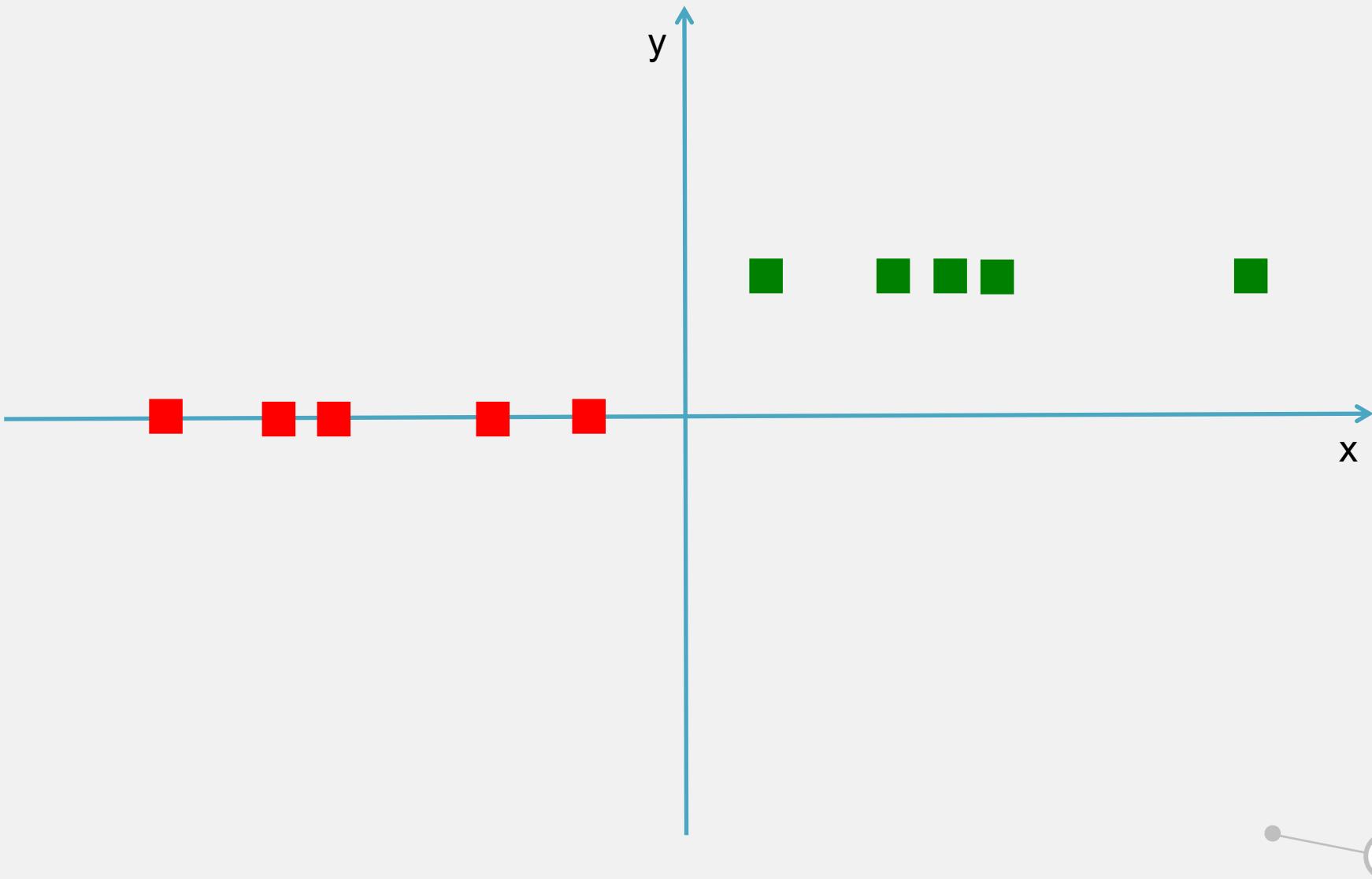
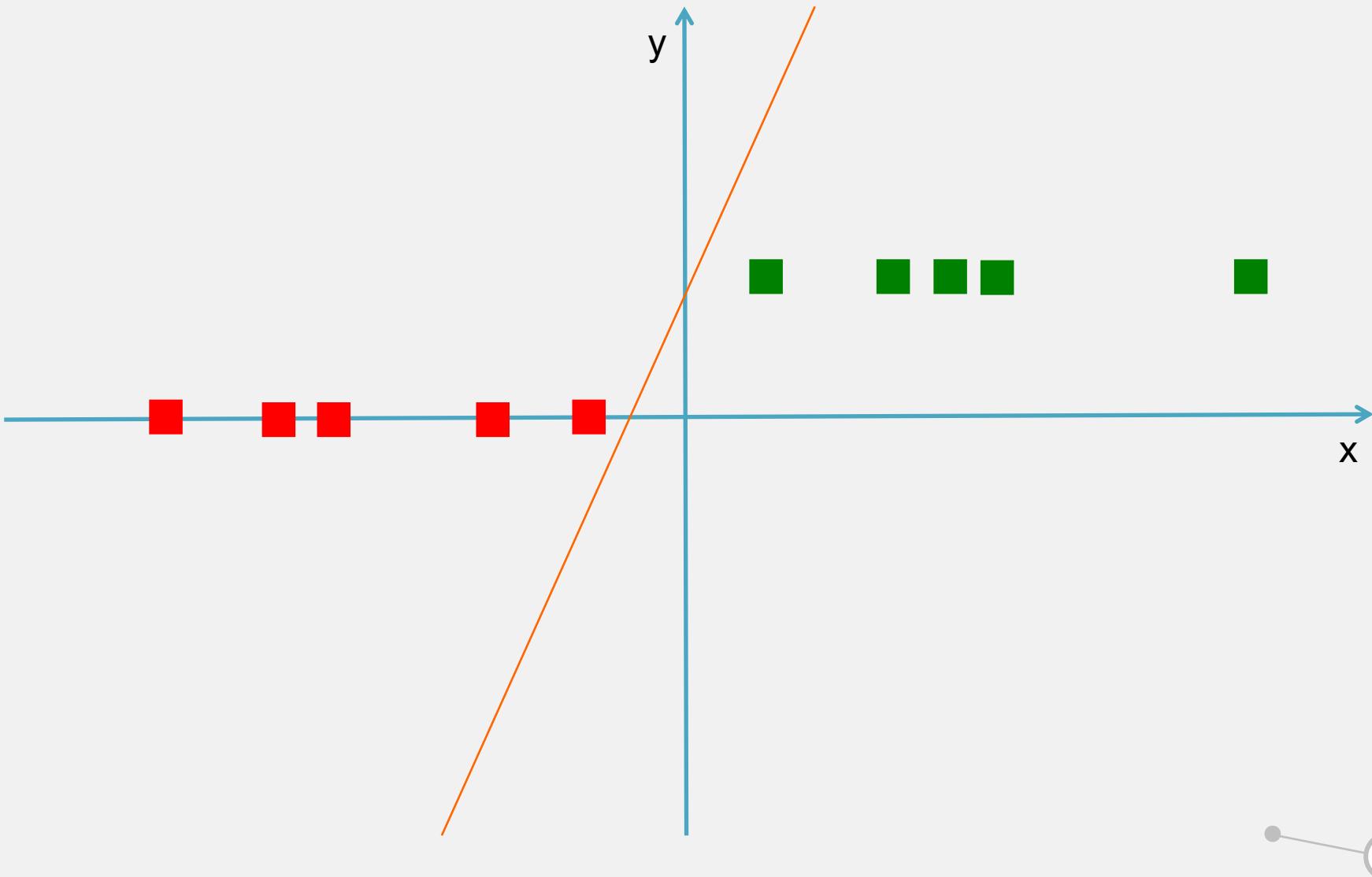


Задачи классификации. Логистическая регрессия

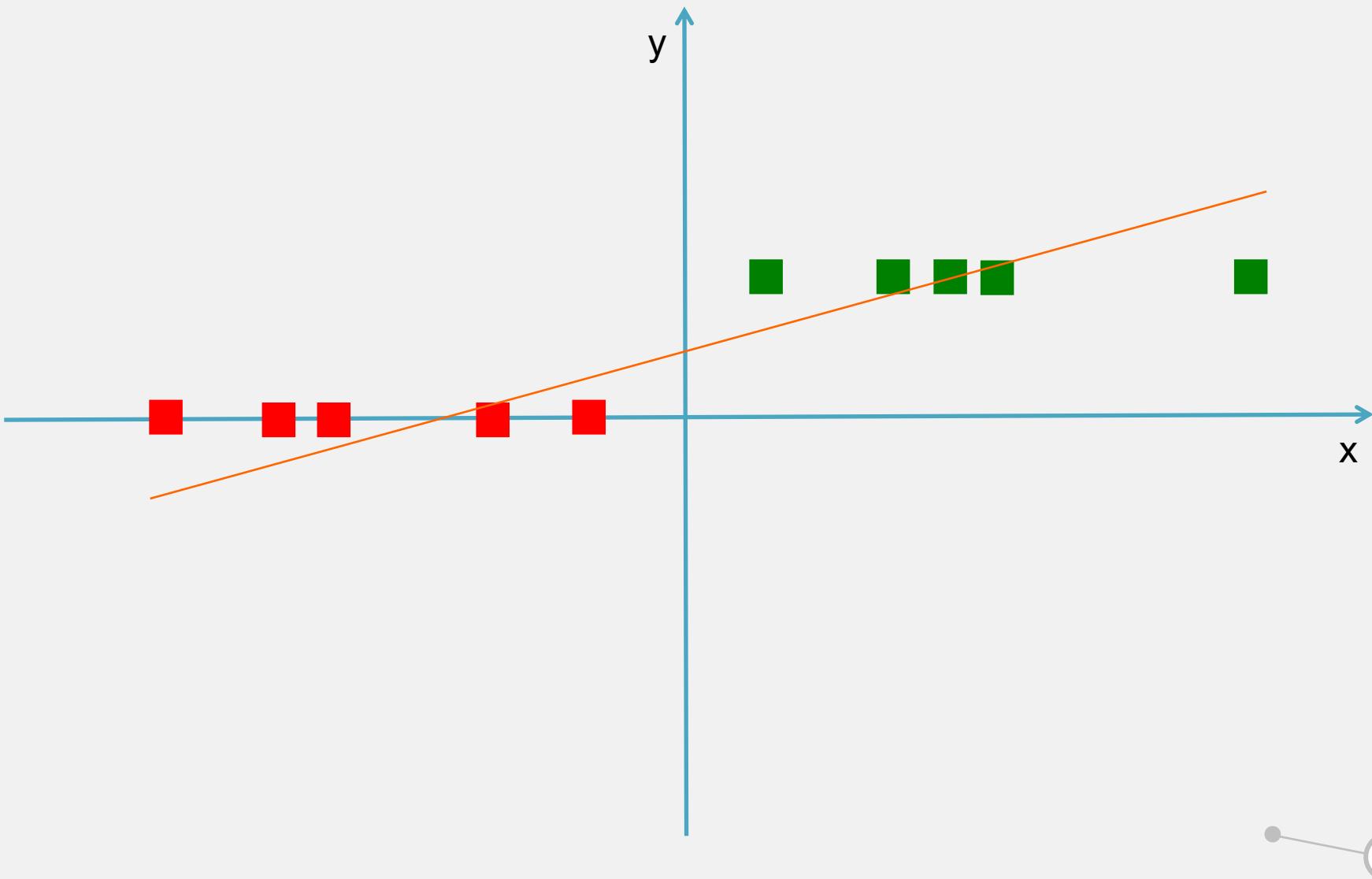
Задача классификации



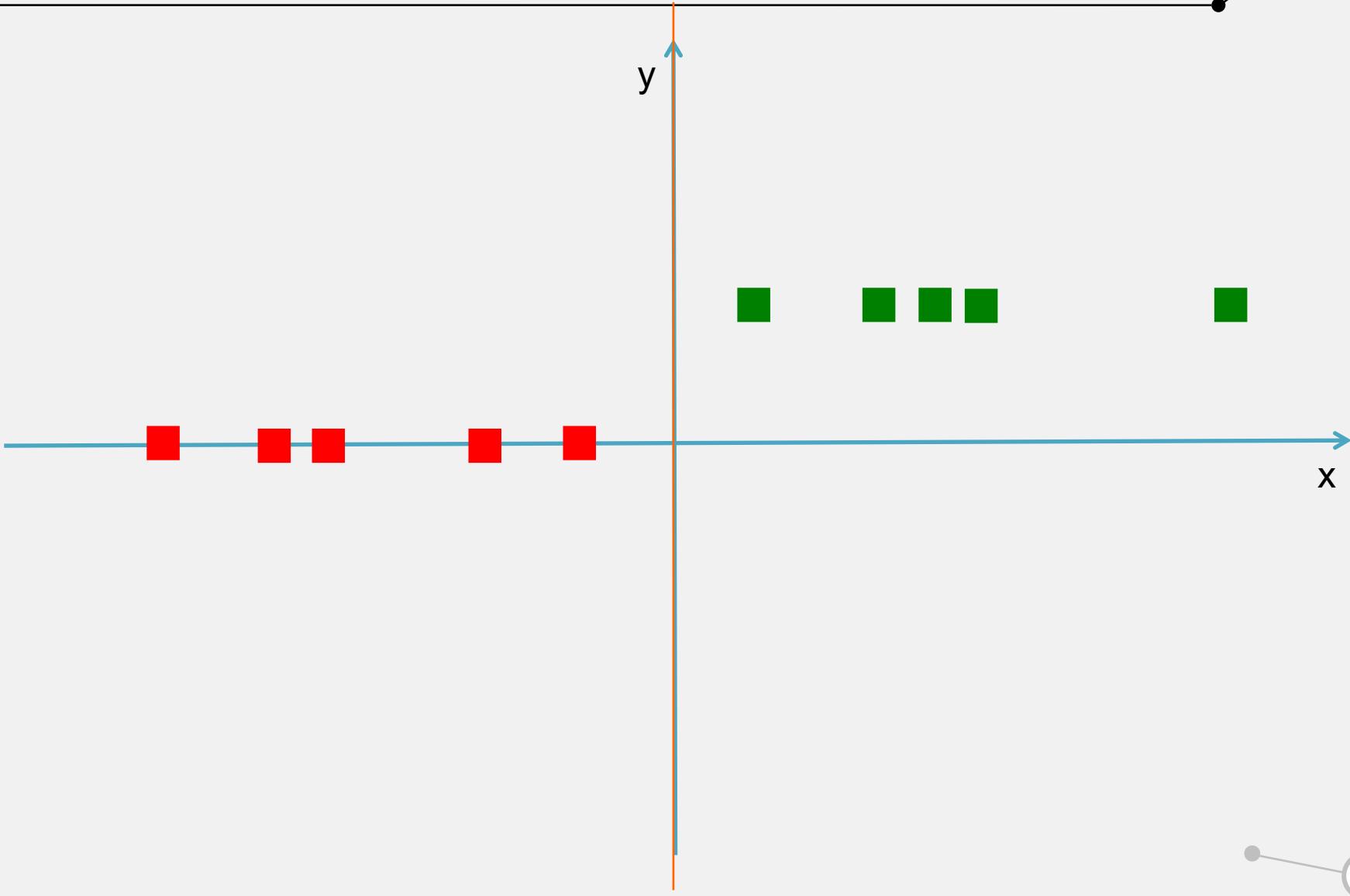
Задача классификации. Линейная модель



Задача классификации. Линейная модель



Задача классификации. Линейная модель



Построение разделяющей поверхности



- Задача классификации на 2 класса $Y=\{1, -1\}$
- Обучающая выборка $X=(x_i, y_i), i=1, L$
- Построить алгоритм классификации $a(x, w)=\text{sign } f(x, w)$

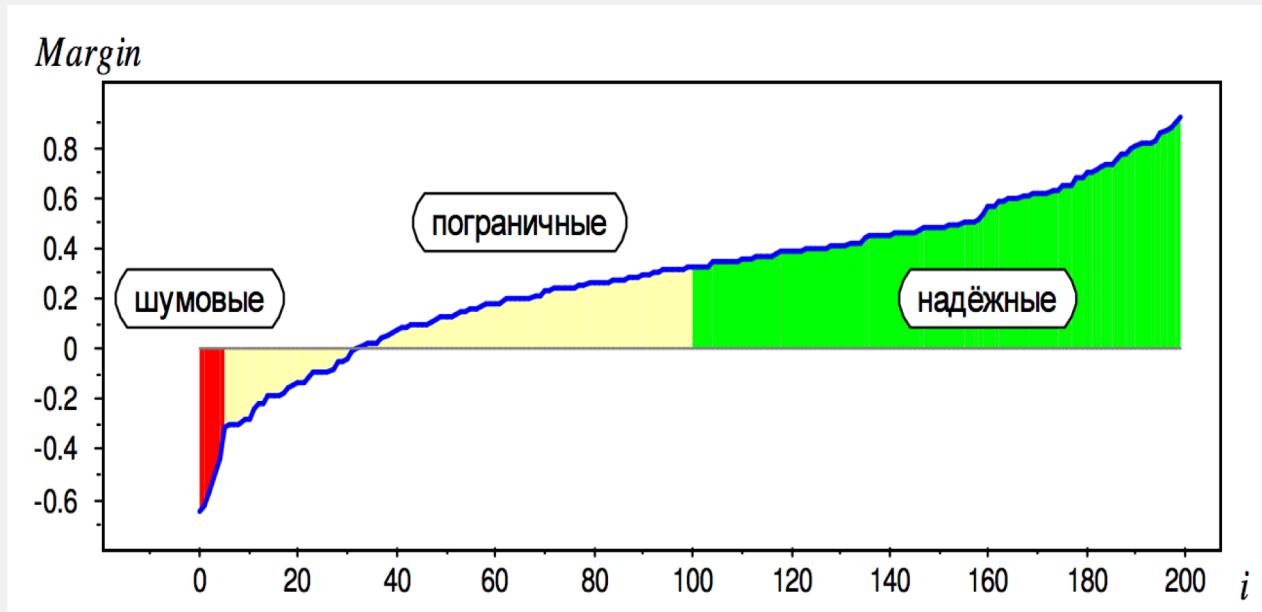
$f(x, w)=0$ – разделяющая поверхность



Отступ



- $f(x, w) = 0$ – разделяющая поверхность
- $M_i(w) = y_i f(x_i, w)$ – отступ объекта i (Margin)
- $M_i(w) < 0 \Rightarrow$ ошибка алгоритма а на объекте i



Функционал эмпирического риска



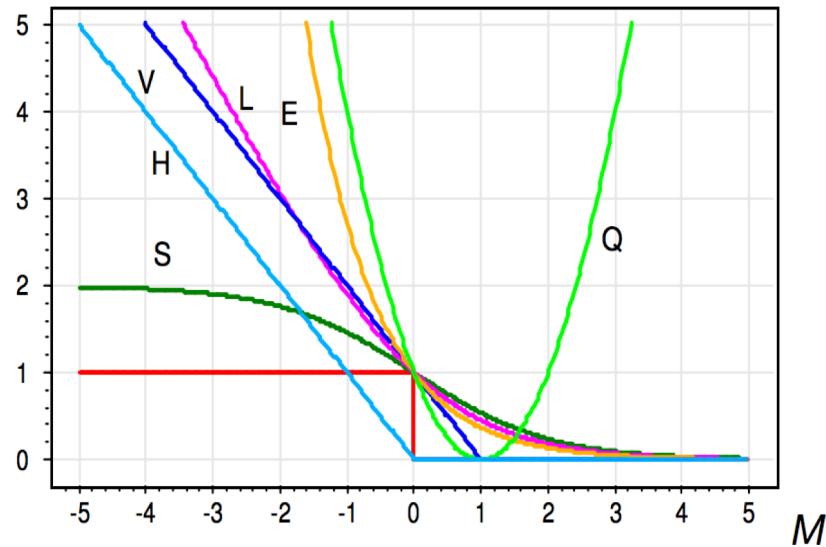
- $M_i(w) < 0 \Rightarrow$ ошибка алгоритма а на объекте i
- Эмпирический риск
 - $Q(w) = \sum_i^L [M_i(w) < 0]$
 - Гладкая аппроксимация $Q \leq Q'(w) = \sum_i^L L(M_i(w))$
 - $Q \rightarrow \min$



Апроксимации



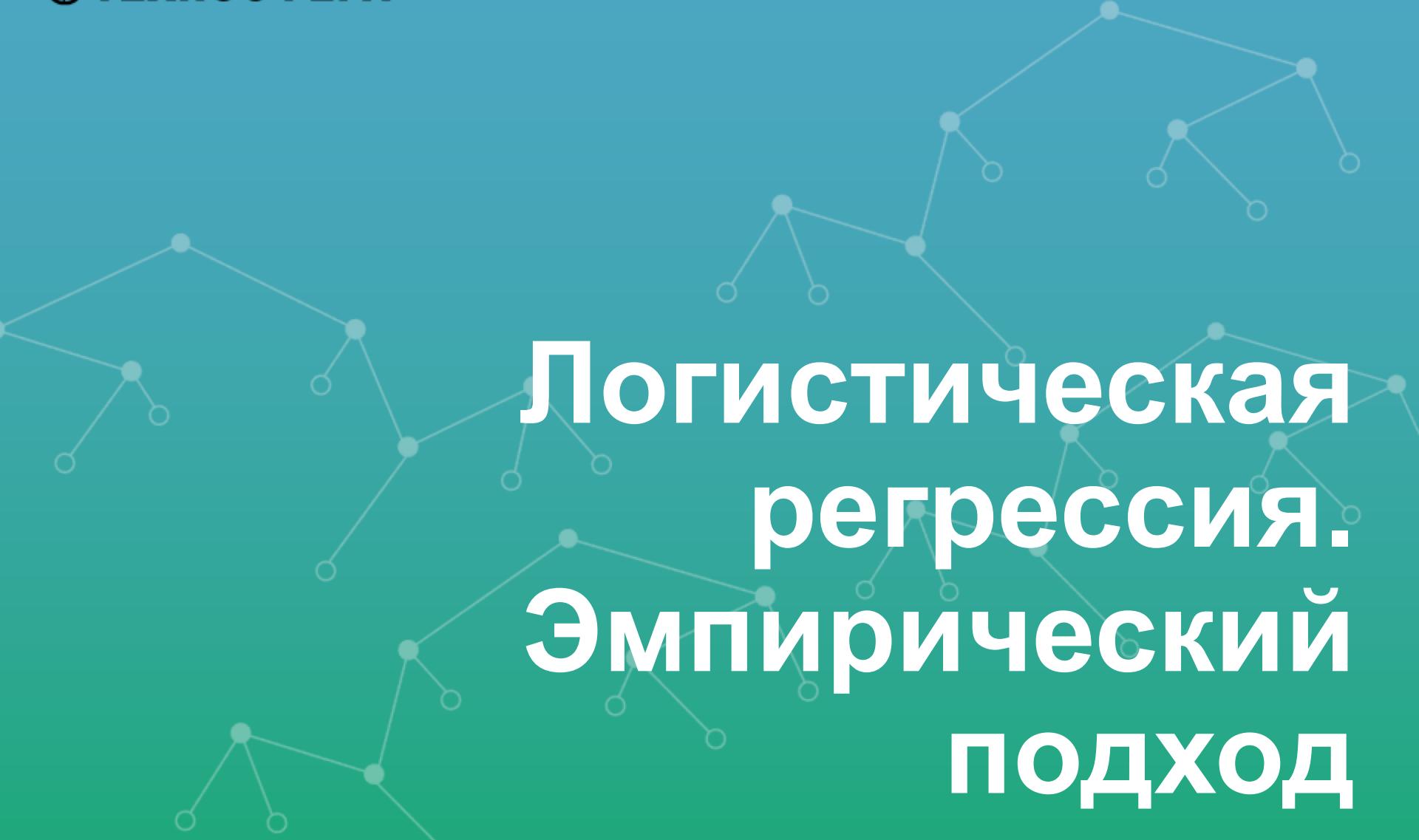
Часто используемые непрерывные функции потерь $\mathcal{L}(M)$:



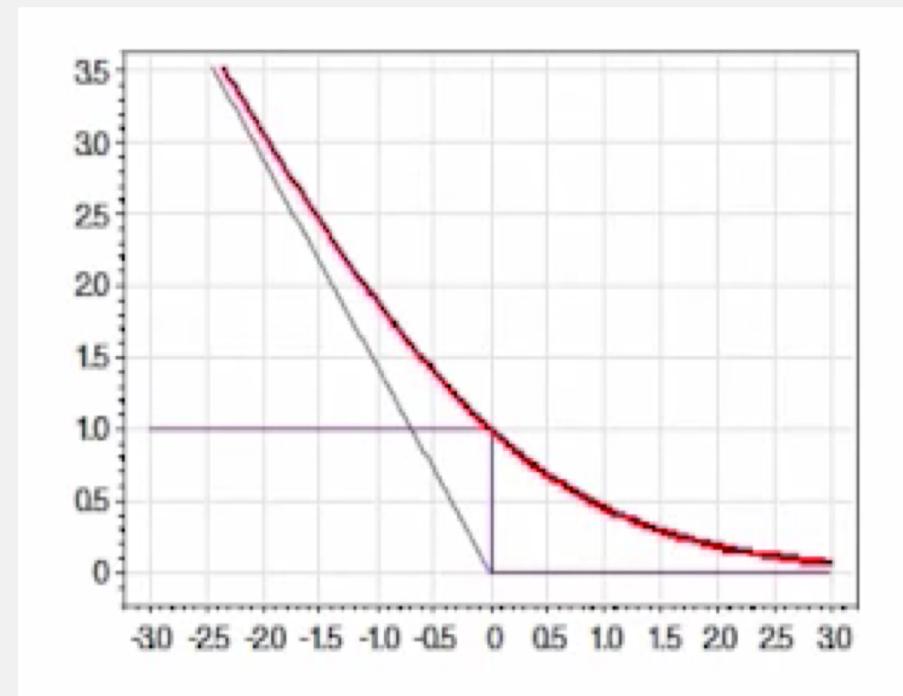
- | | |
|-----------------------------|-----------------------------------|
| $V(M) = (1 - M)_+$ | — кусочно-линейная (SVM); |
| $H(M) = (-M)_+$ | — кусочно-линейная (Hebb's rule); |
| $L(M) = \log_2(1 + e^{-M})$ | — логарифмическая (LR); |
| $Q(M) = (1 - M)^2$ | — квадратичная (FLD); |
| $S(M) = 2(1 + e^M)^{-1}$ | — сигмоидная (ANN); |
| $E(M) = e^{-M}$ | — экспоненциальная (AdaBoost); |



Логистическая регрессия. Эмпирический подход



Апроксимация логистической регрессии



$$L(M) = \log(1+e^{-x})$$



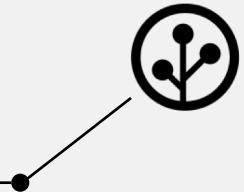
Логистическая регрессия на 1 слайде



- $L = \log(1+e^{-M})$
- $Q'(w) = \sum_i^L \log(1 + \exp(-M_i)) \rightarrow \min_w$
- Оптимизация градиентным спуском
- $a(x_i, w_i) = \text{sign}(x_i, w_i)$



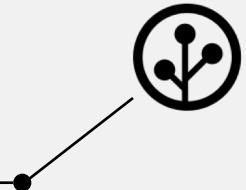
Логистическая регрессия на 1 слайде



- $L = \log(1+e^{-M})$
- $Q'(w) = \sum_i^L \log(1 + \exp(-M_i)) \rightarrow \min_w$
- Оптимизация градиентным спуском
- $a(x_i, w_i) = \text{sign}(x_i, w_i)$



Логистическая регрессия на 1 слайде



- $L = \log(1+e^{-M})$
- $Q'(w) = \sum_i^L \log(1 + \exp(-M_i)) \rightarrow \min_w$
- Оптимизация градиентным спуском
- $a(x_i, w_i) = \text{sign}(x_i, w_i)$
- Какое основание у логарифма ?



Сигмоида

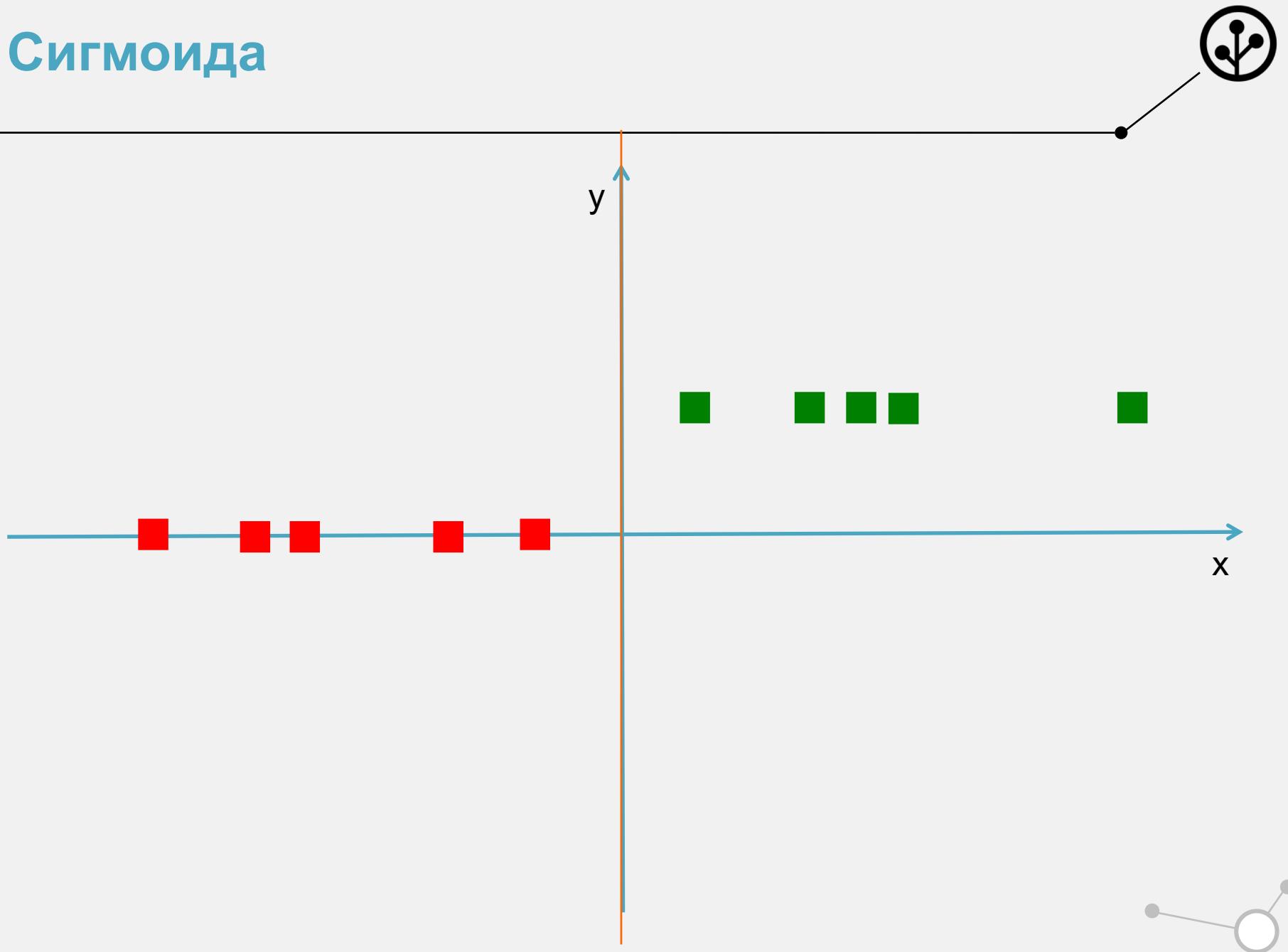


- Посмотрим на функцию

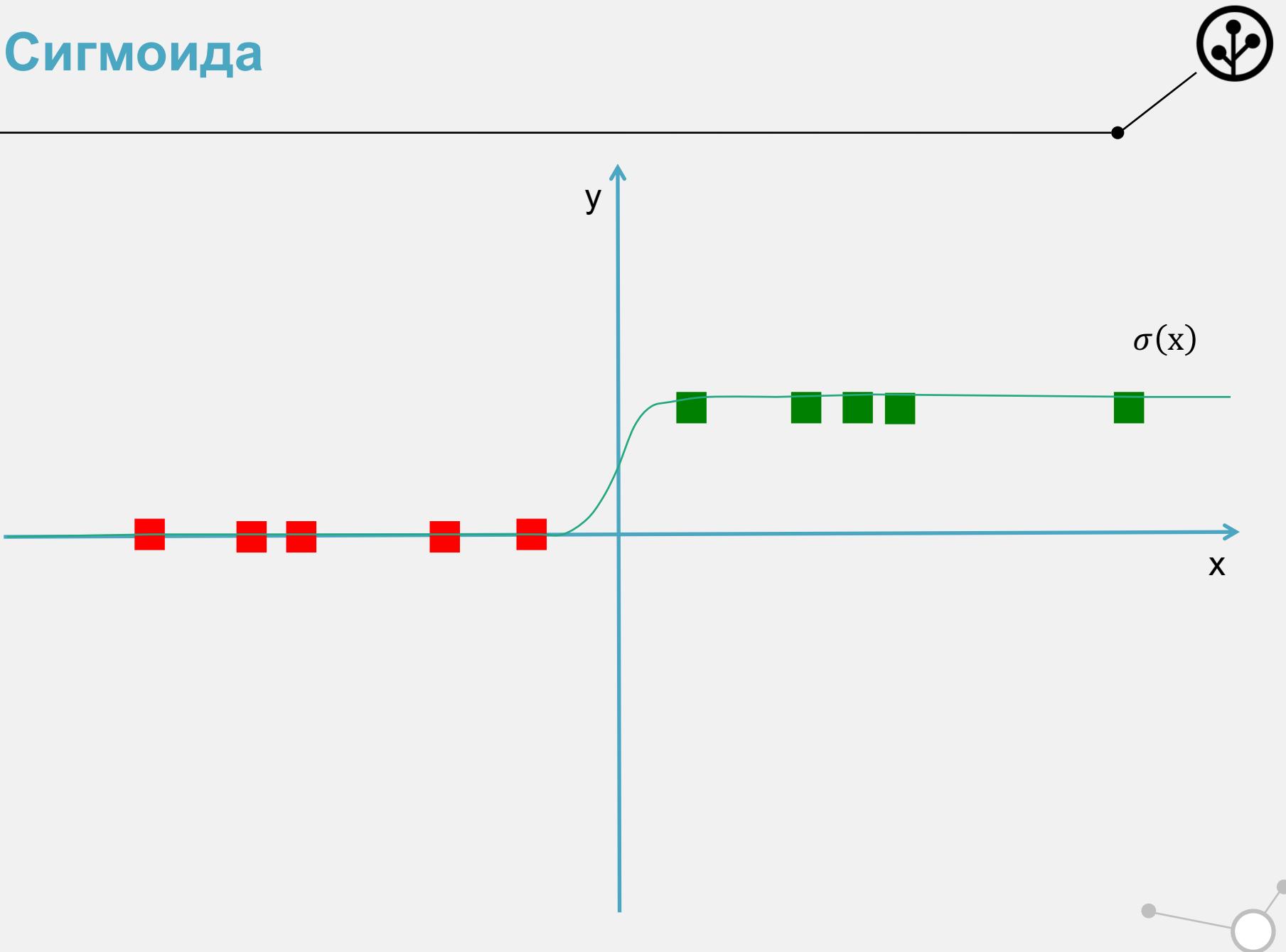
- $\sigma(M) = \frac{1}{1+\exp(-M)}$



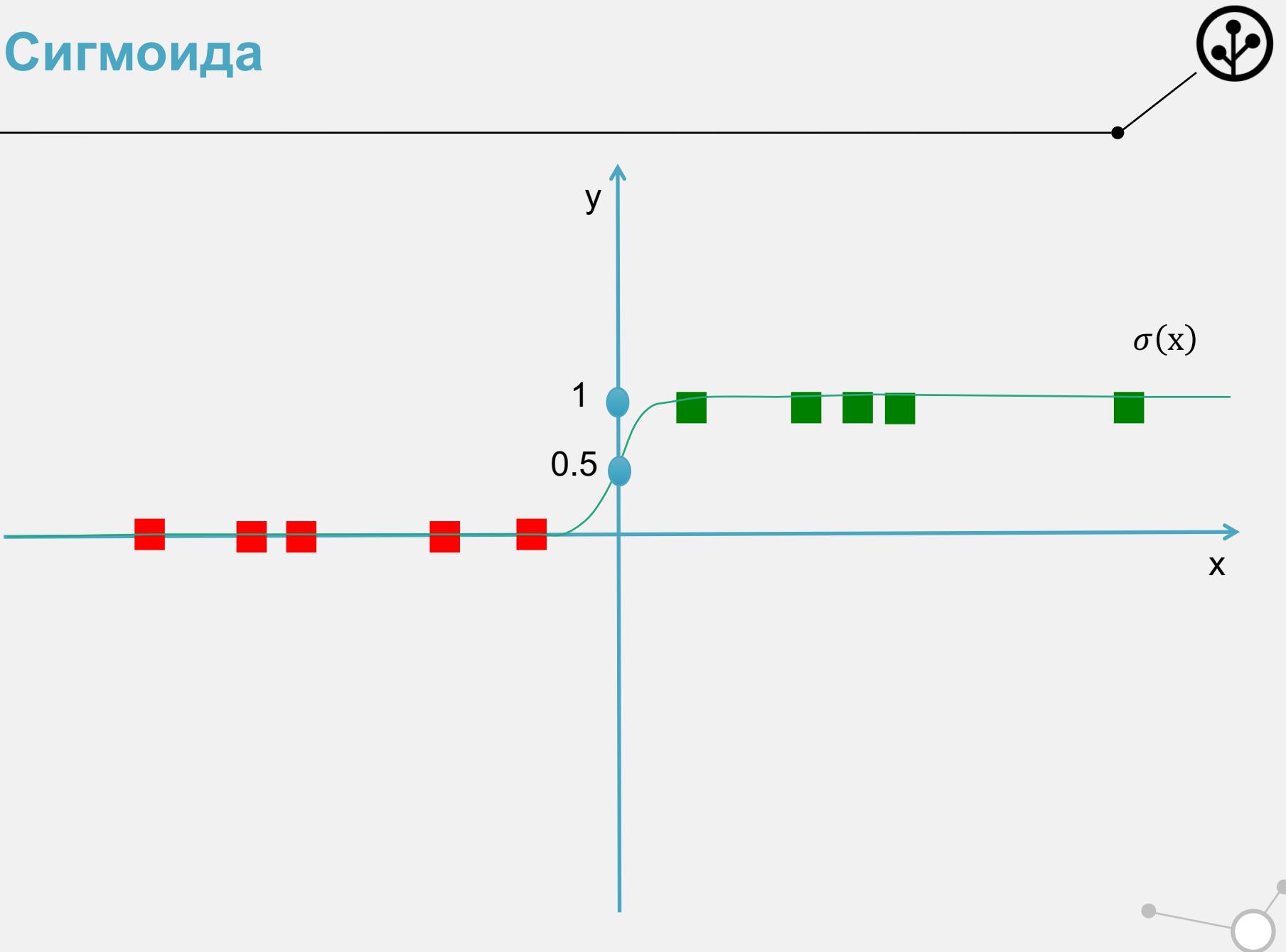
Сигмоида



Сигмоида



Сигмоида



Сигмоида



- Посмотрим на функцию
 - $\sigma(M) = \frac{1}{1+\exp(-M)}$
- Ассимптоты
 - $\sigma(-\infty) = 0$
 - $\sigma(+\infty) = 1$
 - $\sigma(M) = 0.5$
- Looks like вероятность !
- $p(M_i) = \frac{1}{1+\exp(-M_i)}$ - "вероятность правильной классификации объекта I"

Сигмоида



- $Q'(w) = - \sum_i^L \log(p_i) \rightarrow \min_w$
- $Q'(w) = \sum_i^L \log(p_i) \rightarrow \max_w$ – максимизируем 'вероятности' правильных классификаций
- $a(x_i, w_i) = \text{sign}(x_i, w_i) \leftrightarrow a(x_i, w_i) = \text{sign}\left(\frac{1}{1 + \exp(-(x_i, w_i))} - 0.5\right)$
- *Можем даже уточнить алгоритм*
 - $a(x_i, w_i) = \text{sign}\left(\frac{1}{1 + \exp(-(x_i, w_i))} - \text{th}\right)$



Логистическая регрессия



- Другая формализация – $Y = \{0,1\}$
- $Q^{(w)} = - \sum_i^L y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$
- Чаще всего это называют log-loss



Логистическая регрессия



- На этом вроде бы и все
 - Сформировали лосс
 - Умеем оптимизировать
 - Умеем принимать решение
- Но слишком много мы просто приняли на веру
- Почему и откуда именно?
 - $L = \log(1+e^{-M})$
 - $\sigma(M) = \frac{1}{1+\exp(-M)}$

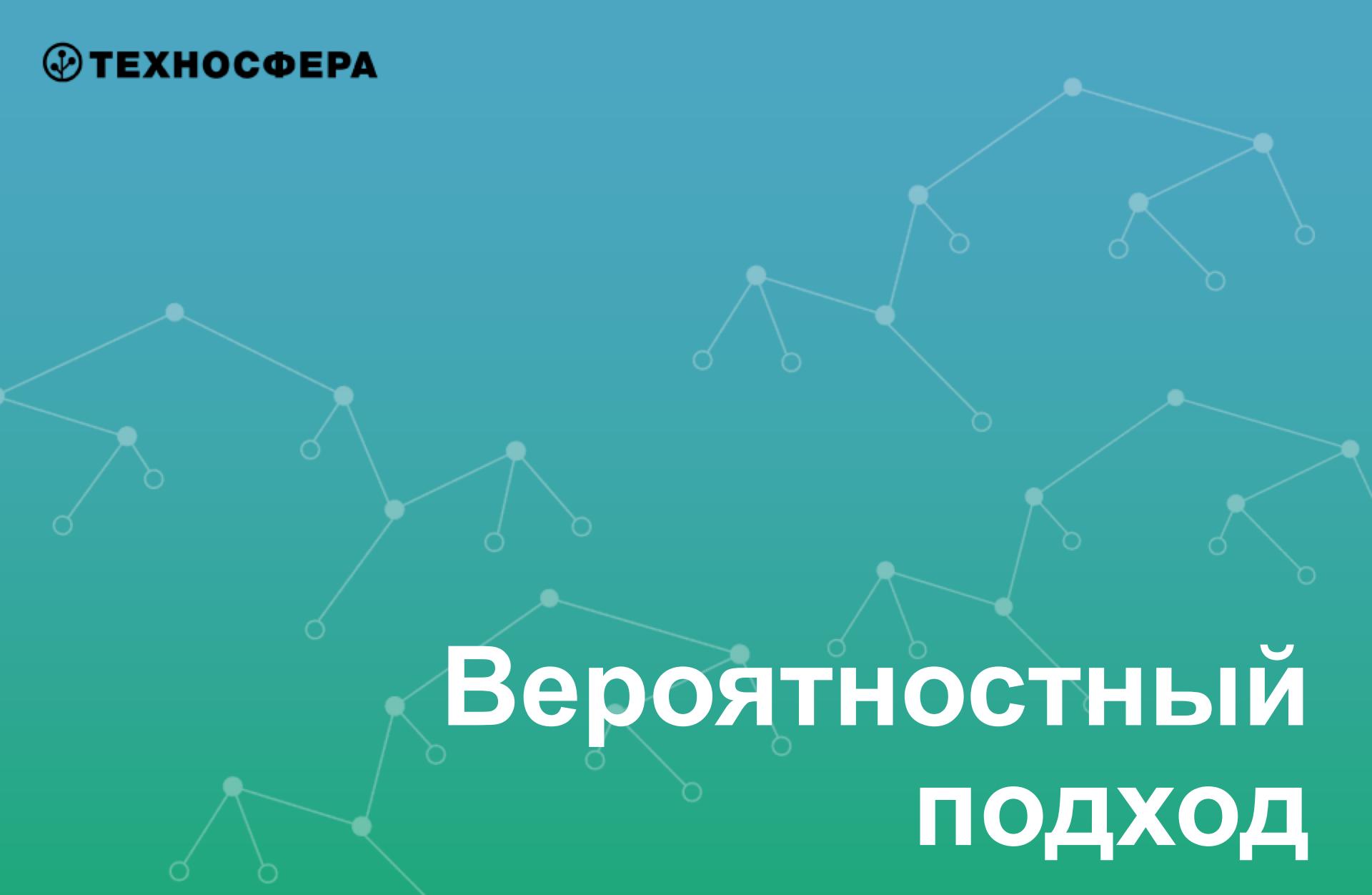


Логистическая регрессия



- На этом вроде бы и все
 - Сформировали лосс
 - Умеем оптимизировать
 - Умеем принимать решение
- Но слишком много мы просто приняли на веру
- Почему и откуда именно?
 - $L = \log(1+e^{-M})$
 - $\sigma(M) = \frac{1}{1+\exp(-M)}$





Вероятностный подход

Math Recap: Метод Максимального Правдоподобия



- Пусть \mathbf{X} – вероятностное пространство с плотностью $p(x|\theta)$
 - θ – параметры распределения
- Примеры
 - $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-u)^2}{2\sigma^2}}$; $\theta_1=u$; $\theta_2=\sigma$
 - $p(x) = e^{-\lambda x}$; $\theta=\lambda$
- Имеем выборку X *iid* реализаций из \mathbf{X}
- Задача – оценить параметры θ по выборке
 - ММП – наиболее правдоподобные ;)



Math Recap: Метод Максимального Правдоподобия



- Введем функцию правдоподобия
 - $L(X) = \prod_i^L p(x_i | \theta)$
 - Совместная вероятность наблюдать каждый из элементов I
 - Перемножаем, т.к. iid
- Метод максимального правдоподобия
 - $\theta' = \arg \max_{\theta} L(X)$
- Вероятность выпадения орла методом ММП?



ММП для нашей задачи



- Веса признаков – параметры нашей задачи
 - $p(x|\theta) \rightarrow p(x|w)$
- Оценить наиболее вероятные веса по выборке
- Решение – методом ММП
 - $w' = \arg \max_w L(X)$
 - $L(X) = \prod_i^L p(x_i|w) \rightarrow \max$
- $L(X) \rightarrow \max \Rightarrow \log L(x) \rightarrow \max$



ММП для нашей задачи



- $L(X) \rightarrow \max \Rightarrow \log L(x) \rightarrow \max$
- $\log \prod_i^L p(x_i|w) = \sum_i^L \log p(x_i|w) \rightarrow \max$
- *Ничего не напоминает ?*



ММП для нашей задачи



- $L(X) \rightarrow \max \Rightarrow \log L(x) \rightarrow \max$
- $\log \prod_i^L p(x_i|w) = \sum_i^L \log p(x_i|w) \rightarrow \max$
- *Ничего не напоминает ?*



Recap: Метрическая логика



- $Q'(w) = - \sum_i^L \log(p_i) \rightarrow \min_w$
- $Q'(w) = \sum_i^L \log(p_i) \rightarrow \max_w$ – максимизируем вероятности правильных классификаций
- $a(x_i, w_i) = \text{sign}(x_i, w_i) \leftrightarrow a(x_i, w_i) = \text{sign}\left(\frac{1}{1 + \exp(-(x_i, w_i))} - 0.5\right)$
- *Можем даже уточнить алгоритм*
 - $a(x_i, w_i) = \text{sign}\left(\frac{1}{1 + \exp(-(x_i, w_i))} - th\right)$



ММП для нашей задачи



- $L(X) \rightarrow \max \Rightarrow \log L(x) \rightarrow \max$
- $\log \prod_i^L p(x_i|w) = \sum_i^L \log p(x_i|w) \rightarrow \max$
- *Ничего не напоминает ?*
- Обосновали log-Loss !



Порождающая модель $p(x)$



Осталось определить $p(x_i|w)$

- В данной постановке x_i – пара объект-ответ (x_i, y_i)
- Пусть признаки – только бинарные $\{0, 1\}$
 - Тогда порождающее распределение – Бернулли
 - Можно доказать, что для Бернулли
 - $\frac{p(x, y=+1|w)}{p(x, y=-1|w)} = \exp(w, x)$
 - Доказательство (beyond the scope)



Порождающая модель $p(x)$



Осталось определить $p(x_i|w)$

- $\frac{p(x,y=+1|w)}{p(x,y=-1|w)} = \exp(w, x)$
- Из полной вероятности
 - $p(x, y = +1|w) + p(x, y = -1|w) = 1$
- $p(x, y = +1|w) = \frac{1}{1+\exp\{-(w,x)\}}; p(x, y = -1|w) = \frac{1}{1+\exp\{(w,x)\}}$



Порождающая модель $p(x)$

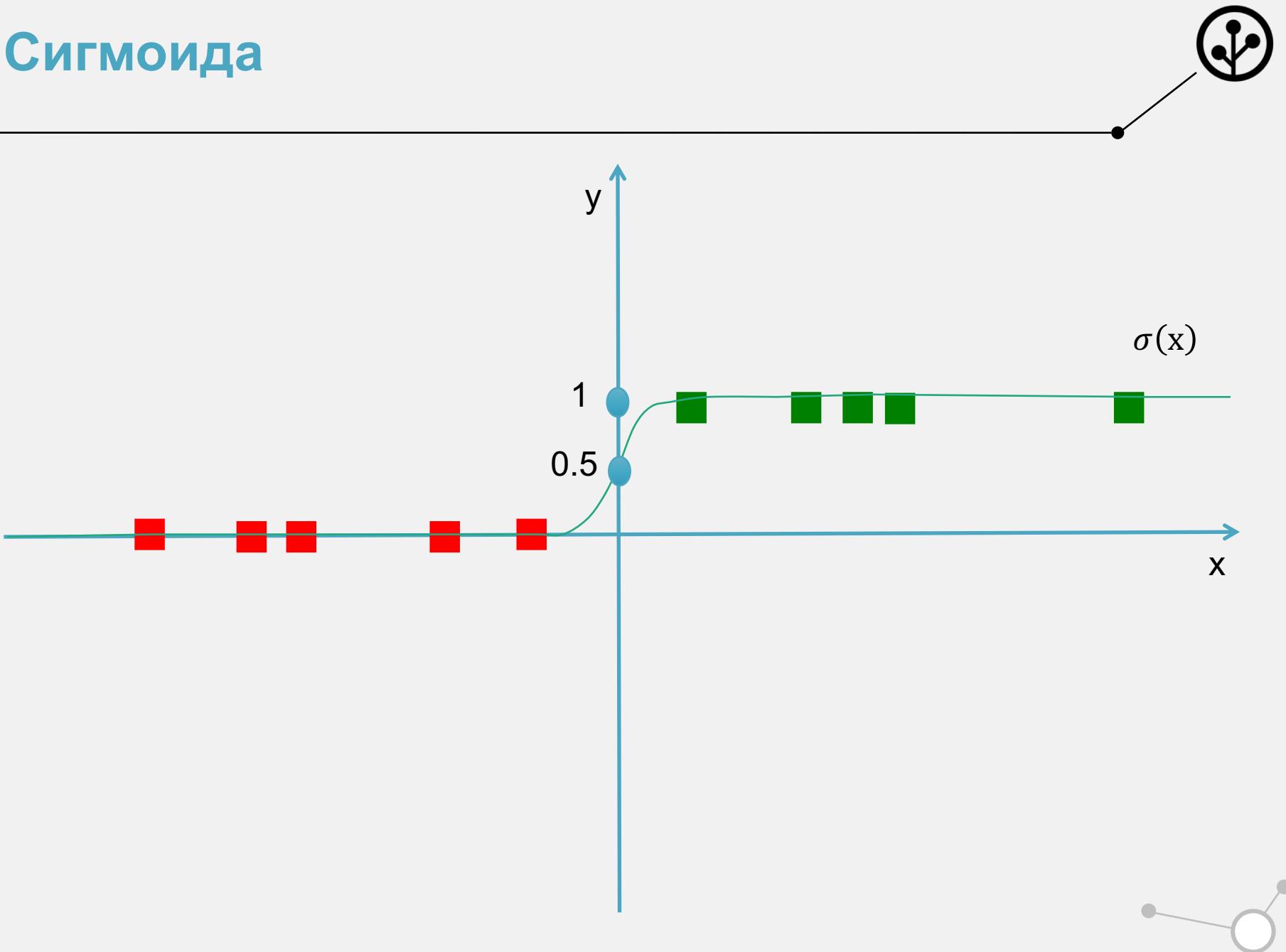


Осталось определить $p(x_i|w)$

- $\frac{p(x,y=+1|w)}{p(x,y=-1|w)} = \exp(w, x)$
- Из полной вероятности
 - $p(x, y = +1|w) + p(x, y = -1|w) = 1$
- $p(x, y = +1|w) = \frac{1}{1+\exp\{-(w,x)\}}; p(x, y = -1|w) = \frac{1}{1+\exp\{(w,x)\}}$
- Ничего не напоминает ?
 - $p(x, y|w) = \frac{1}{1+\exp\{ -M\}}$



Сигмоида



Порождающая модель $p(x)$



Осталось определить $p(x_i|w)$

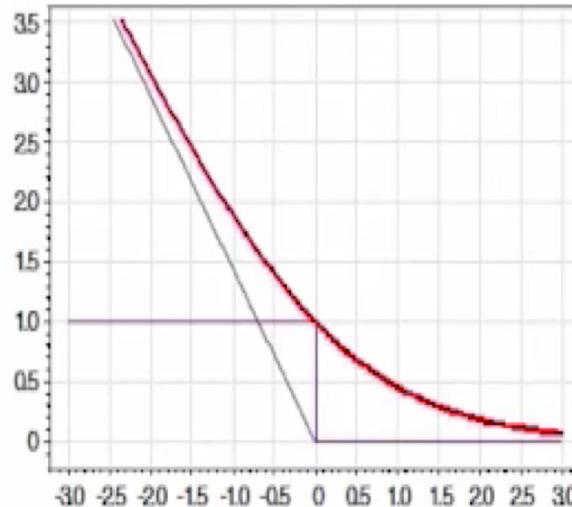
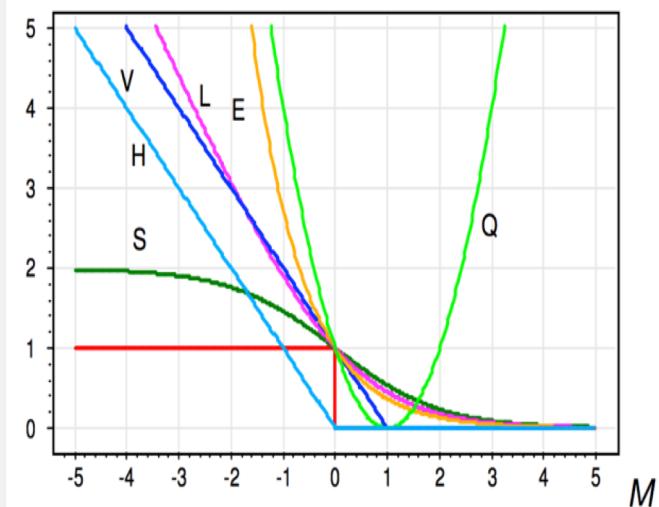
- $\frac{p(x,y=+1|w)}{p(x,y=-1|w)} = \exp(w, x)$
- Из полной вероятности
 - $p(x, y = +1|w) + p(x, y = -1|w) = 1$
- $p(x, y = +1|w) = \frac{1}{1+\exp\{-(w,x)\}}; p(x, y = -1|w) = \frac{1}{1+\exp\{(w,x)\}}$
- *Ничего не напоминает ? Обосновали сигмоиду!*
 - $p(x, y|w) = \frac{1}{1+\exp\{ -M\}}$



Логистическая регрессия



- $\sum_i^L \log p(x_i|w) = \sum_i^L \log \sigma(M) \rightarrow \max$
- - $\sum_i^L \log(1 + \exp(-M)) \rightarrow \max$
- $\sum_i^L \log(1 + \exp(-M)) \rightarrow \min$



Логистическая регрессия на 1 слайде



- $L = \log(1+e^{-M})$
- $Q'(w) = \sum_i^L \log(1 + \exp(-M_i)) \rightarrow \min_w$
- Оптимизация градиентным спуском
- $a(x_i, w_i) = sign [\sigma(x_i, w_i) - th]$



Логистическая регрессия на 1 слайде



- $L = \log(1+e^{-M})$
- $Q'(w) = \sum_i^L \log(1 + \exp(-M_i)) \rightarrow \min_w$
- Оптимизация градиентным спуском
- $a(x_i, w_i) = \text{sign} [\sigma(x_i, w_i) - th]$



Recap



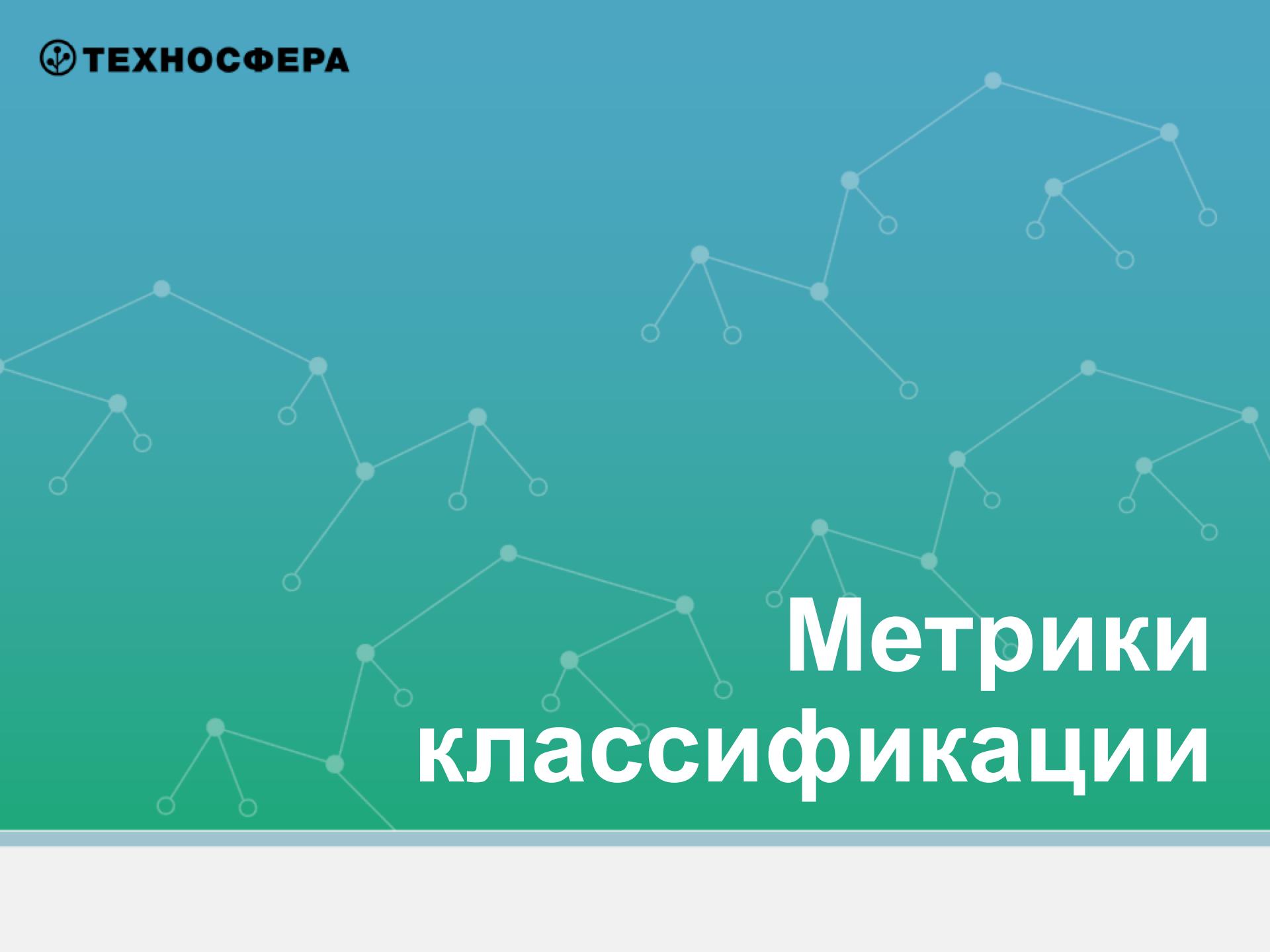
- Логистическая регрессия – алгоритм линейной классификации
- Апроксимирует эмпирический риск лог-лоссом
- **ИЛИ** порождает признаки из вероятностного пространства
 - Показали для Бернулли – бинарных признаков
 - *Но не для всех – сигмоиду нельзя трактовать как чистую вероятность. Но можно применять =)*
- Оптимизирует лог-лосс
 - **Или** решает задачу ММП
- Удобна для задач скоринга, спама
 - Важен не только аутпут, но и мера принадлежности



Recap

- Think empirically
- ИЛИ
- Think probabilistically





Метрики классификации

Метрики оценки



	true positive	false positive	true negative	false negative
--	------------------	-------------------	------------------	-------------------

y_pred	1			
y_true	1			



Метрики оценки



	true positive	false positive	true negative	false negative
--	------------------	-------------------	------------------	-------------------

y_pred	1	1		
y_true	1	0		



Метрики оценки



	true positive	false positive	true negative	false negative
--	------------------	-------------------	------------------	-------------------

y_pred	1	1	0	
y_true	1	0	0	



Метрики оценки



	true positive	false positive	true negative	false negative
--	------------------	-------------------	------------------	-------------------

y_pred	1	1	0	0
y_true	1	0	0	1



Метрики оценки: точность [precision]



$$precision = \frac{tp}{tp + fp}$$



Метрики оценки: полнота [recall]



$$precision = \frac{tp}{tp + fn}$$



Метрики оценки: f1-мера

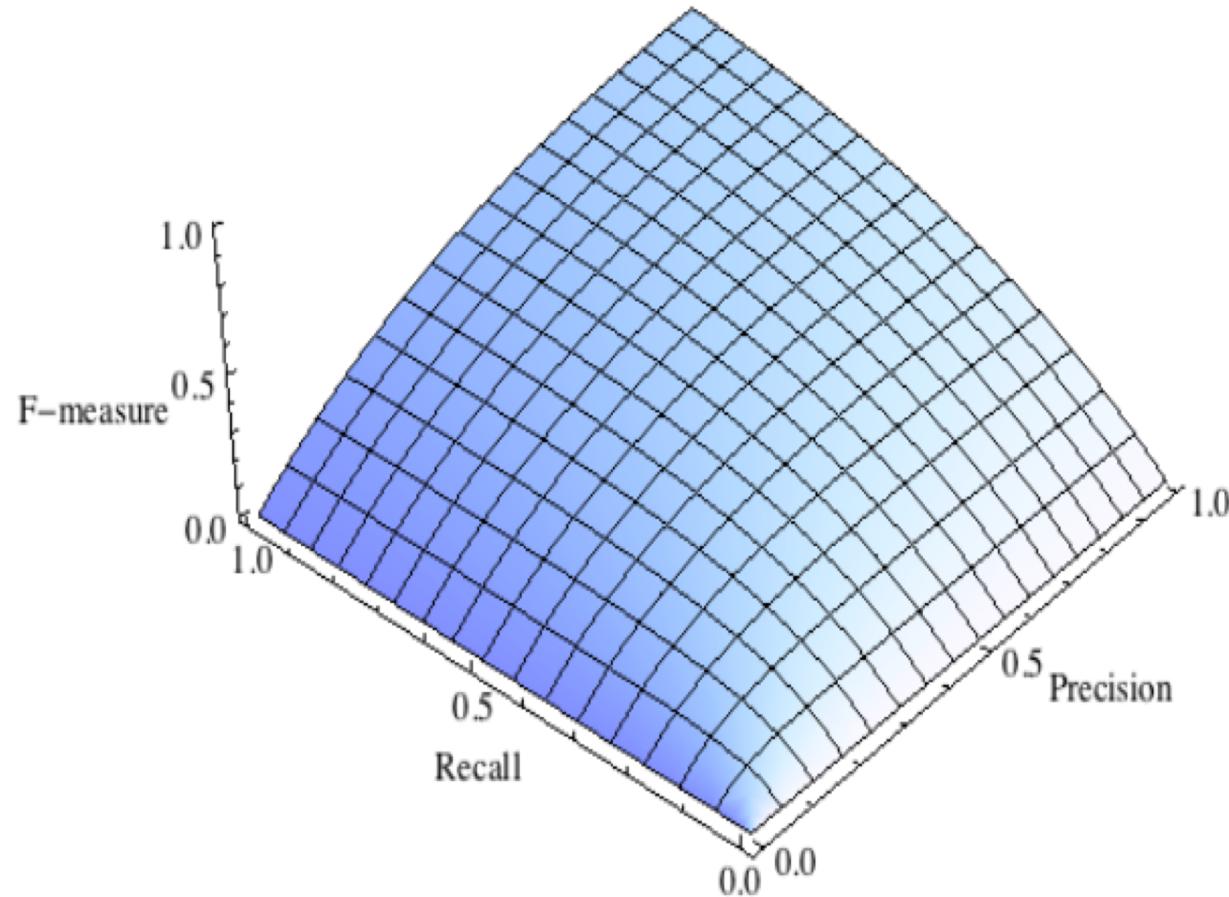


$$f1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

- По сути – среднее гармоническое



Метрики оценки: f1-мера



Сбалансированная F-мера



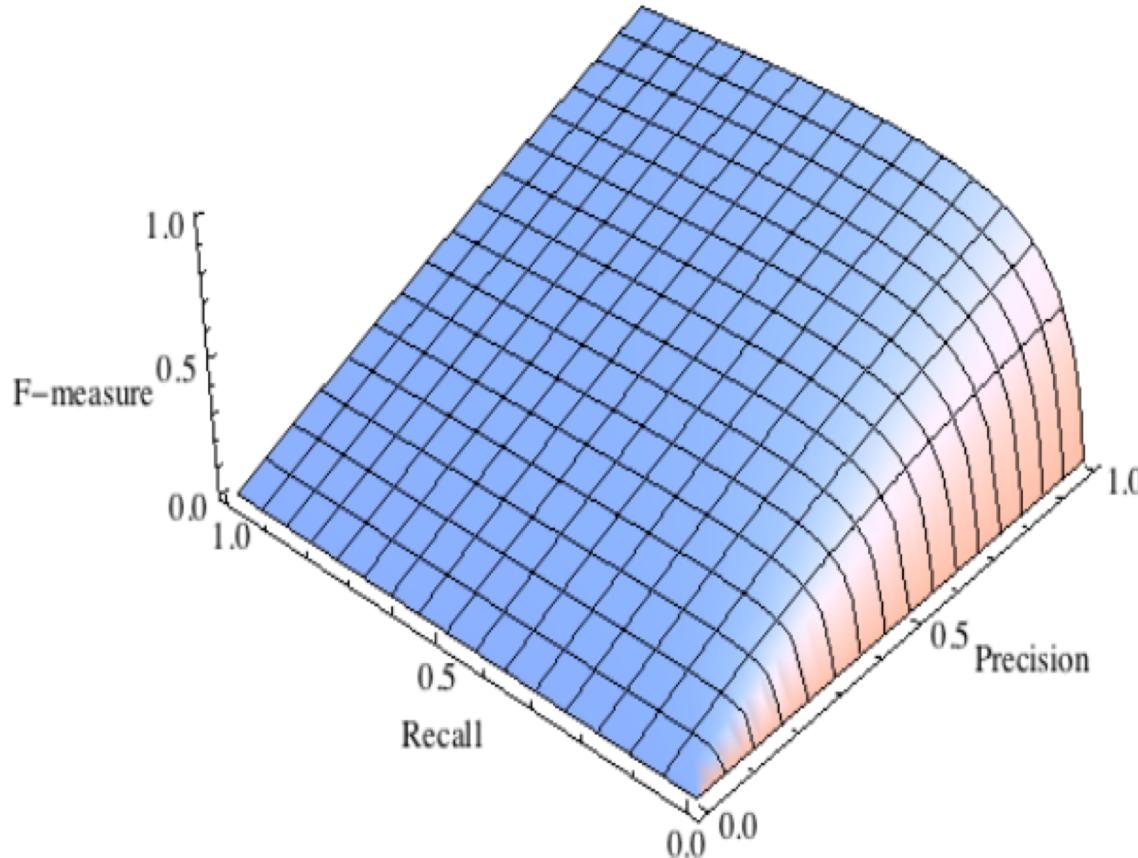
Метрики оценки: F-мера



$$F\beta = \frac{1 + \beta^2}{\frac{1}{precision} + \frac{\beta^2}{recall}}$$



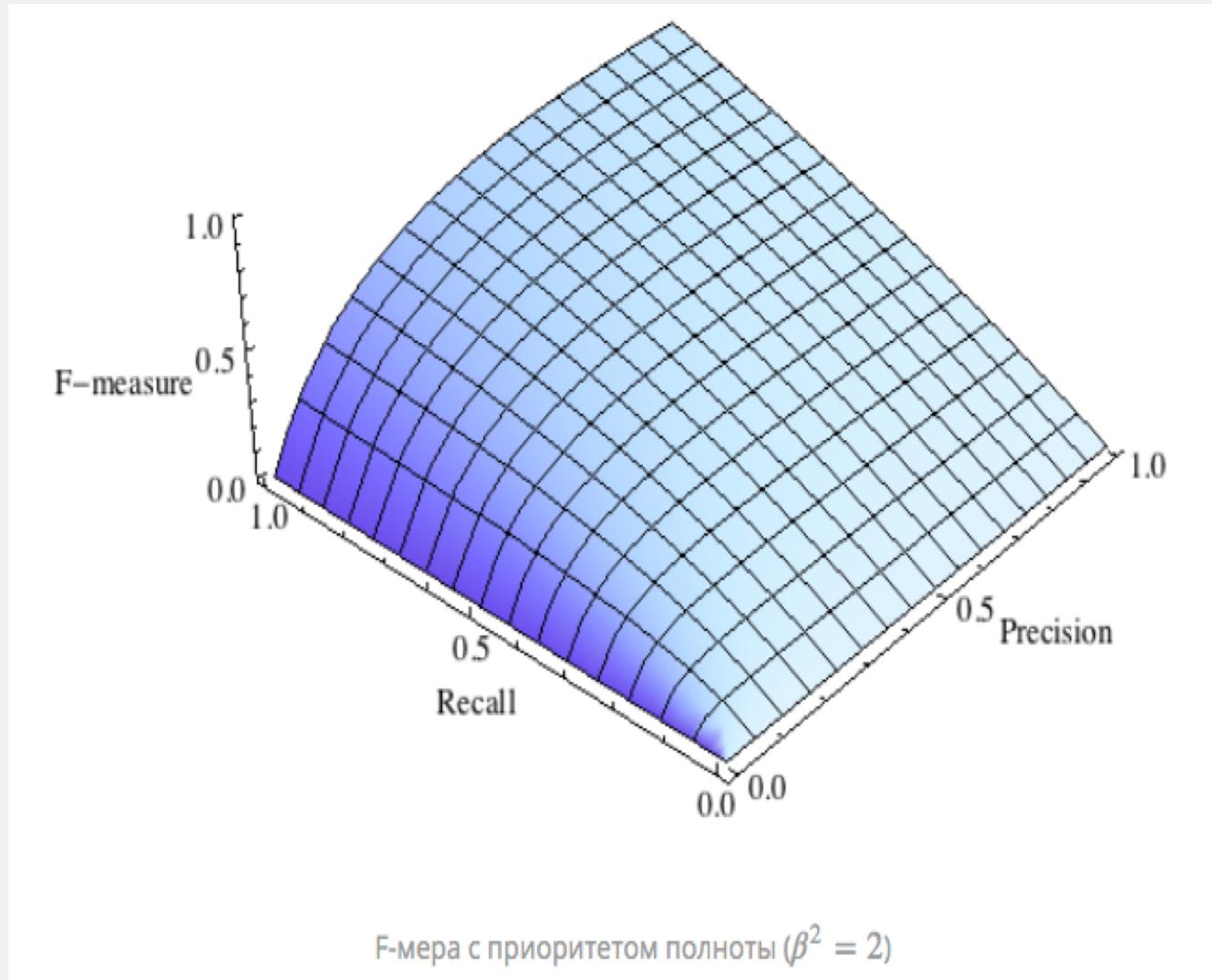
Метрики оценки: f1-мера



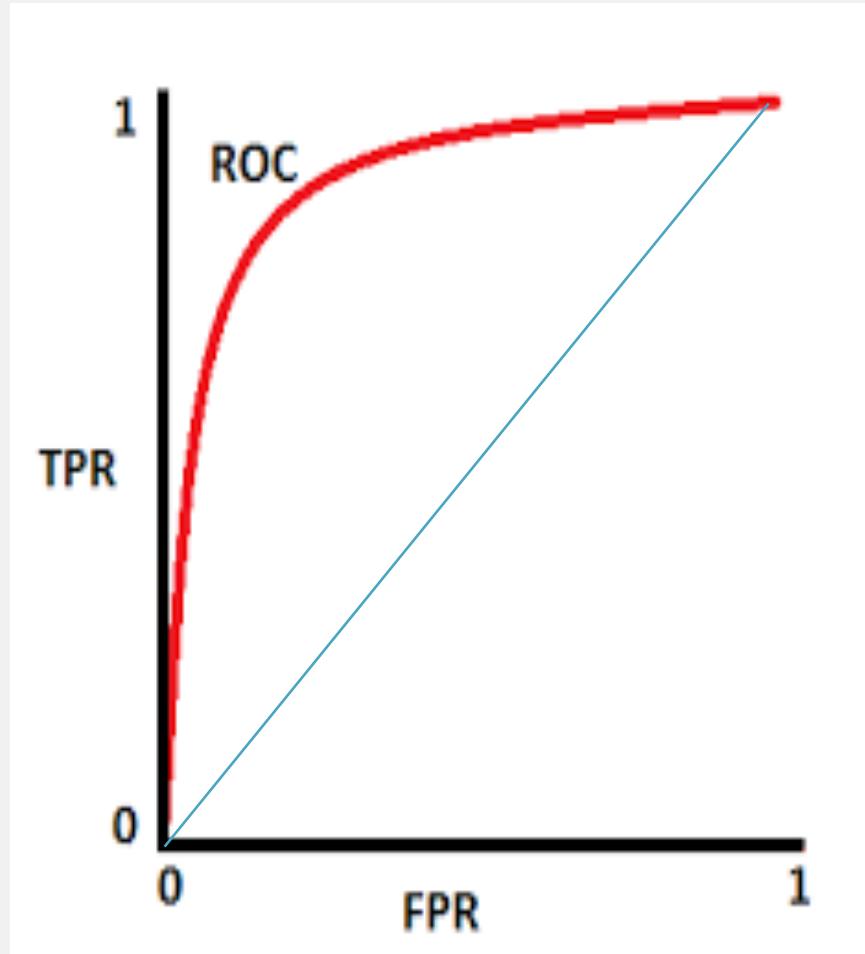
F-мера с приоритетом точности ($\beta^2 = \frac{1}{4}$)



Метрики оценки: f1-мера



Метрики оценки: ROC AUC



- ROC AUC – площадь под кривой
 - ROC = Receiver Operator Characteristics
 - AUC = Area Under Curve
- $0 < \text{ROC AUC} < 1$
- Random Classifier: $\text{ROC AUC} = 0.5$



Метрики оценки: ROC AUC



- Внутри квадрата – все пары точек 1 и 0 из выборки
- ROC отделяет верно упорядоченные алгоритмом пары точек
- ROC AUC – число верно отранжированных пар
- Исчерпывающая интерпретация:
 - **Вероятность, что алгоритм верно упорядочит случайно выбранную пару (1,0)**



Как подобрать порог

- Оптимизация F1 кросс-валидацией
- Подбор на ROC кривой исходя из целевых
 - FPR
 - TPR



Спасибо за внимание!



Конец презентации