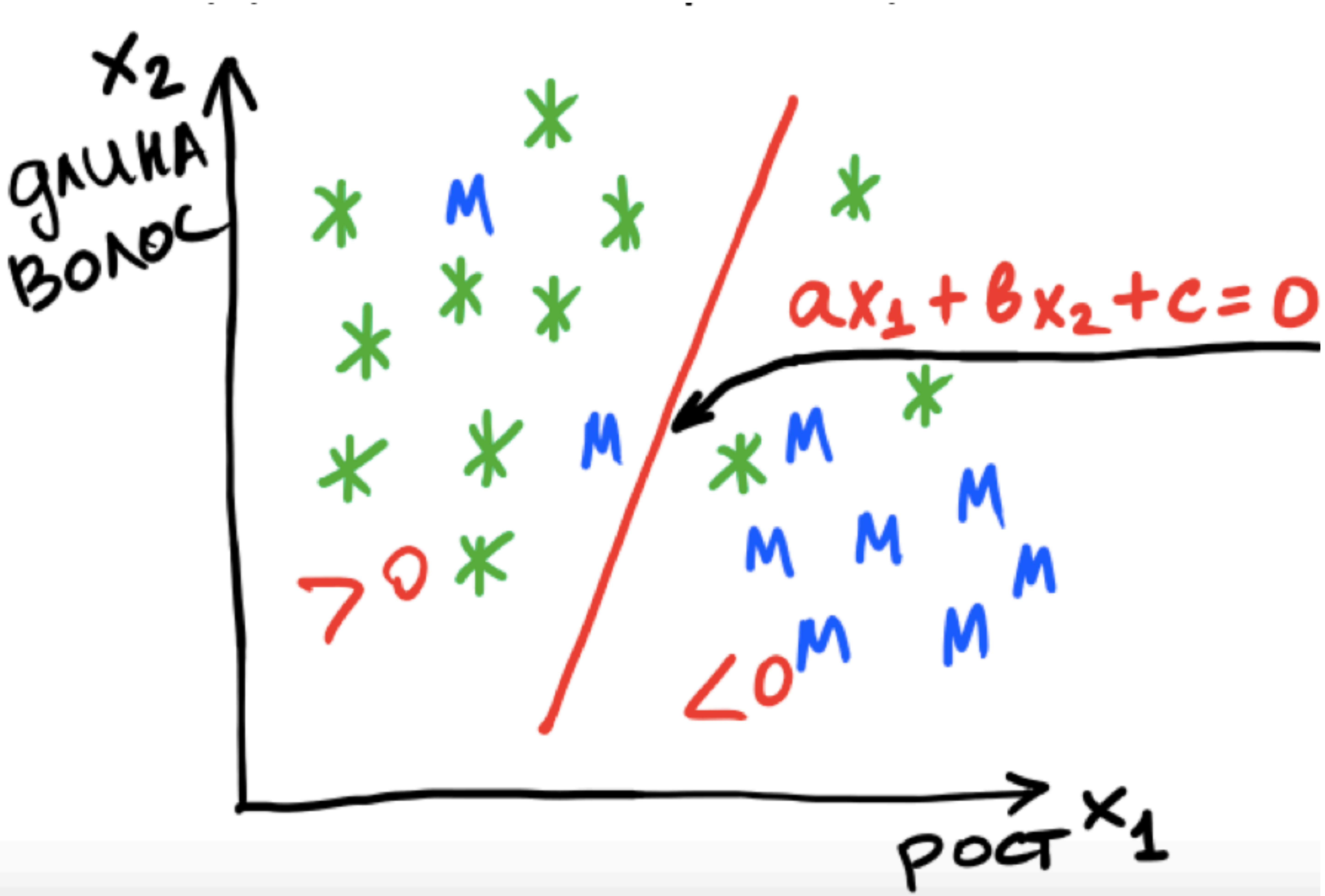
Подбираем веса для признаков так, чтобы объекты из одного класса имели вес со знаком +, а другого -

$$a(x, w) = sign(\sum_{j=1}^{l} w_j x_j)$$

Берем из этого только знак

Модель линейного классификатора

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



Линейный классификатор

$$a(x) = \mathrm{sign}\left(w_0 + \sum_{j=1}^{d} w_j x^j\right)$$

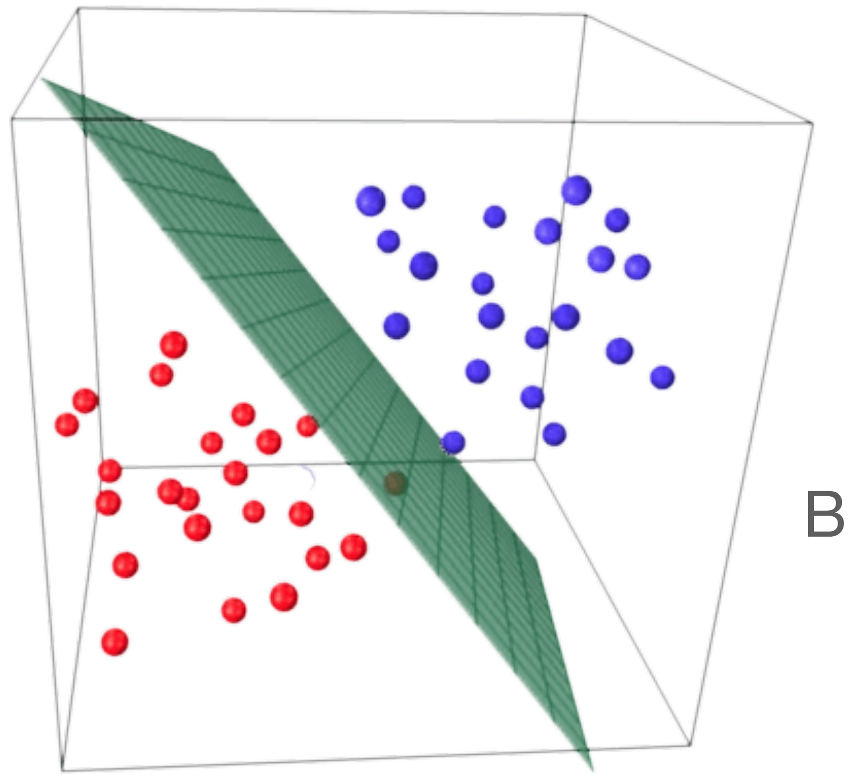
Свободный коэффициент

Веса

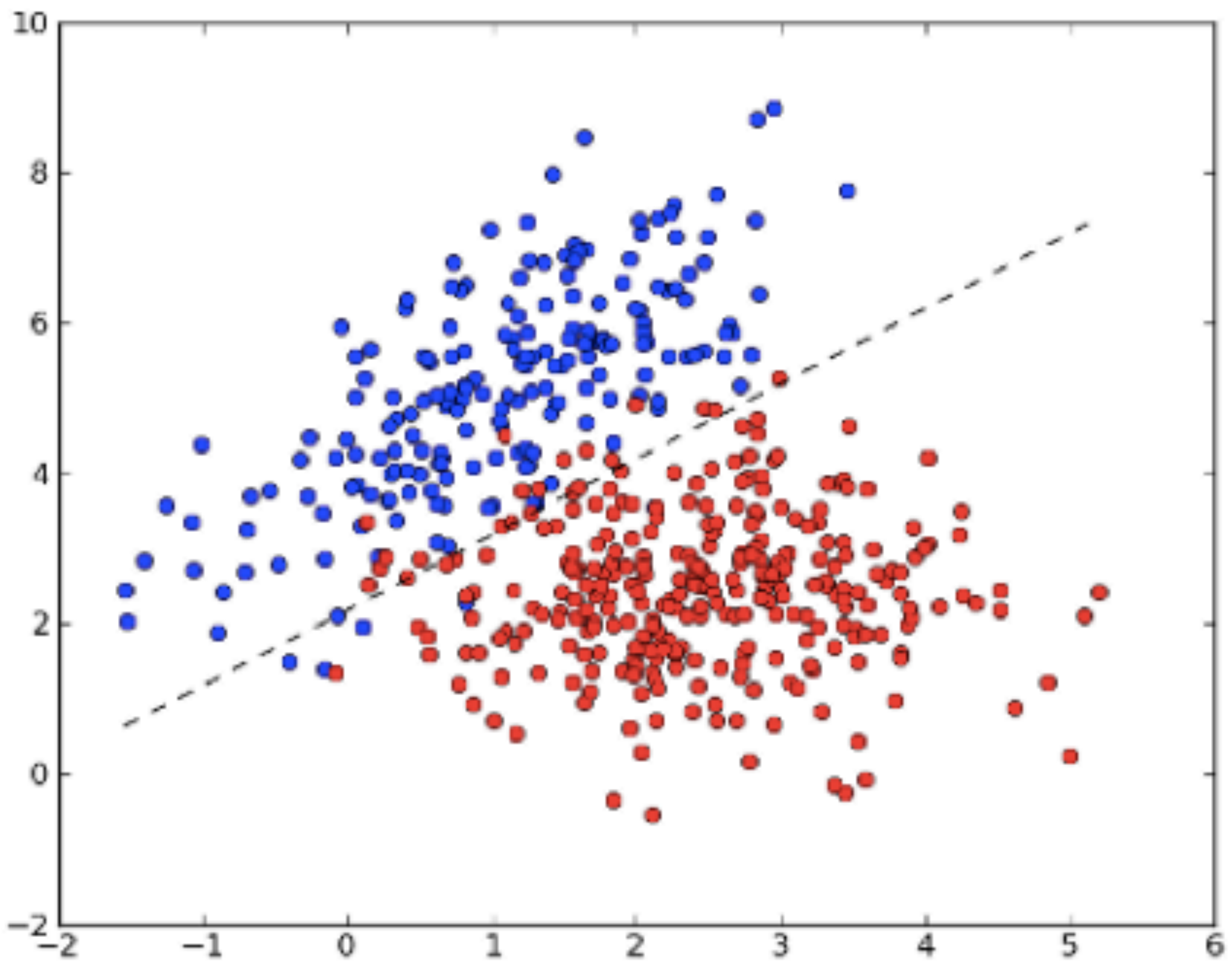
Признаки

Линейный классификатор
Геометрия

- Линейный классификатор проводит гиперплоскость
- $\langle w, x \rangle < 0$ — объект «слева» от неё
- $\langle w, x \rangle > 0$ — объект «справа» от неё



В 3-мерном случае



Обучение классификатора

Как обучить?

Обучение - минимизация доли ошибок классификатора:

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

где $[a(\boldsymbol{x}_i) \neq y_i] = 1$, если предсказание на объекте неверное, и 0, если верное.

Обучение классификатора

Как обучить?

Обучение - минимизация доли ошибок классификатора:

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

где $[a(x_i) \neq y_i] = 1$, если предсказание на объекте неверное, и 0, если верное.

Делим кол-во ошибок на кол-во объектов. Получаем долю ошибок

Отступ классификатора (margin)

$$Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{a}(\mathbf{x}_i) \neq \mathbf{y}_i] \rightarrow min$$

Нам это не нравится

$M_{\boldsymbol{i}} = y_{\boldsymbol{i}} \cdot (w_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{i}})$ - ***отступ** на i-м объекте.*

Если классификация неверная, какой будет знак здесь: $y_{\boldsymbol{i}} \cdot (w_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{i}})$?

Отступ классификатора (margin)

Если классификация неверная, знак $y_i \cdot (w_i, \boldsymbol{x_i})$ всегда будет **минус**

- Если наши множители имеют разный знак - М будет **отрицательный**
- Если наши множители одного знака - М будет **положительный**

Margin всегда отрицательный при ошибке

Обучение классификатора

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

$$=$$

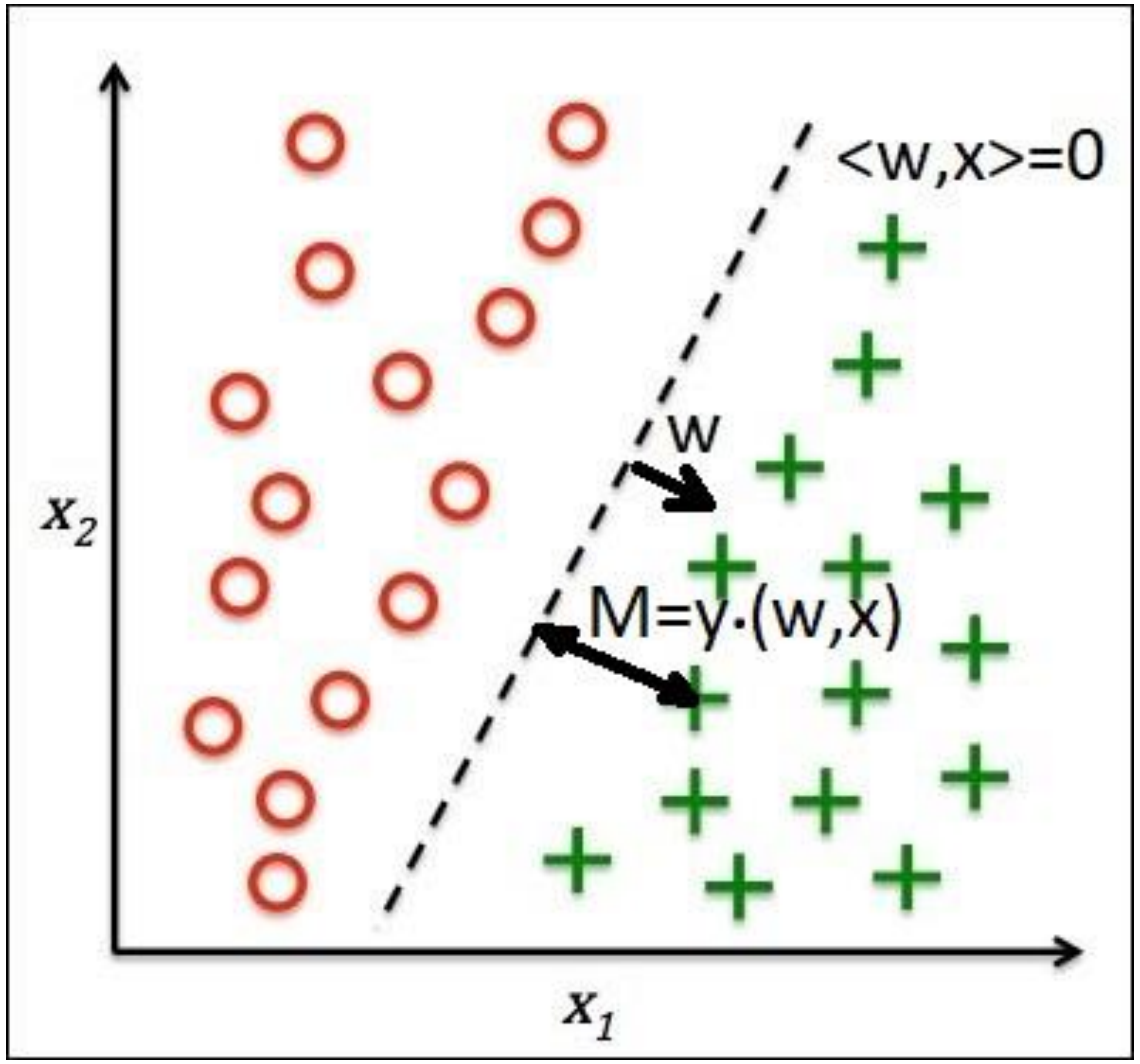
$$Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}[M_i < 0] \rightarrow min$$

Отступ классификатора (margin)

Абсолютная величина отступа М обозначает **степень уверенности классификатора в ответе**

(чем ближе М к нулю, тем меньше уверенность в ответе)

Отступ = насколько мы далеко от разделяющей поверхности



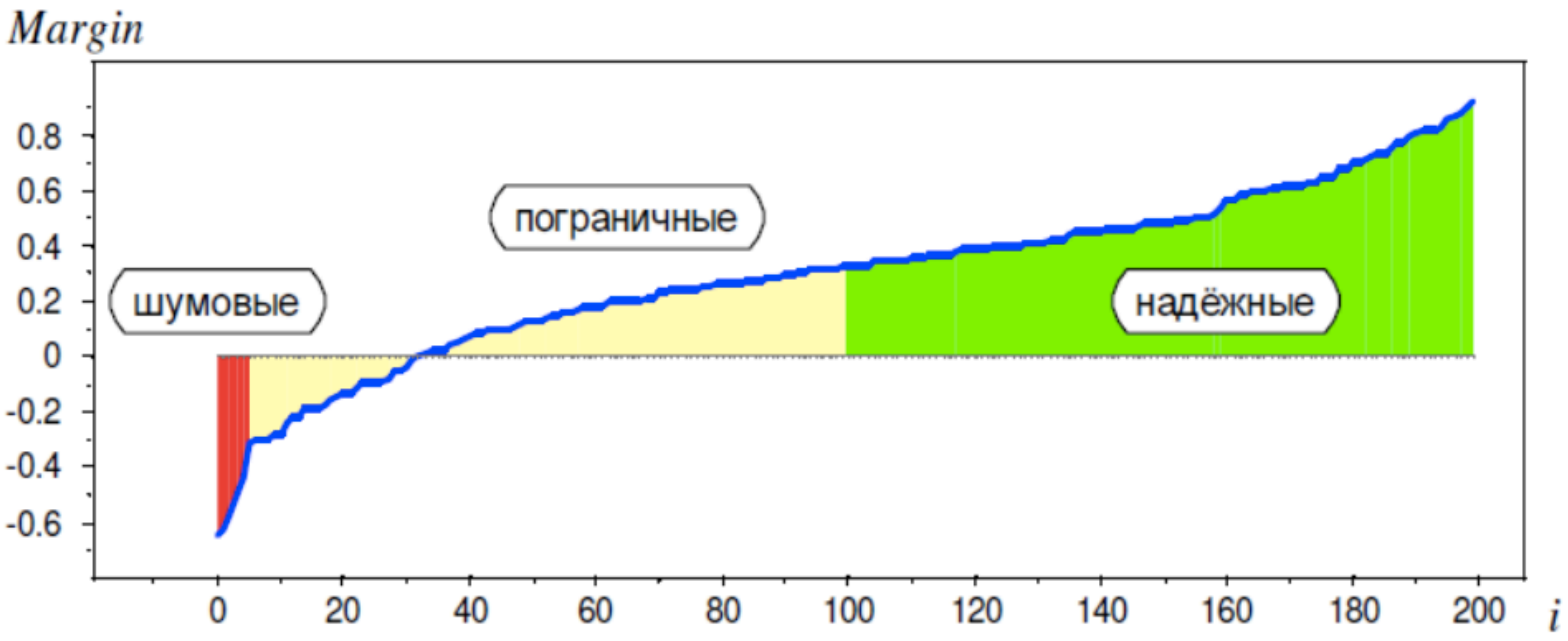
Отступ

Ранжирование значения

Надежные - большой положительный отступ

Пограничные - небольшой положит. или отрицательный

Шумовые - большой отрицательный (сильно ошиблись, выброс)



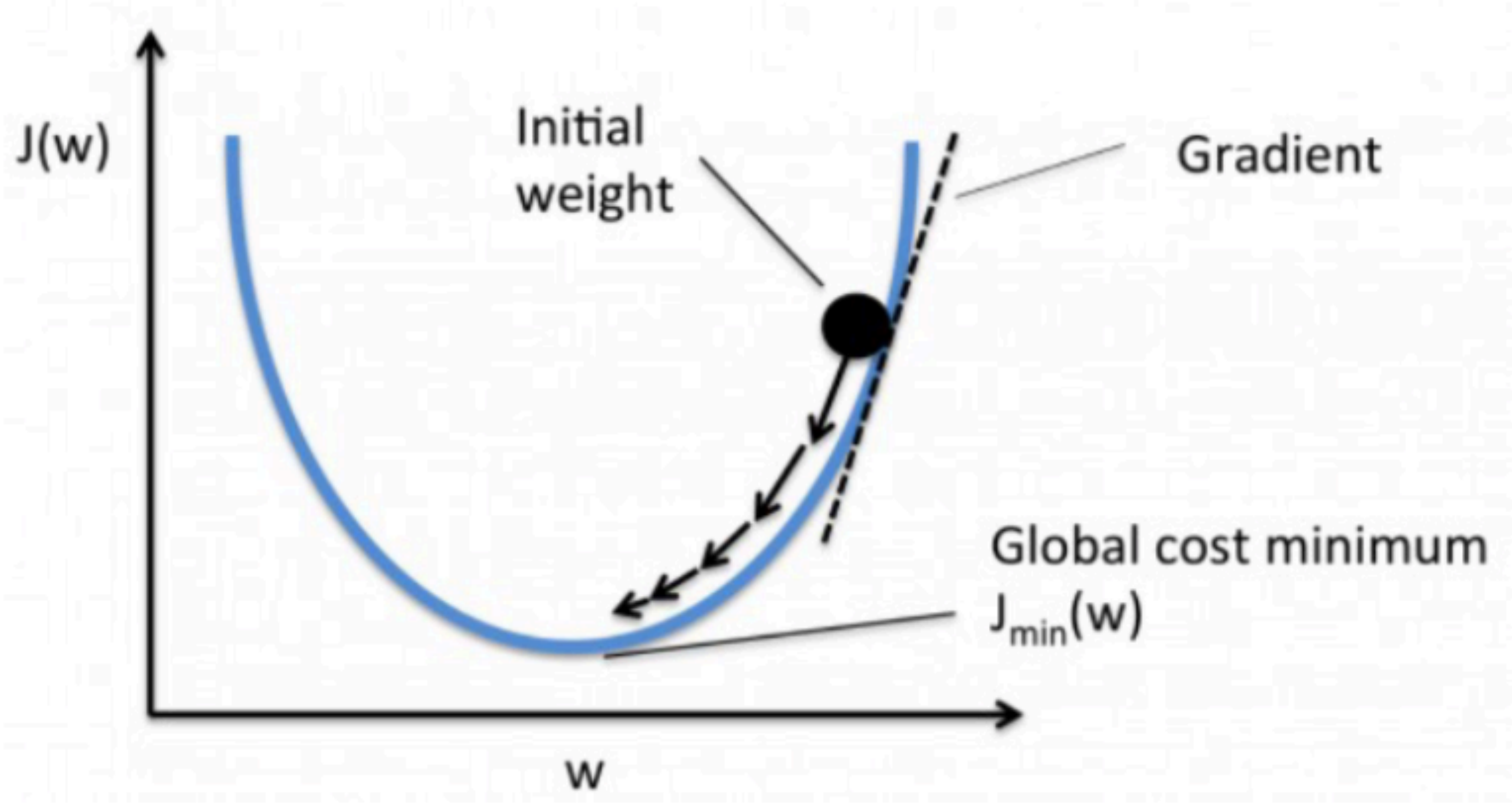
Линейная регрессия
Ну-ка вспомним

Как мы минимизировали ошибки в линейной регрессии?

Линейная регрессия
Ну-ка вспомним

Как мы минимизировали ошибки в линейной регрессии?

Находили **минимум функции MSE с помощью градиентного спуска**



Линейная классификация

Но в классификации у нас не MSE, а какие-то отступы…

$$Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}[M_i < 0] \to min$$

Что здесь минимизировать? **Пороговую функцию потерь**

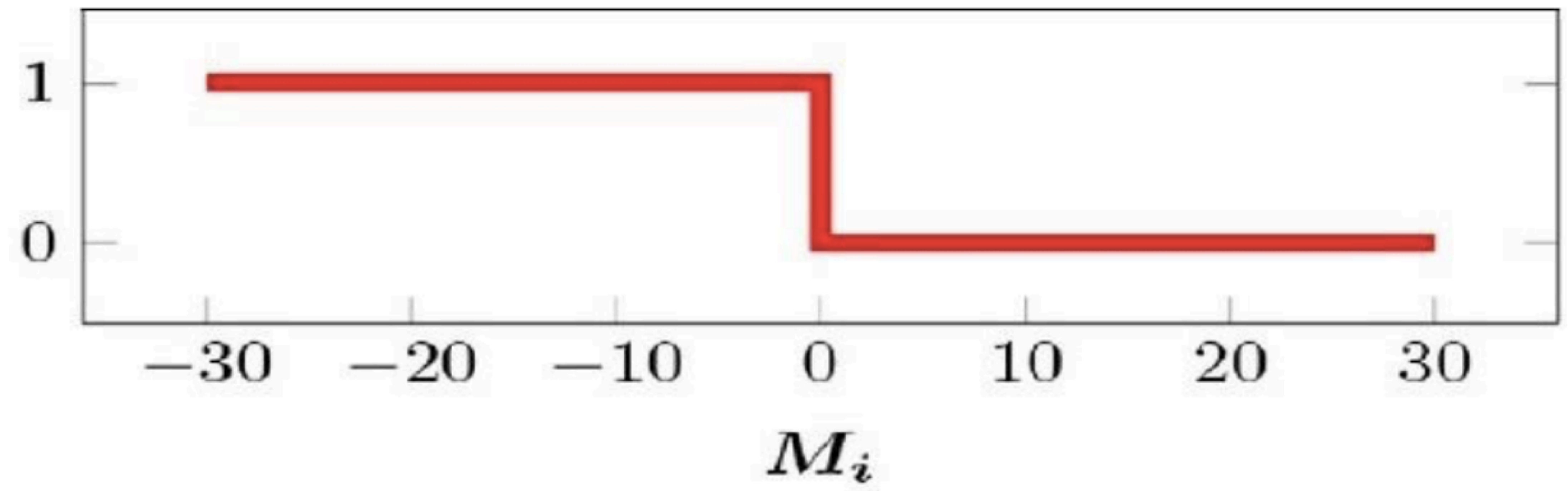
Линейная классификация

Но в классификации у нас не MSE, а какие-то отступы…

$$Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}[M_i < 0] \rightarrow min$$

Что здесь минимизировать? **Пороговую функцию потерь.**

Но она разрывна! Тут никакого градиентного спуска…



Функция потерь

Вместо пороговой функции потерь придется взять **другую, гладкую (непрерывную)**, чтобы от нее считать производные.

Взять может не любую функцию, а такую, которая будет ограничивать нашу функцию сверху

Функция потерь

Вместо пороговой функции потерь придется взять **другую, гладкую (непрерывную)**, чтобы от нее считать производные.

Взять может не любую функцию, а такую, которая будет ограничивать нашу функцию сверху

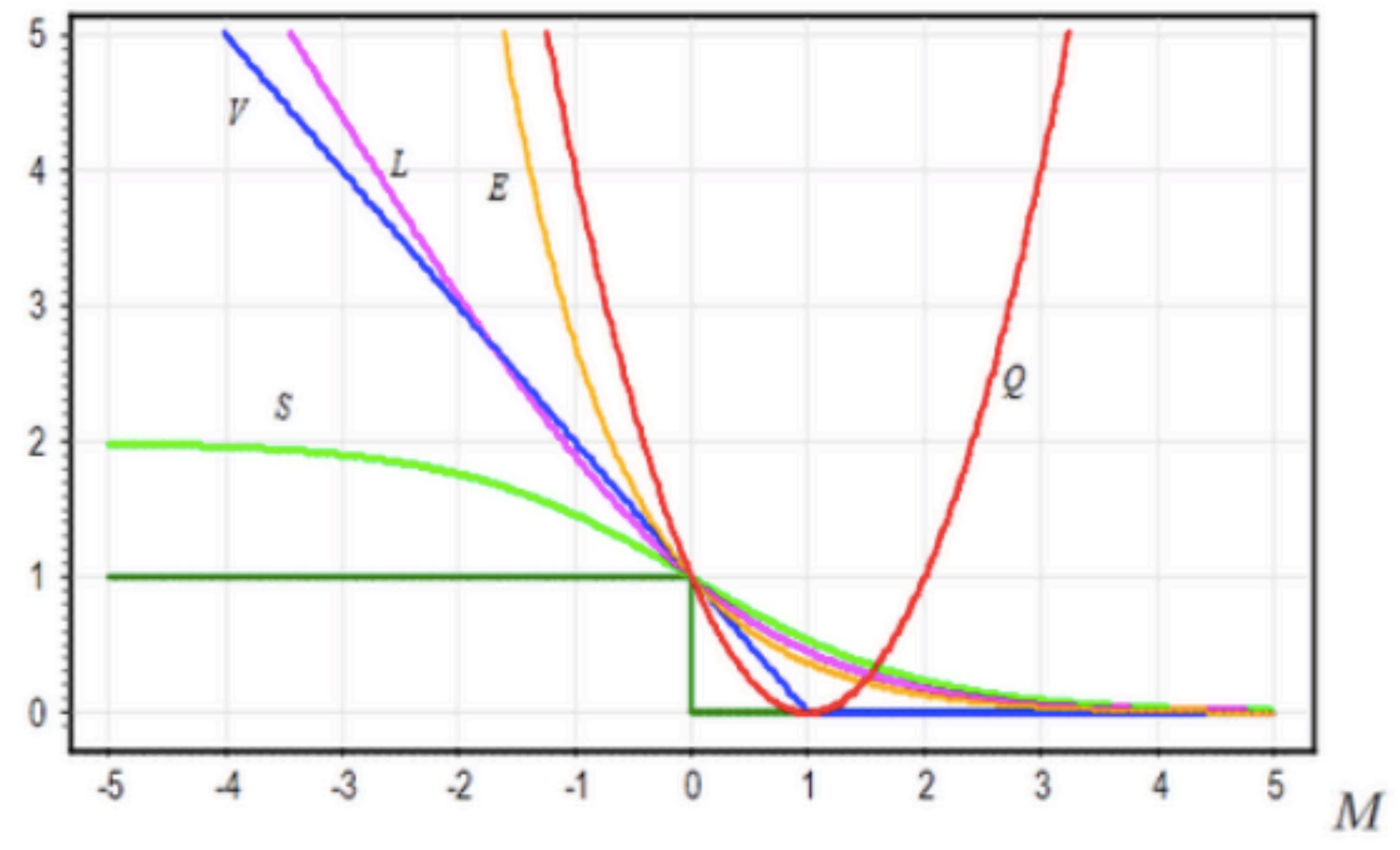


Рис. 7. Непрерывные аппроксимации пороговой функции потерь $[M < 0]$.

$Q(M) = (1 - M)^2$ — квадратичная;
$V(M) = (1 - M)_+$ — кусочно-линейная;
$S(M) = 2(1 + e^M)^{-1}$ — сигмоидная;
$L(M) = \log_2(1 + e^{-M})$ — логистическая;
$E(M) = e^{-M}$ — экспоненциальная.

Функция потерь

Вместо пороговой функции потерь придется взять **другую, гладкую (непрерывную)**, чтобы от нее считать производные.

Взять может не любую функцию, а такую, которая будет ограничивать нашу функцию сверху

В зависимости от взятой функции потерь мы получим **разные ответы**, это **разные классификаторы**.

Функция потерь

В зависимости от функции мы получаем какую-то модель.

Например, с помощью логистической функции потерь - логистическую регрессию

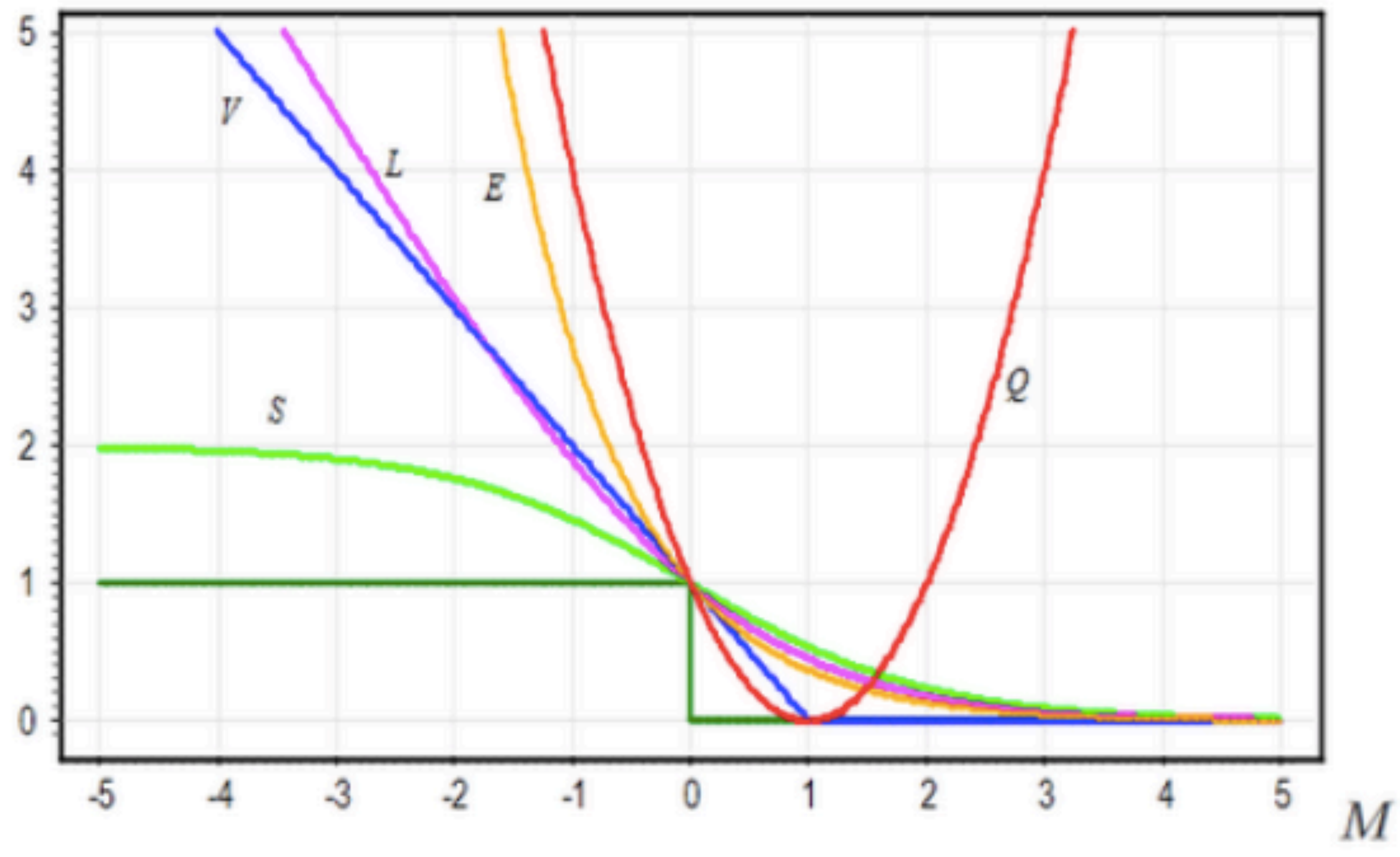


Рис. 7. Непрерывные аппроксимации пороговой функции потерь $[M < 0]$.

$$
\begin{aligned}
Q(M) &= (1 - M)^2 && \text{— квадратичная;} \\
V(M) &= (1 - M)_+ && \text{— кусочно-линейная;} \\
S(M) &= 2(1 + e^{M})^{-1} && \text{— сигмоидная;} \\
L(M) &= \log_2(1 + e^{-M}) && \text{— логистическая;} \\
E(M) &= e^{-M} && \text{— экспоненциальная.}
\end{aligned}
$$

Повторим

1. Мы не можем использовать для классификации функцию линейной регрессии. Вместо этого мы берем знак (sign)

$$a(\boldsymbol{x}, w) = \sum_{i=1}^{l} w_j x_j \longrightarrow \boldsymbol{a}(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{sign}\left(\sum_{j=1}^{l} \boldsymbol{w}_j \boldsymbol{x}_j\right)$$

2. Тогда наша функция потерь такая: $Q(a, X) = \frac{1}{n}\sum_{i=1}^{n} [\boldsymbol{M}_i < \boldsymbol{0}] \to min$

Но она пороговая.

Придется брать аппроксимирующую функцию.

Повторим

3. Аппроксимирующая функция должна ограничивать нашу пороговую сверху. Этому условию уовлетворяет несколько функций.

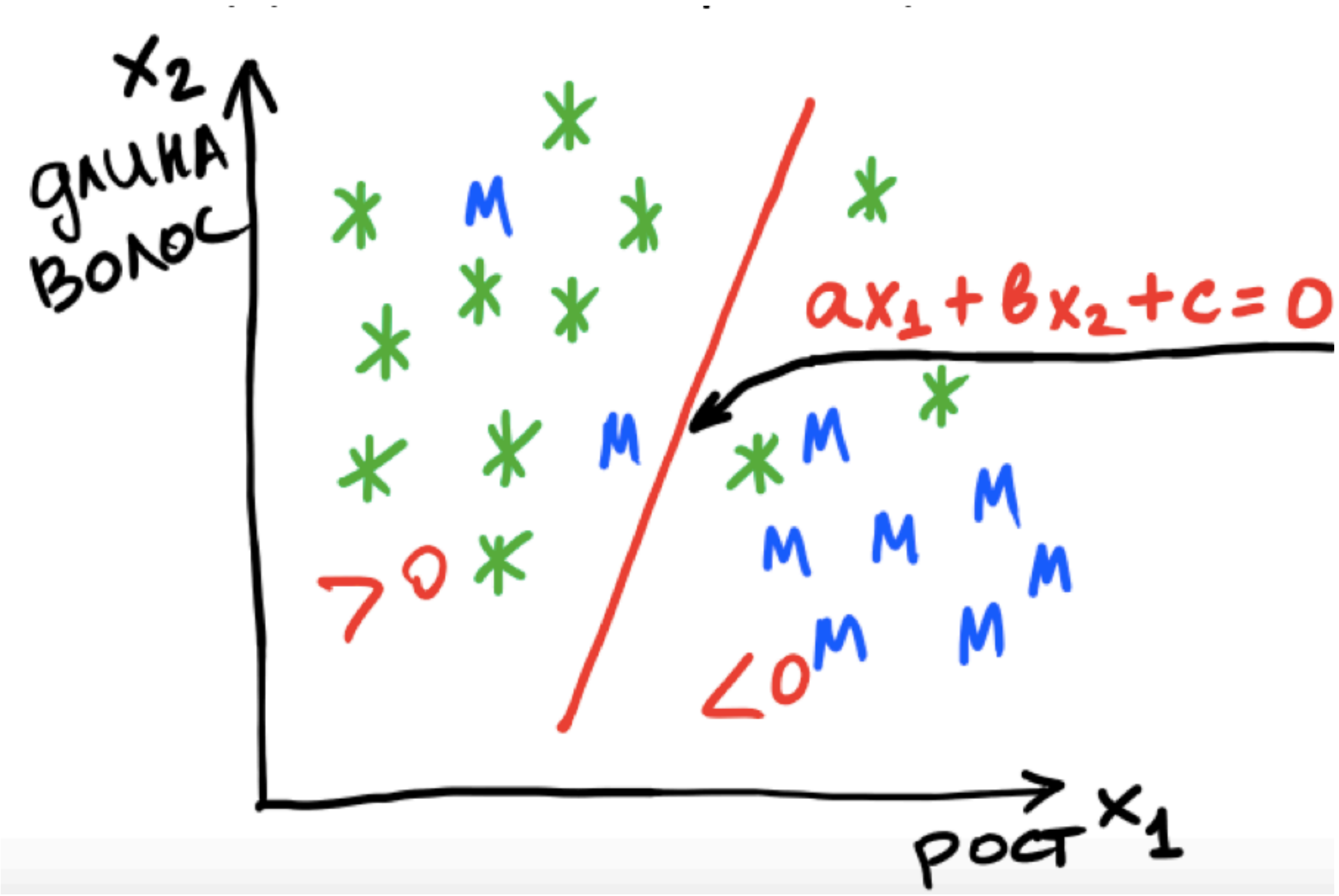
В зависимости от функции, мы получим разные ответы. След., разные классификаторы

4. Ура, классификатор обучен! Он умеет выдавать ответы для новых данных. **Как понять, хороший ли у нас классификатор?**

Метрики качества

Метрика качества классификации

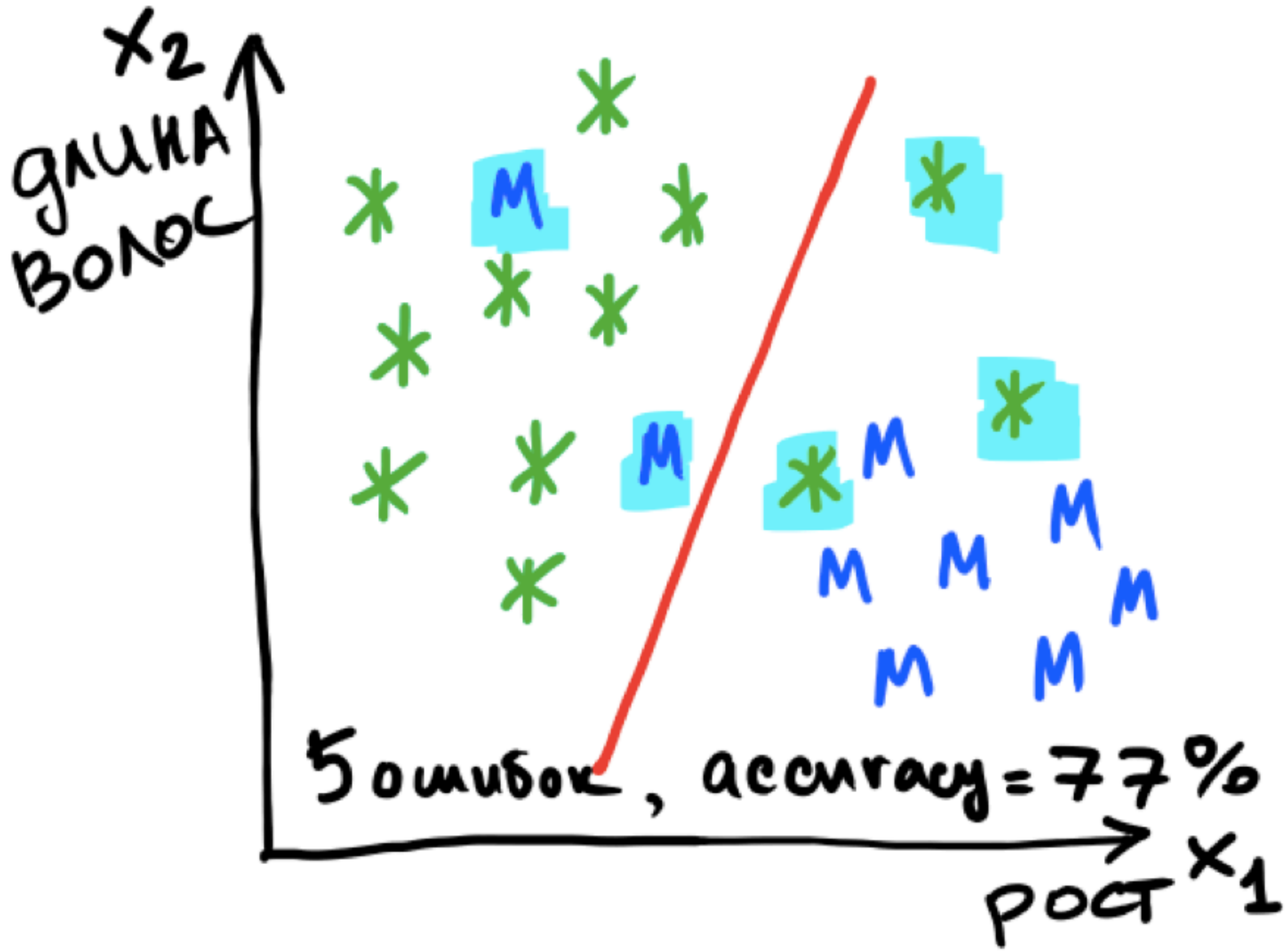
Accuracy



Просто посчитаем долю правильных ответов на новых данных (test)

Метрика качества классификации

Accuracy



Метрика качества классификации
Accuracy

Плюсы:

- просто считать
- очевидно

Минусы?

Метрика качества классификации

Accuracy

Плюсы:

- просто считать
- очевидно

Минусы:

При несбалансированных классах плохо отражает качество алгоритма.

Метрика качества классификации

Задачка

Пример: Кредитный скоринг.

Выборка из 200 клиентов, 100 вернули, 100 не вернули

Модель 1: одобряет 100 кредитов

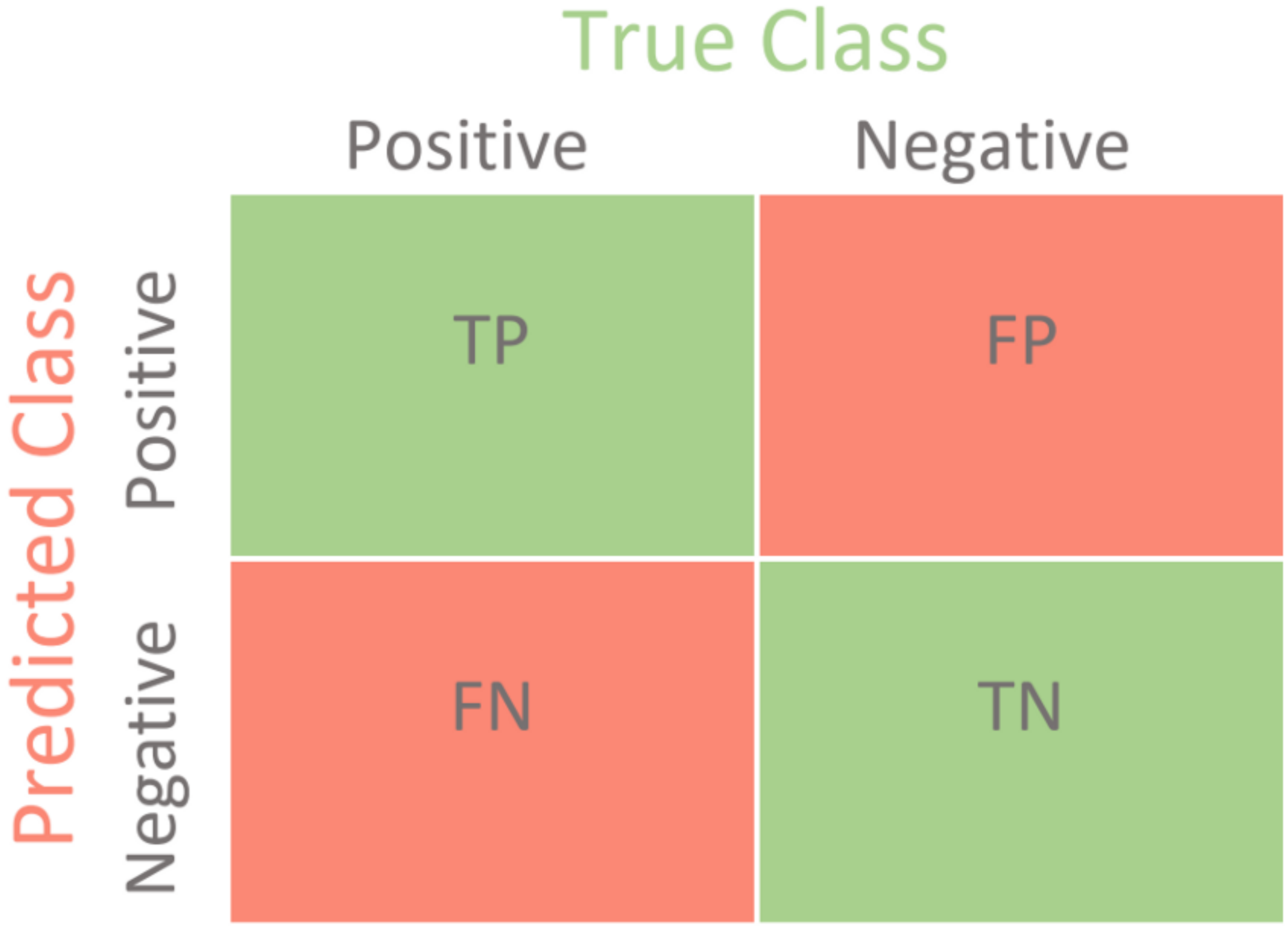
80 кредитов вернули
20 кредитов не вернули

Какая лучше?

Модель 2: одобряет 50 кредитов

48 кредитов вернули
2 кредита не вернули

Матрица ошибок



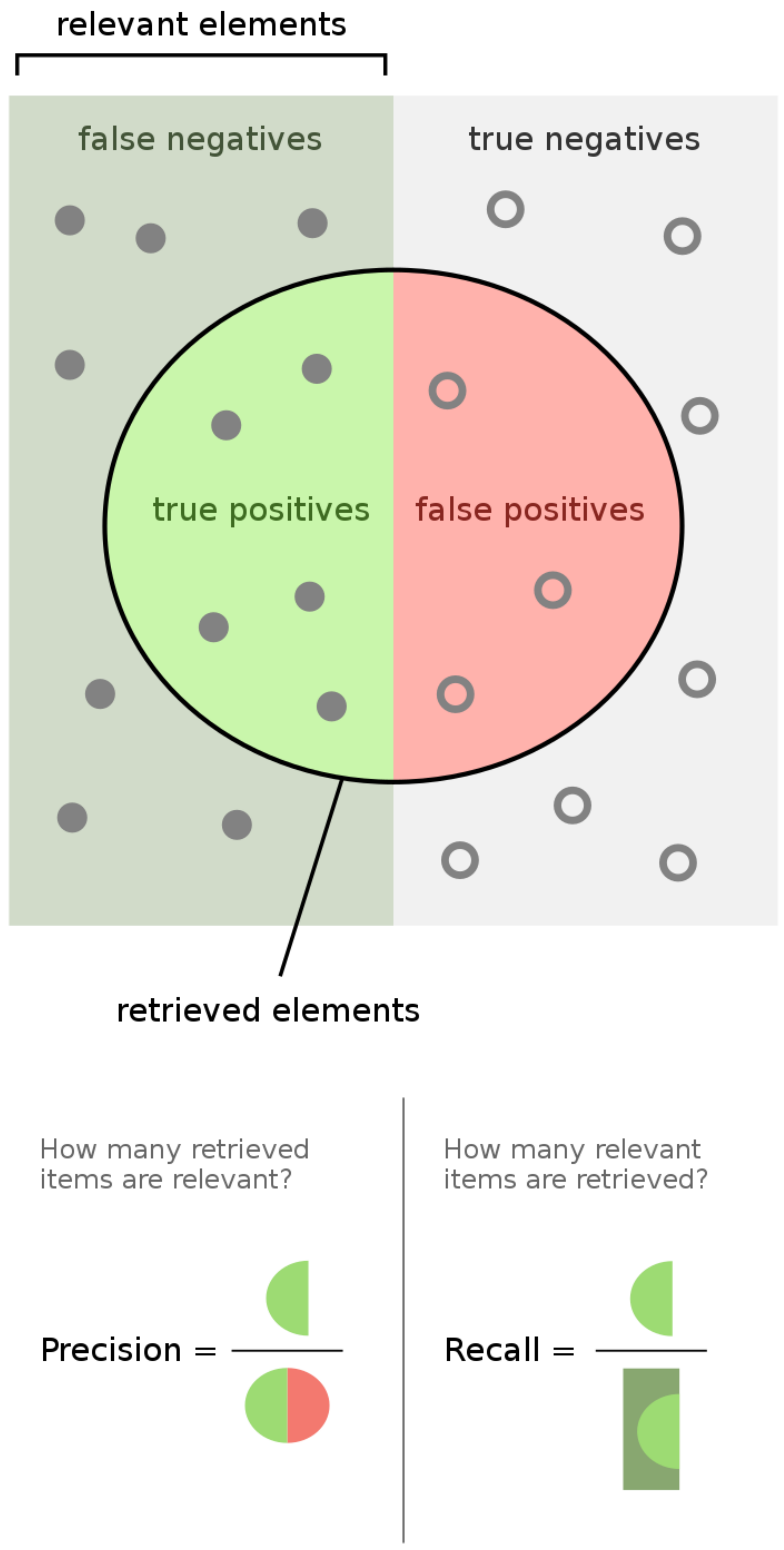
Precision, Recall

Точность (precision):

$$Precision(a, X) = \frac{TP}{TP + FP}$$

Полнота (recall):

$$Recall(a, X) = \frac{TP}{TP + FN}$$



Precision

Модель $a_1(x)$:

$$\text{precision}(a_1, X) = 0.8$$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ Не получили кредит	20	80

Модель $a_2(x)$:

$$\text{precision}(a_2, X) = 0.96$$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	48	2
$\mathbf{a(x) = -1}$ Не получили кредит	52	98

Recall

Модель $a_1(x)$:

$\mathrm{rec}all(a_1, X) = 0.8$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	80	20
$\mathbf{a(x) = -1}$ Не получили кредит	20	80

Модель $a_2(x)$:

$\mathrm{recall}(a_2, X) = 0.48$

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$\mathbf{a(x) = 1}$ Получили кредит	48	2
$\mathbf{a(x) = -1}$ Не получили кредит	52	98

F-мера

F-мера учитывает и точность, и полноту

$$F(a, X) = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$