Основы машинного обучения

Лекция 10

Метрики качества классификации.

Логистическая регрессия. Метод опорных векторов.

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2022

Метрики качества ранжирования

Классификатор

- Линейный классификатор:

$$a(x) = \operatorname{sign}(\langle w, x \rangle - t) = 2[\langle w, x \rangle > t] - 1$$

- $\langle w, x \rangle$ — оценка принадлежности классу +1
- Нередко $t = 0$

Оценка принадлежности

- Высокий порог:
  - Мало объектов относим к +1
  - Точность выше
  - Полнота ниже
- Низкий порог:
  - Много объектов относим к +1
  - Точность ниже
  - Полнота выше

Оценка принадлежности

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности

-1	**-1**	**+1**	**-1**	**-1**	**-1**	**+1**	**+1**	**-1**	**+1**
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности

-1	+1	-1	+1	-1	+1	-1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

Оценка принадлежности

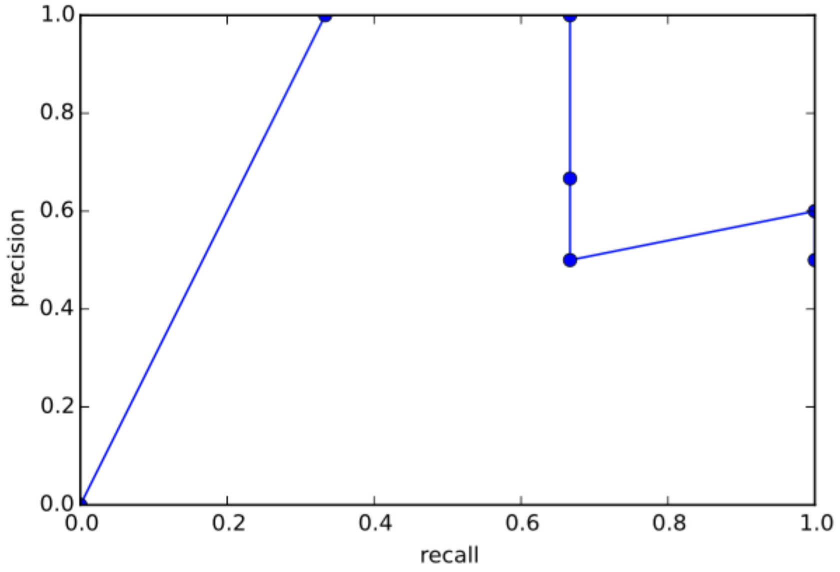
- Как оценить качество $b(x)$?
- Порог выбирается позже
- Порог зависит от ограничений на точность или полноту

Оценка принадлежности

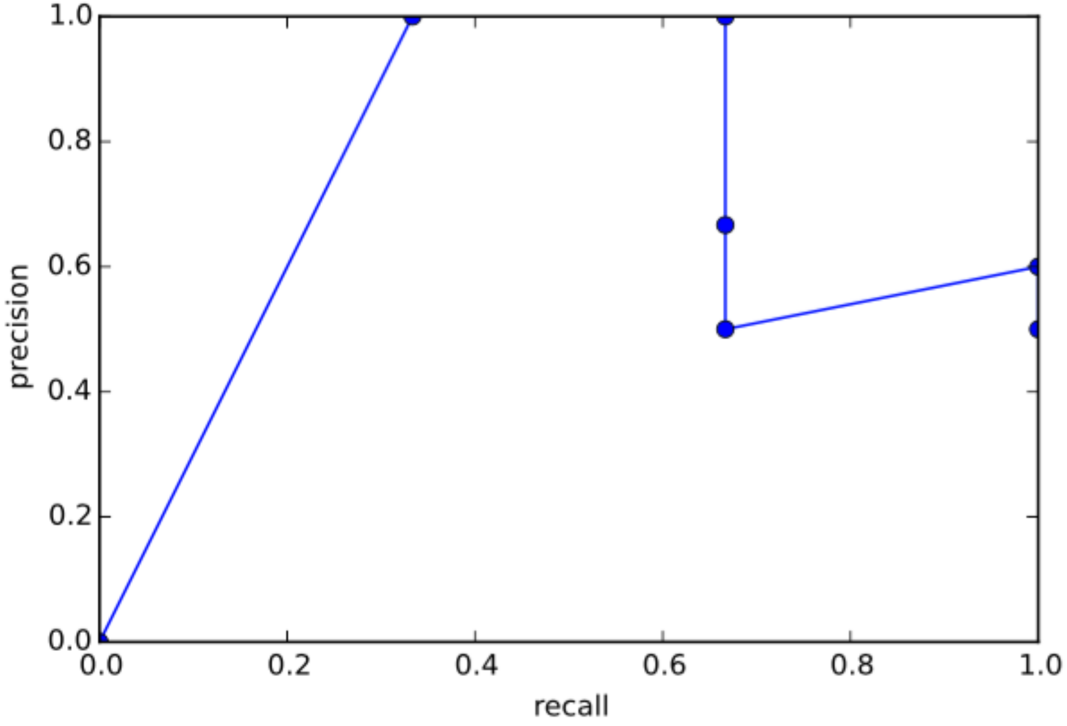
- Пример: кредитный скоринг
- $b(x)$ — оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- $\text{precision} = 0.1, \text{recall} = 0.7$
- В чем дело — в пороге или в алгоритме?

PR-кривая

- Кривая точности-полноты
- Ось X — полнота
- Ось Y — точность
- Точки — значения точности и полноты при последовательных порогах

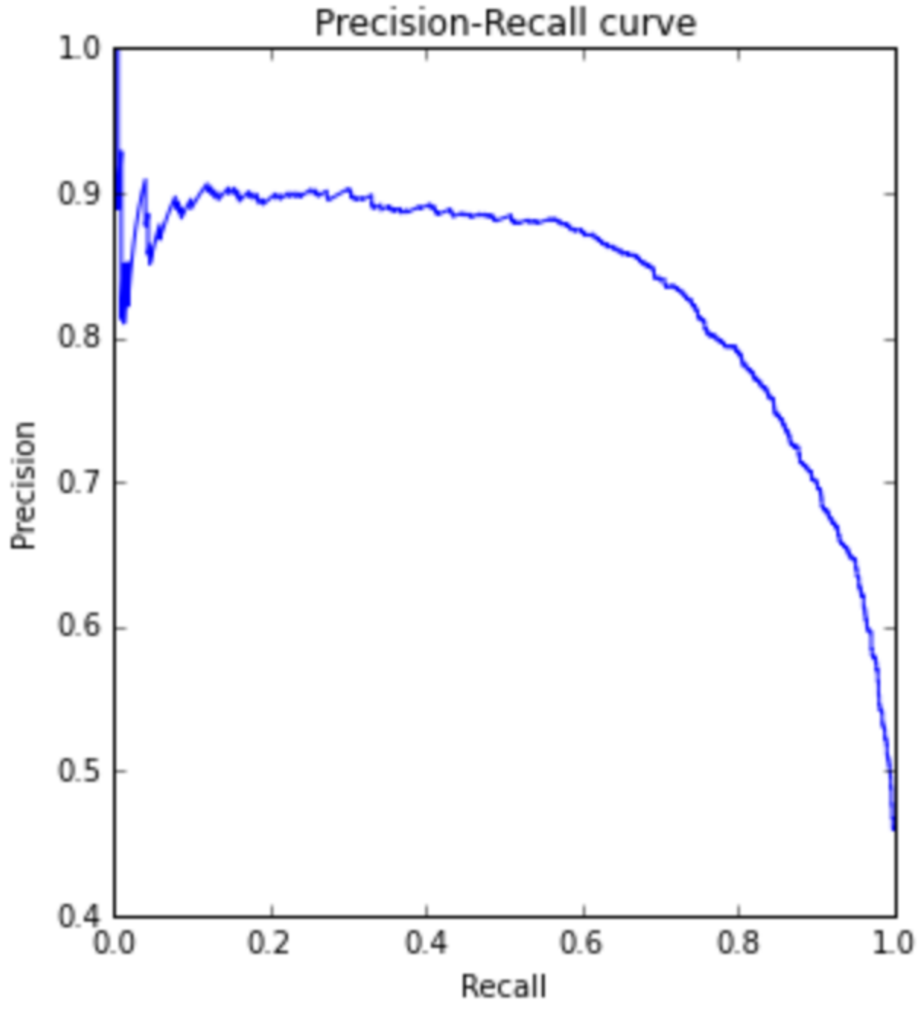


PR-кривая



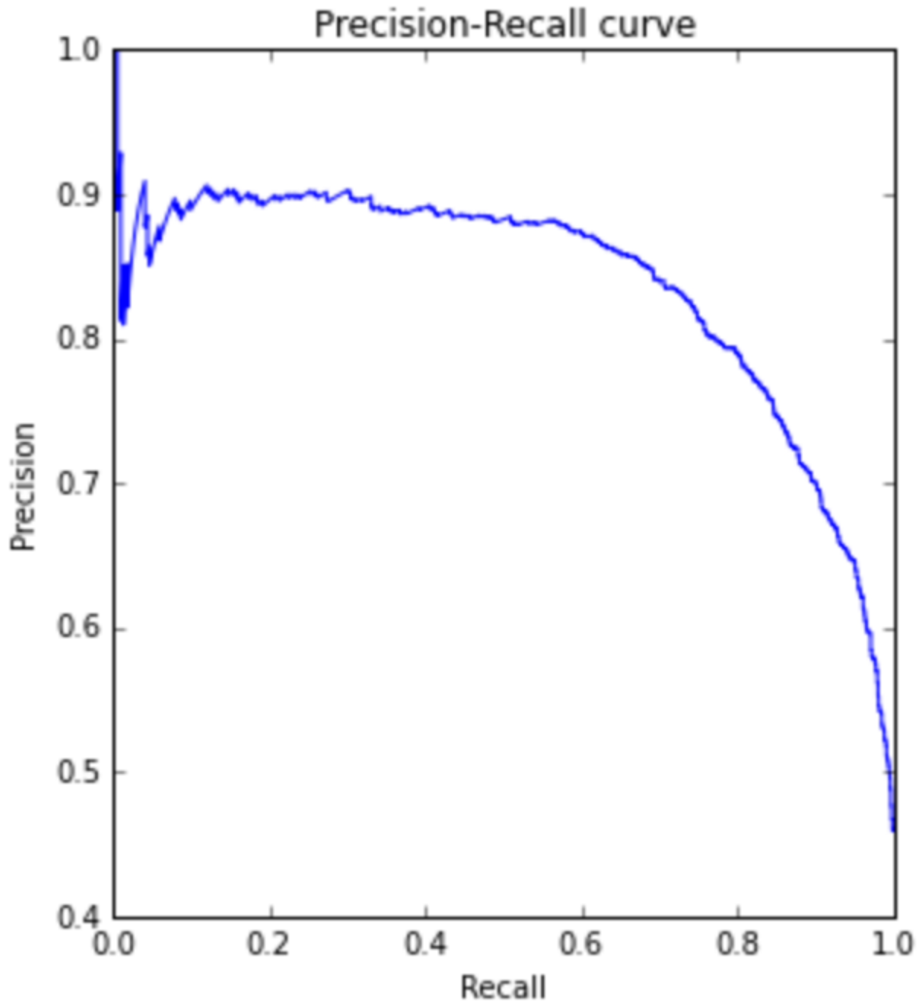
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

PR-кривая в реальности

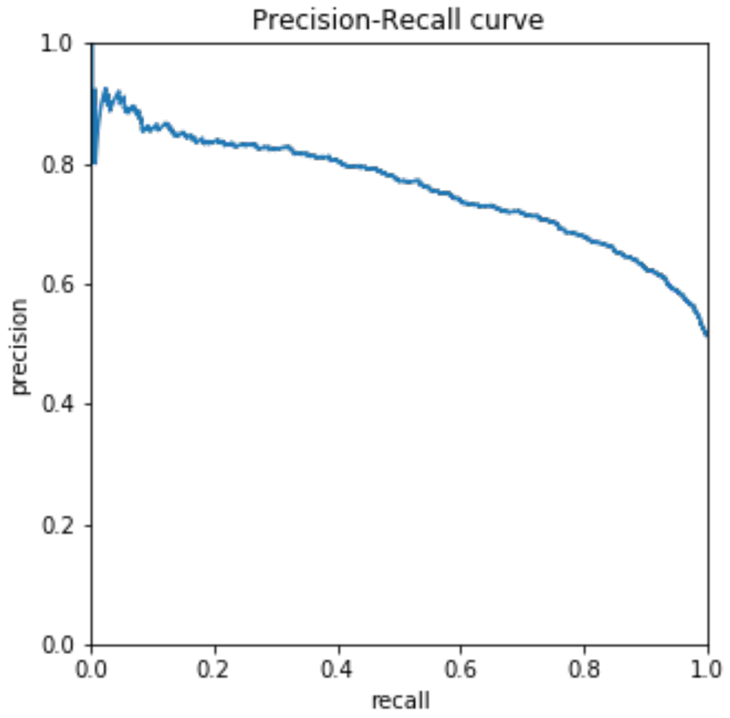


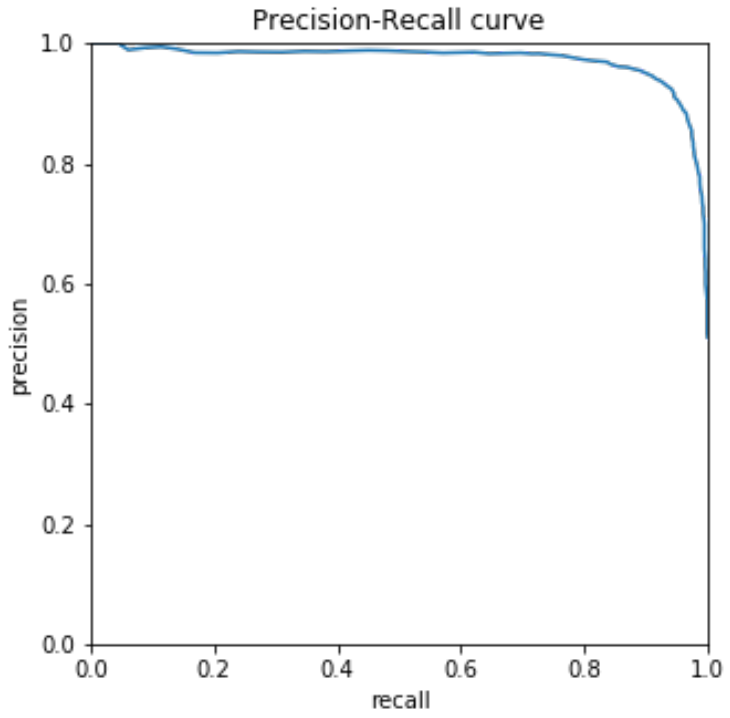
PR-кривая

- Левая точка: (0, 1)
- Правая точка: $(1, r)$, r — доля положительных объектов
- Для идеального классификатора проходит через (1, 1)
- AUC-PRC — площадь под PR-кривой



PR-кривая


Precision-Recall curve
precision
recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall curve
precision
recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

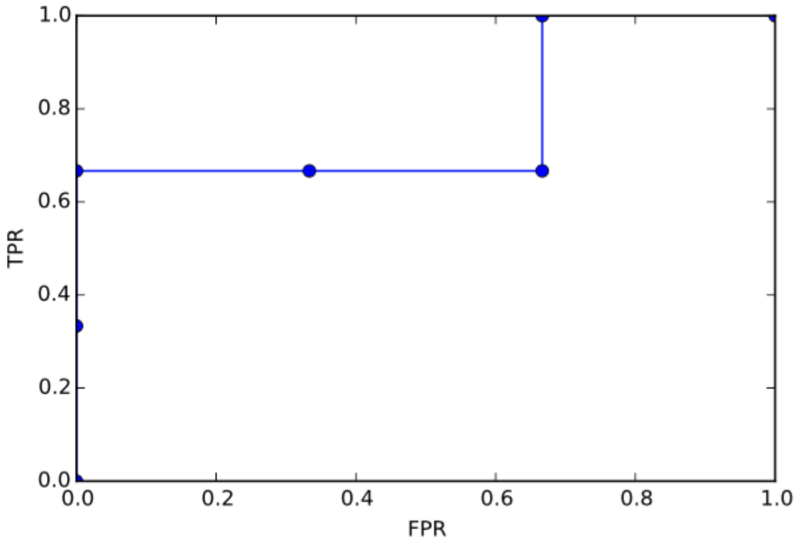
ROC-кривая

- Receiver Operating Characteristic
- Ось X — False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

- Ось Y — True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$



ROC-кривая

- Receiver Operating Characteristic
- Ось X — False Positive Rate

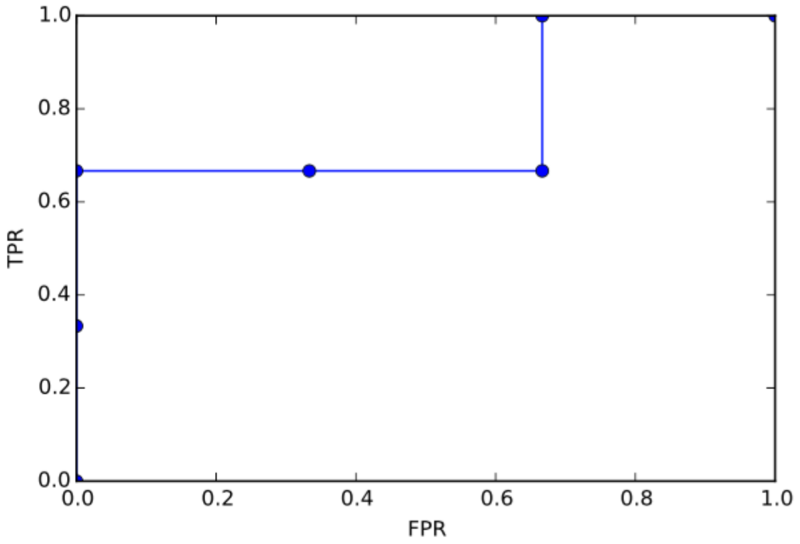
$$FPR = \frac{FP}{\boxed{FP + TN}}$$

Число отрицательных объектов

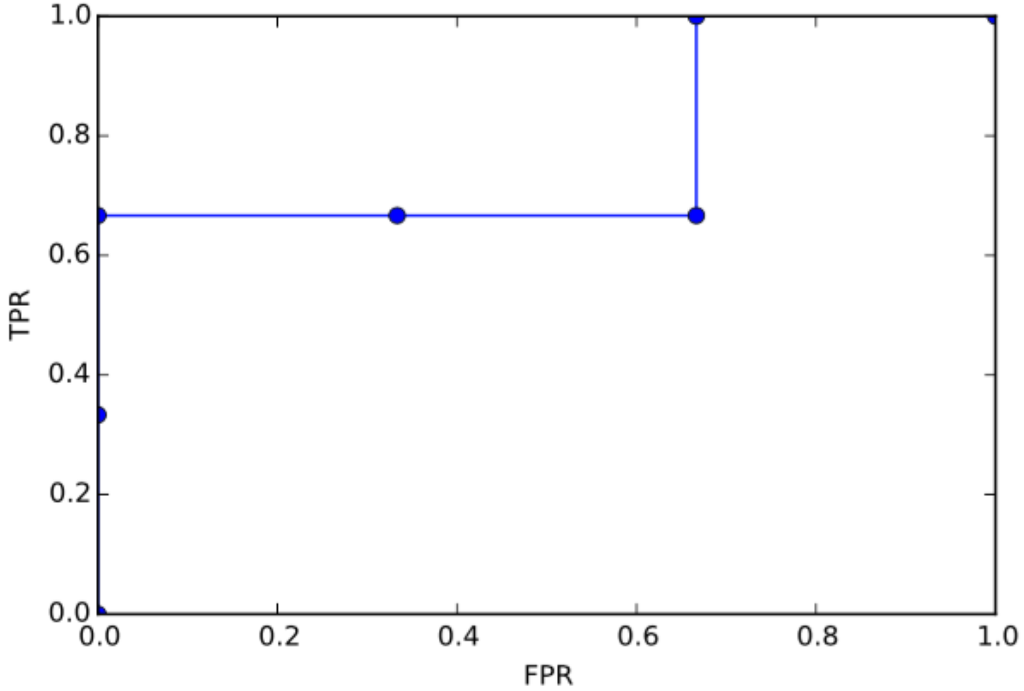
- Ось Y — True Positive Rate

$$TPR = \frac{TP}{\boxed{TP + FN}}$$

Число положительных объектов

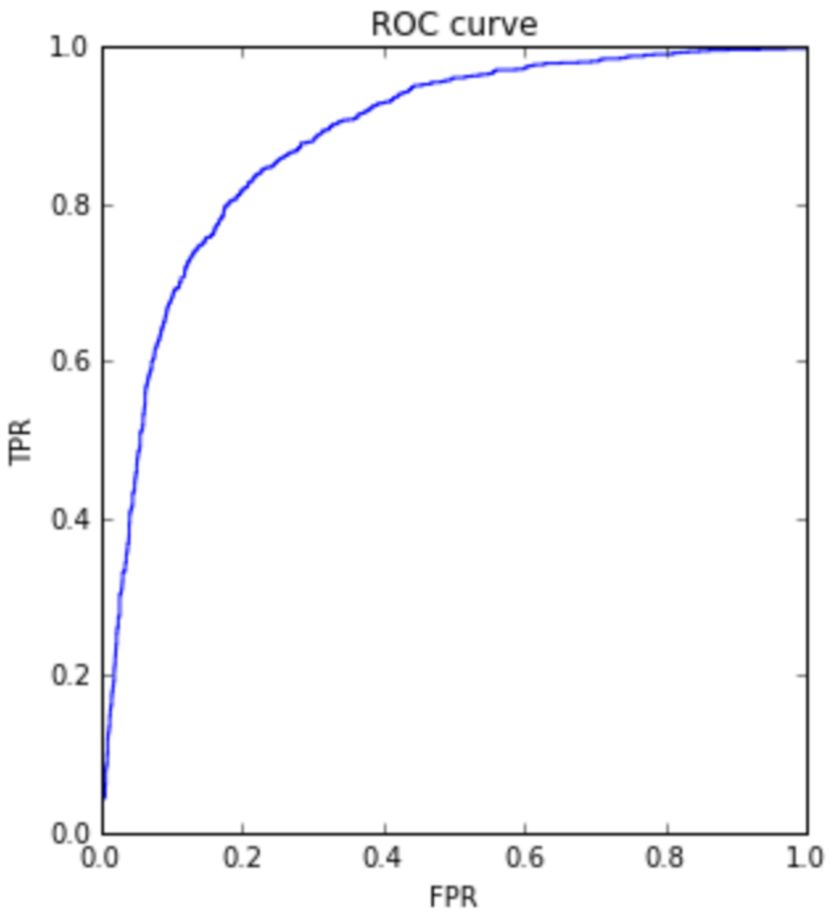


ROC-кривая



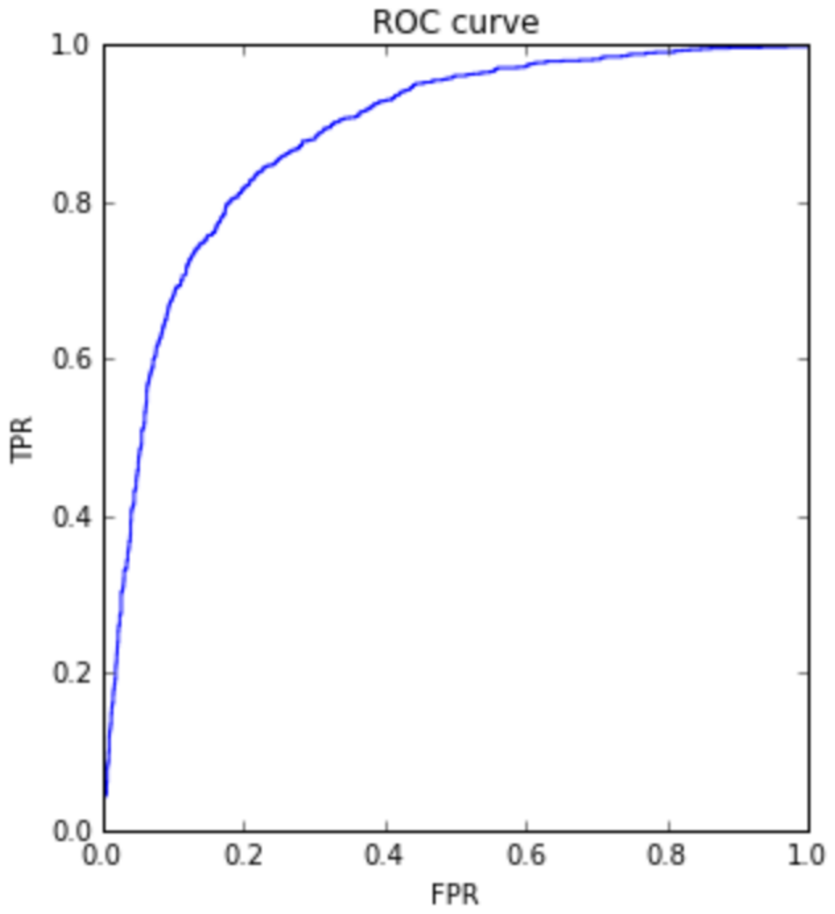
$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

ROC-кривая в реальности

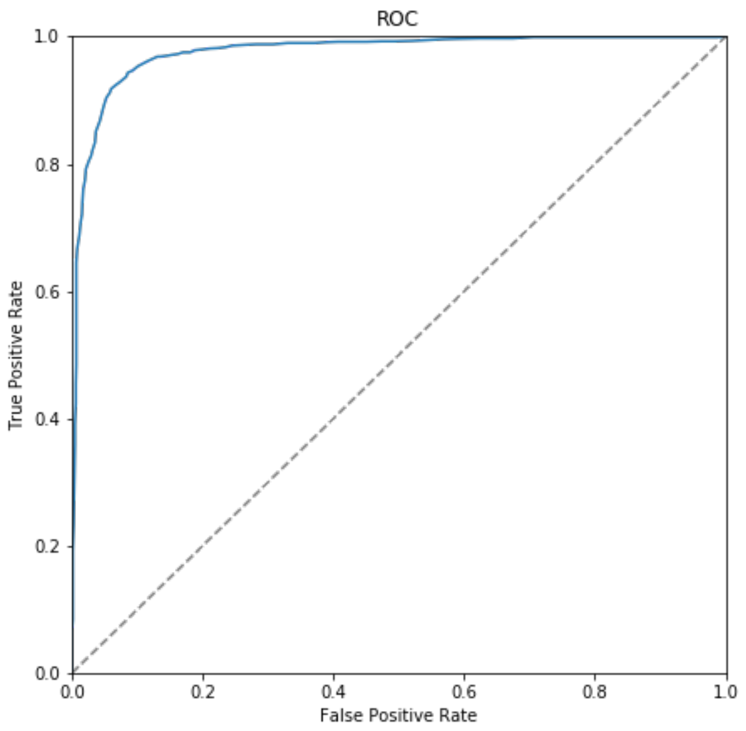


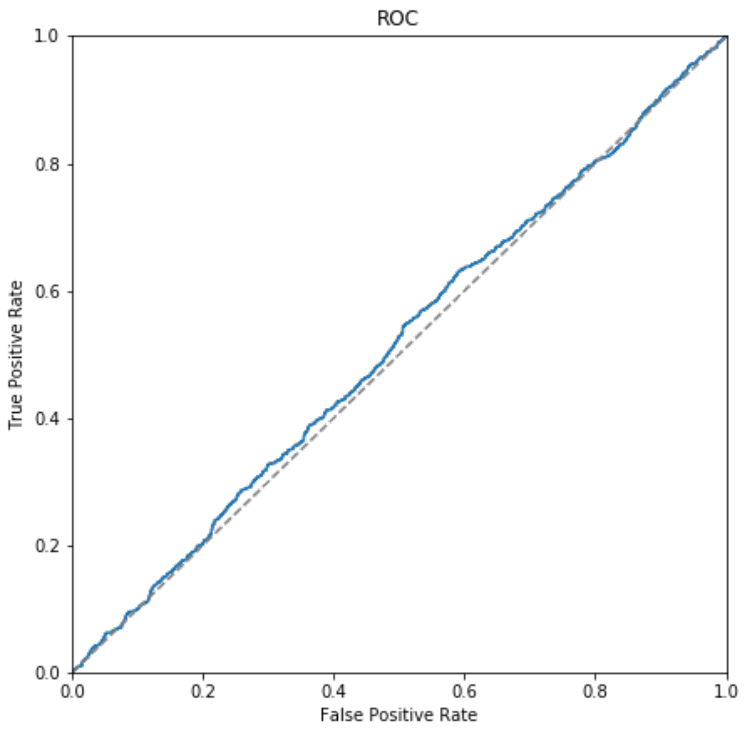
ROC-кривая

- Левая точка: (0, 0)
- Правая точка: $(1, 1)$
- Для идеального классификатора проходит через (0, 1)
- AUC-ROC — площадь под ROC-кривой



ROC-кривая


ROC
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate


ROC
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

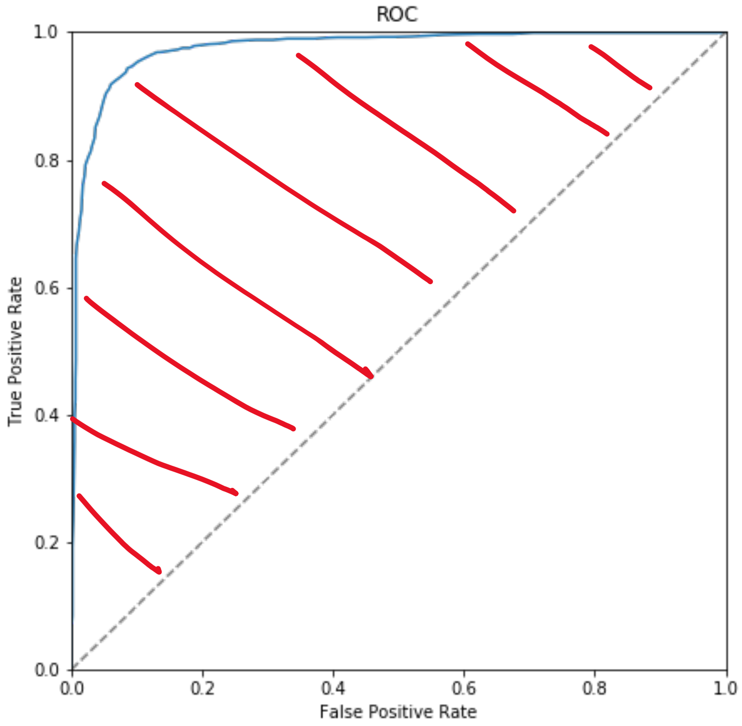
AUC-ROC

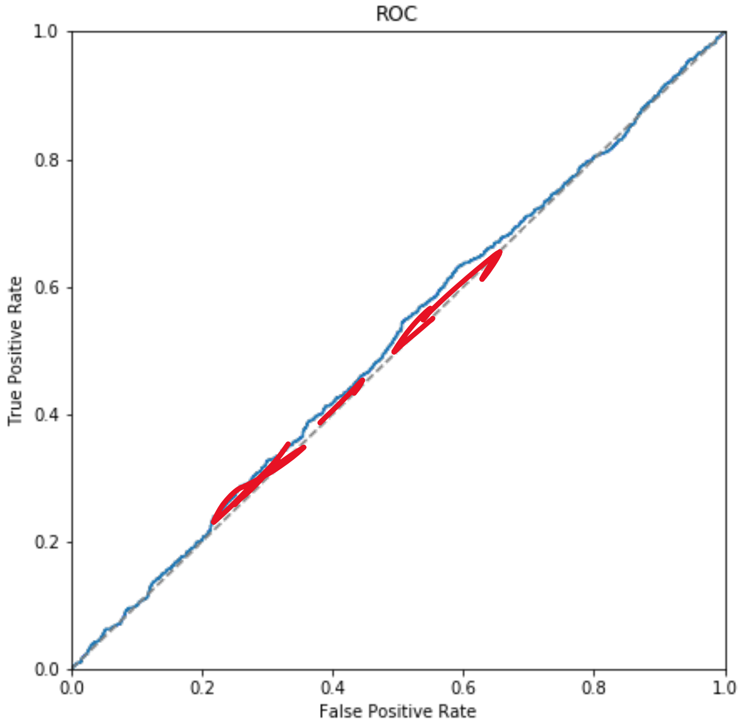
$$FPR = \frac{FP}{FP+TN}; \qquad TPR = \frac{TP}{TP+FN}$$

- FPR и TPR нормируются на размеры классов
- AUC-ROC не поменяется при изменении баланса классов
- Идеальный алгоритм: $\mathrm{AUC{-}ROC} = 1$
- Худший алгоритм: $\mathrm{AUC{-}ROC} \approx 0.5$
- Интересные интерпретации: например, это примерно доля пар правильно упорядоченных объектов

Коэффициент Джини

$$\text{Gini} = 2 * (\text{AUC–ROC} - 0.5)$$





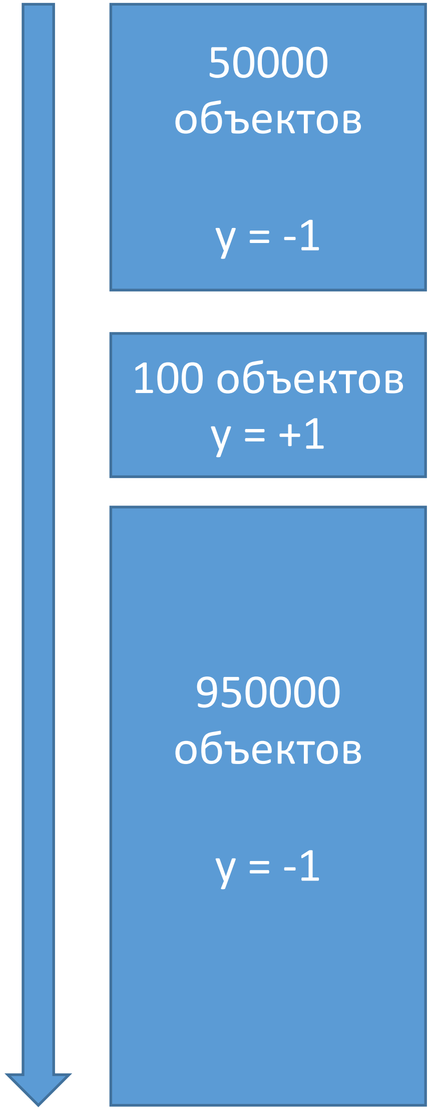
AUC-PRC

$$\text{precision} = \frac{TP}{TP+FP}; \qquad \text{recall} = \frac{TP}{TP+FN}$$

- Точность поменяется при изменении баланса классов
- AUC-PRC идеального алгоритма зависит от баланса классов
- Проще интерпретировать, если выборка несбалансированная
- Лучше, если задачу надо решать в терминах точности и полноты

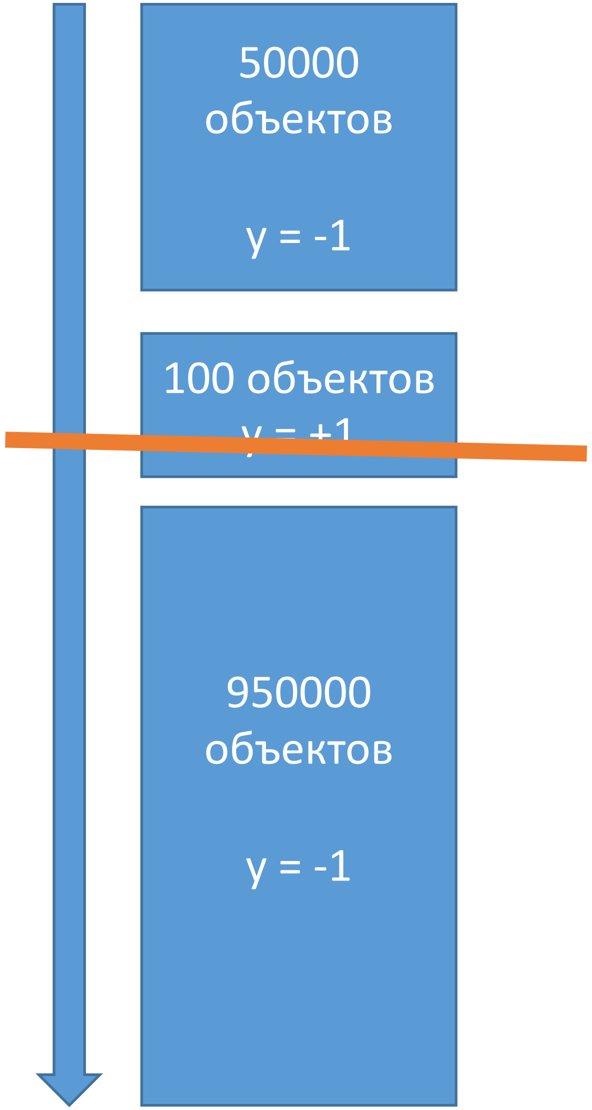
Пример

- AUC-ROC = 0.95
- AUC-PRC = 0.001



Пример

- Выберем конкретный классификатор
- $a(x) = 1$ — 50095 объектов
- Из них FP = 50000, TP = 95
- TPR = 0.95, FPR = 0.05
- precision = 0.0019, recall = 0.95



Логистическая регрессия: простое объяснение

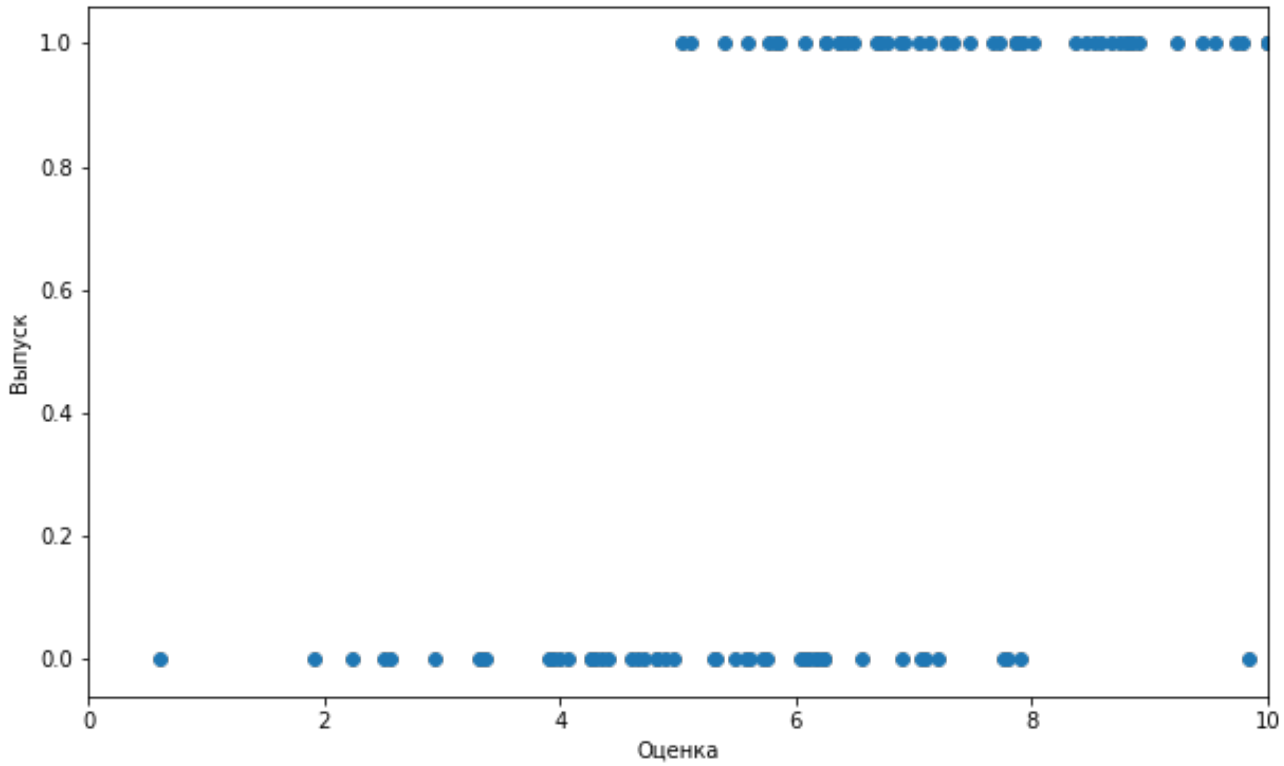
Логистическая регрессия

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

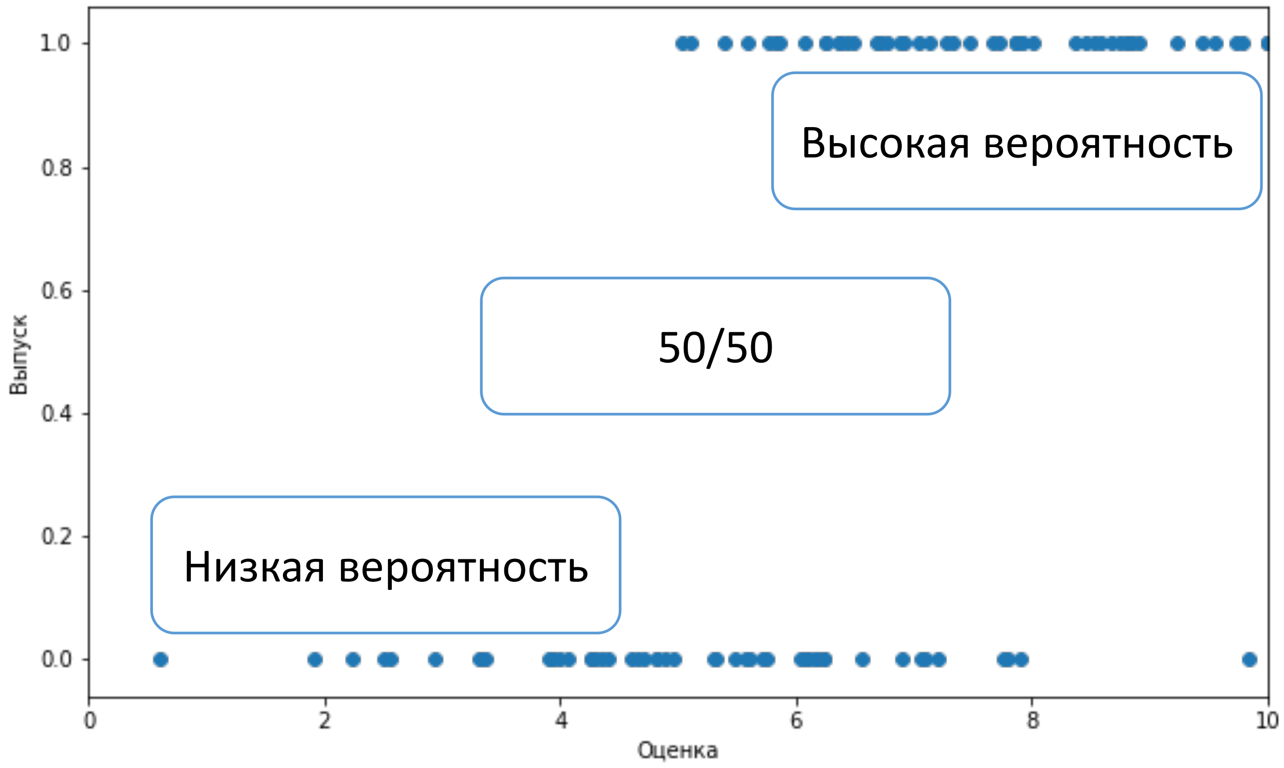
- Минимизация верхней оценки:

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

Предсказание вероятностей



Предсказание вероятностей



Предсказание вероятностей

- Кредитный скоринг
- Стратегия: выдавать кредит только клиентам с $b(x) > 0.9$
- 10% невозвращённых кредитов — нормально

Предсказание вероятностей

- Баннерная реклама
- $b(x)$ — вероятность, что пользователь кликнет по рекламе
- $c(x)$ — прибыль в случае клика
- $c(x)b(x)$— хотим оптимизировать

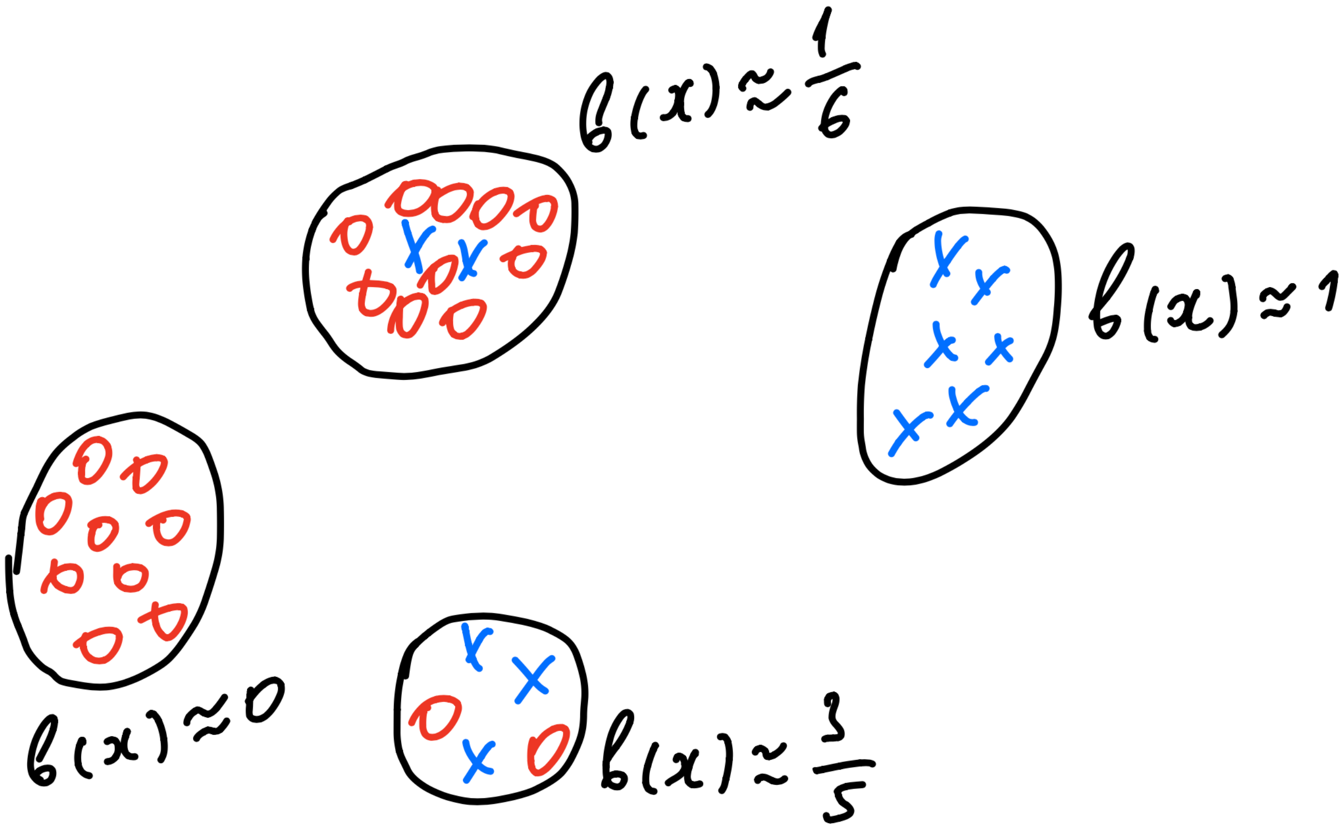
Предсказание вероятностей

- Прогнозирование оттока клиентов
- Медицинская диагностика
- Поисковое ранжирование (насколько веб-страница соответствует запросу?)

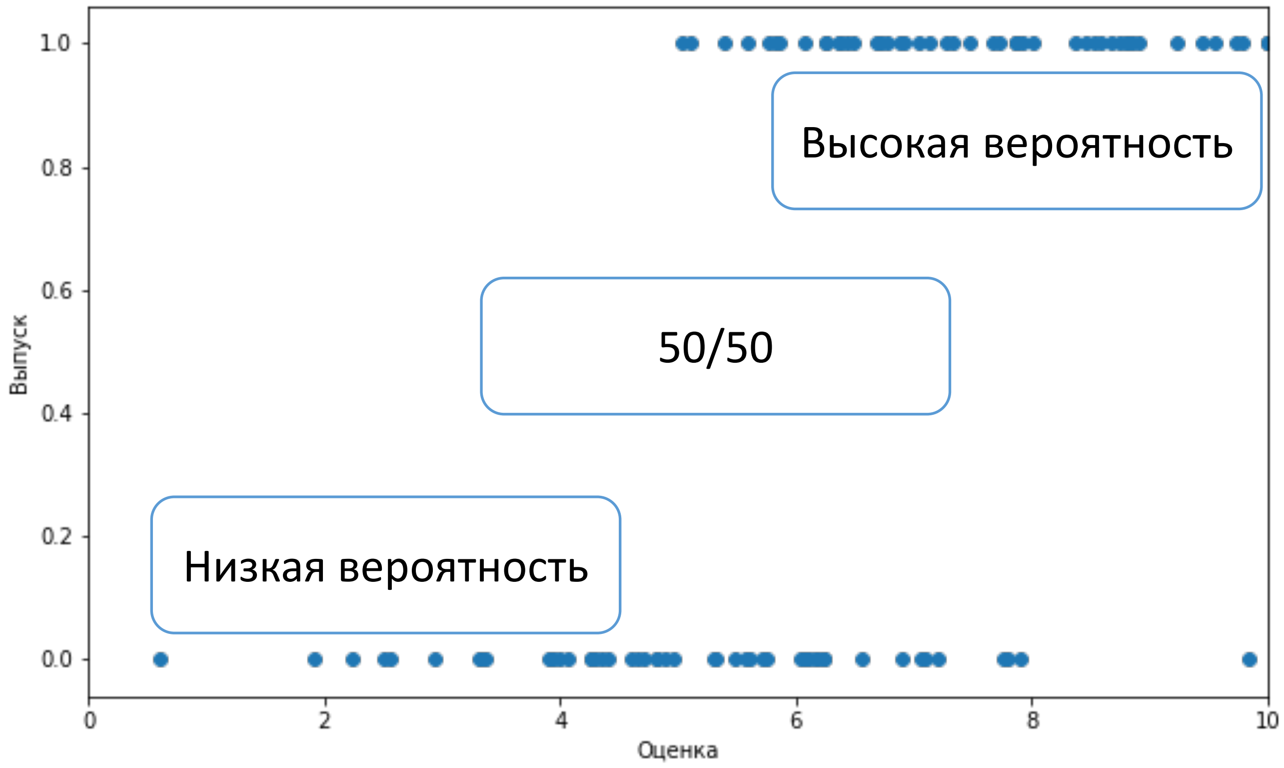
Предсказание вероятностей

Будем говорить, что модель $b(x)$ предсказывает вероятности, если среди объектов с $b(x) = p$ доля положительных равна p.

Предсказание вероятностей



Предсказание вероятностей



Линейный классификатор

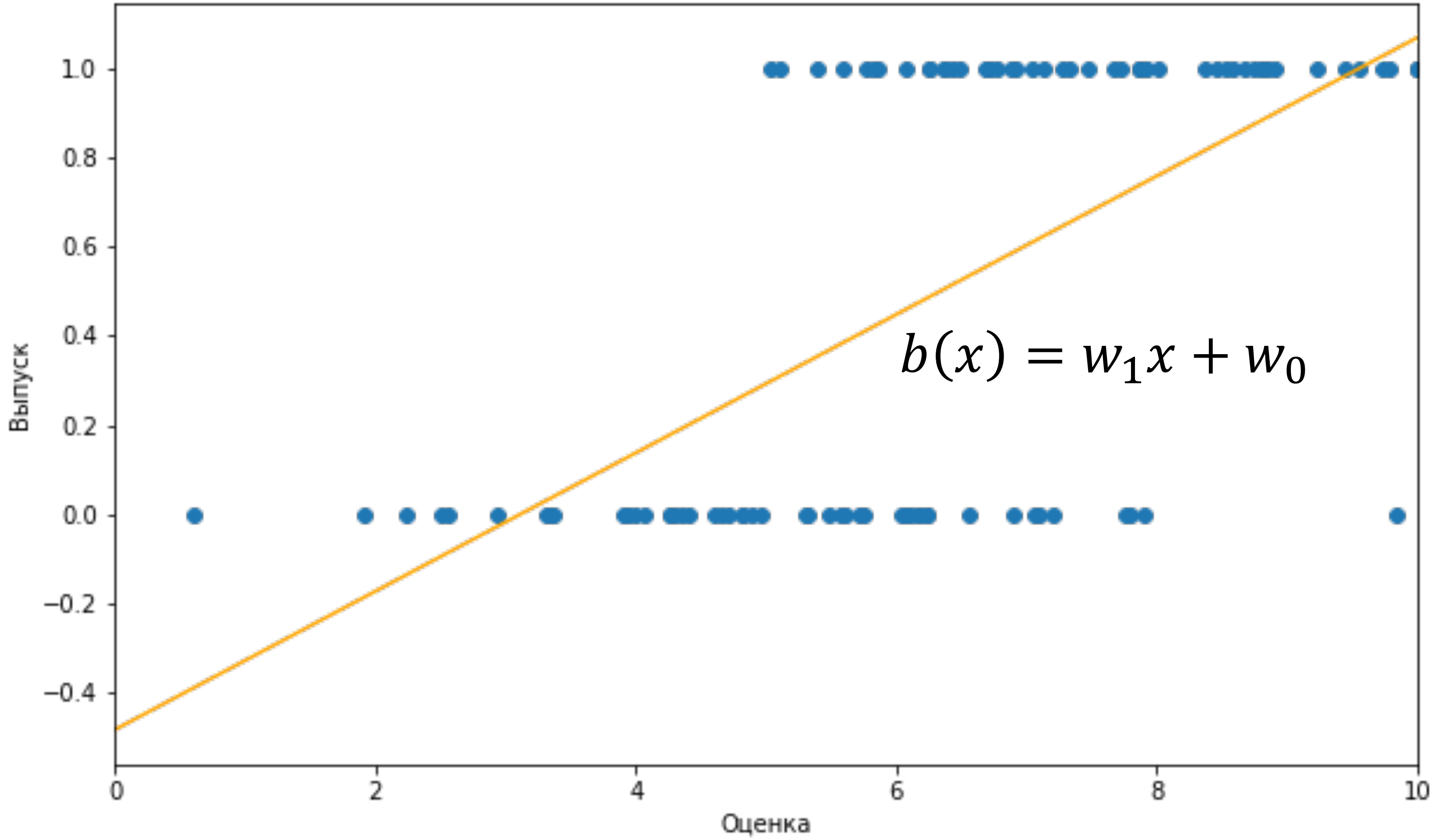
$$a(x) = \operatorname{sign} \langle w, x \rangle$$

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

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

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

Предсказание вероятностей

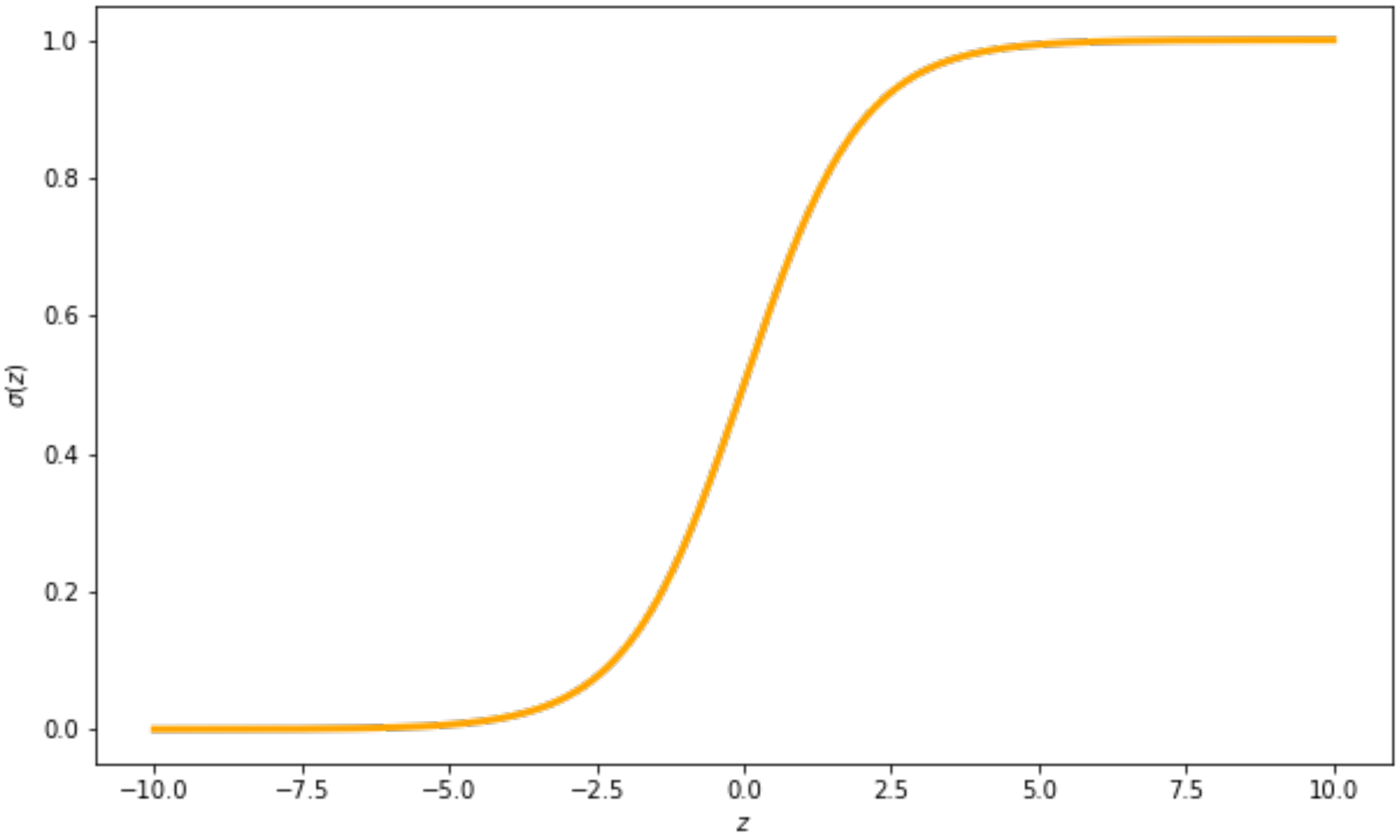


Линейный классификатор

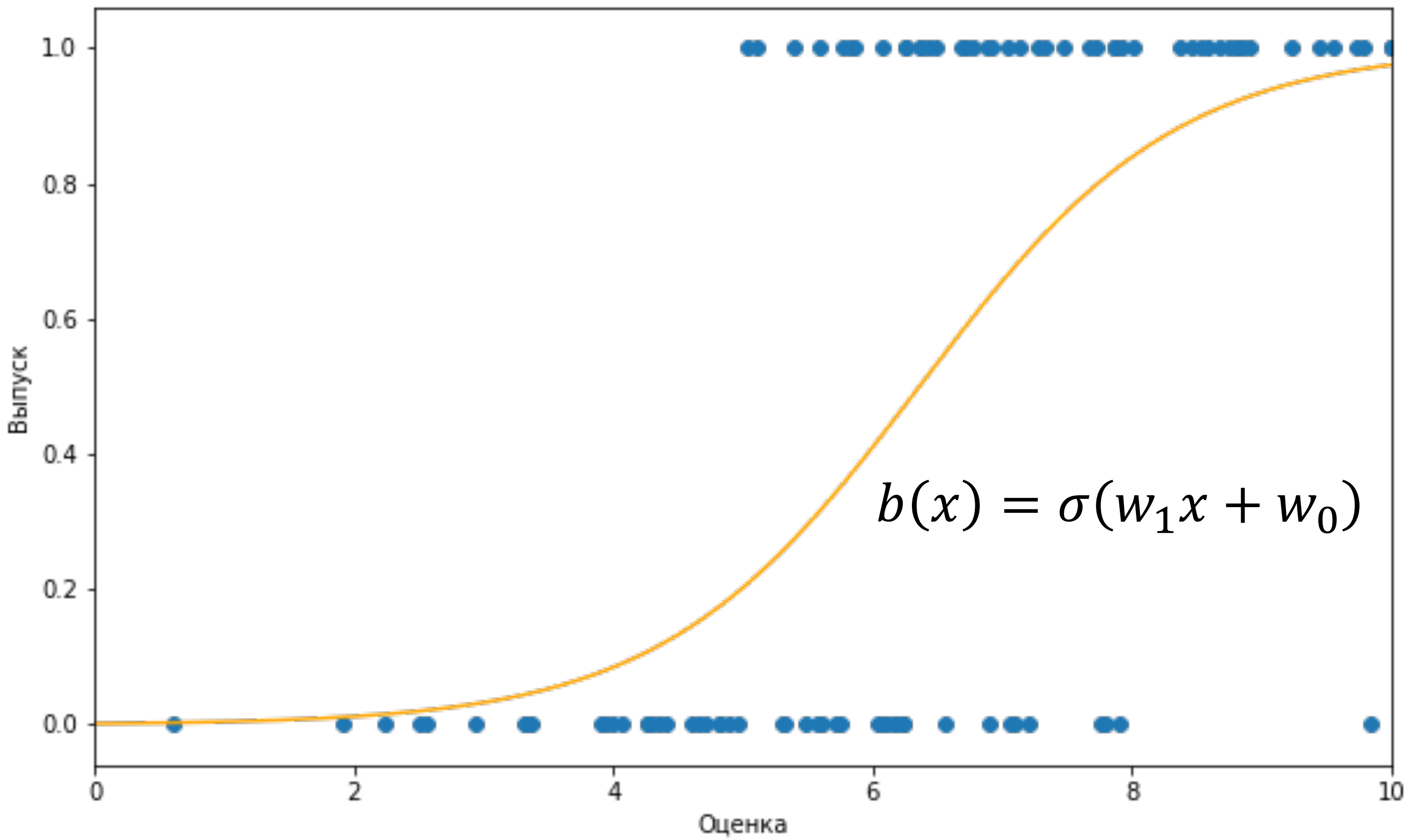
- Переведём выход модели на отрезок [0, 1]
- Например, с помощью сигмоиды:

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

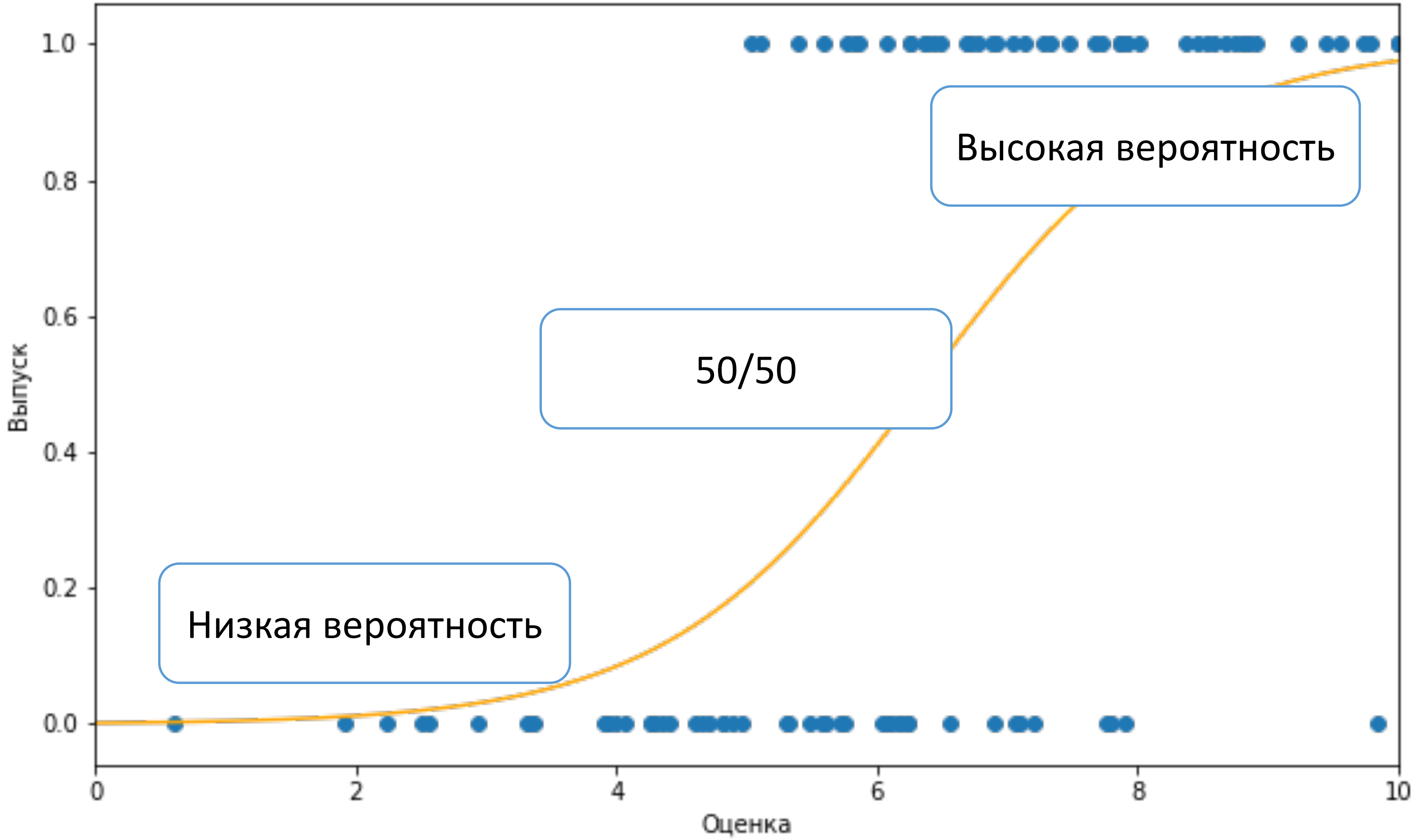
Сигмоида



Предсказание вероятностей



Предсказание вероятностей



Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

Предсказание вероятностей

- Модель для оценивания вероятностей:

$$b(x) = \sigma(\langle w, x \rangle)$$

- Как обучать?
- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$ или $\langle w, x_i \rangle \to +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$ или $\langle w, x_i \rangle \to -\infty$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$ или $\langle w, x_i \rangle \to +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$ или $\langle w, x_i \rangle \to -\infty$
- То есть задача — сделать отступы на всех объектах максимальными

$$y_i \langle w, x_i \rangle \to \max_w$$

Предсказание вероятностей

- Если $y_i = +1$, то $\sigma(\langle w, x_i \rangle) \to 1$
- Если $y_i = -1$, то $\sigma(\langle w, x_i \rangle) \to 0$

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

Предсказание вероятностей

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

- Если $y_i = +1$ и $\sigma(\langle w, x_i\rangle) = 0$, то штраф равен 1
- Если $y_i = +1$, то заменить $\sigma(\langle w, x_i\rangle) = 1$ на $\sigma(\langle w, x_i\rangle) = 0.5$ так же плохо, как заменить $\sigma(\langle w, x_i\rangle) = 0.5$ на $\sigma(\langle w, x_i\rangle) = 0$
- Надо строже!

Предсказание вероятностей

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

- 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**

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

Логистическая регрессия

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

$$-\sum_{i=1}^{\ell}\left\{[y_i = 1]\log\frac{1}{1 + \exp(-\langle w, x\rangle)} + [y_i = -1]\log\left(1 - \frac{1}{1 + \exp(-\langle w, x\rangle)}\right)\right\} =$$

$$-\sum_{i=1}^{\ell}\left\{[y_i = 1]\log\frac{1}{1 + \exp(-\langle w, x\rangle)} + [y_i = -1]\log\left(\frac{1}{1 + \exp(\langle w, x\rangle)}\right)\right\} =$$

$$\sum_{i=1}^{\ell}\{[y_i = 1]\log(1 + \exp(-\langle w, x\rangle)) + [y_i = -1]\log(1 + \exp(\langle w, x\rangle))\} =$$

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

Метод опорных векторов

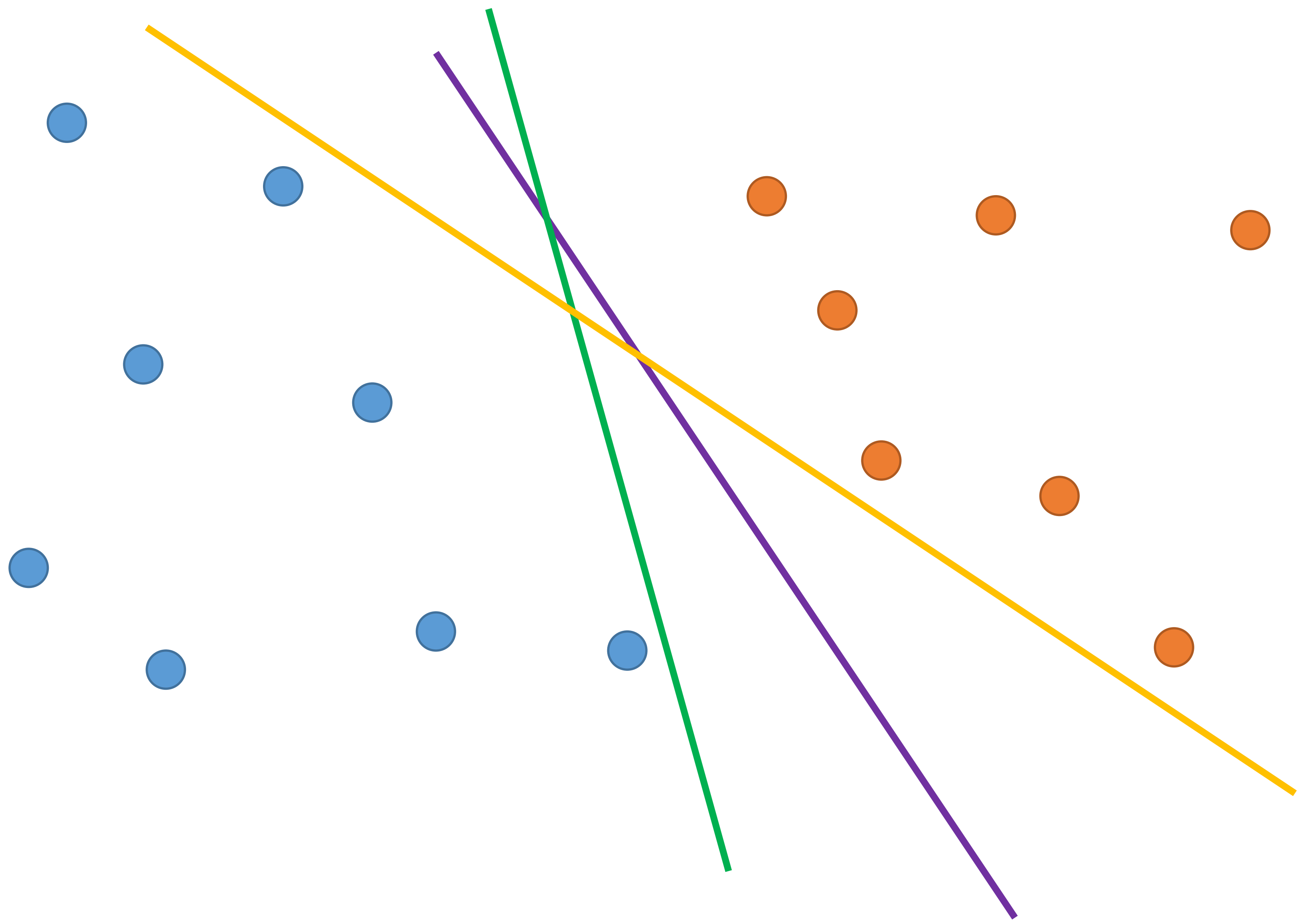
Hinge loss

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

- Минимизация верхней оценки:

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

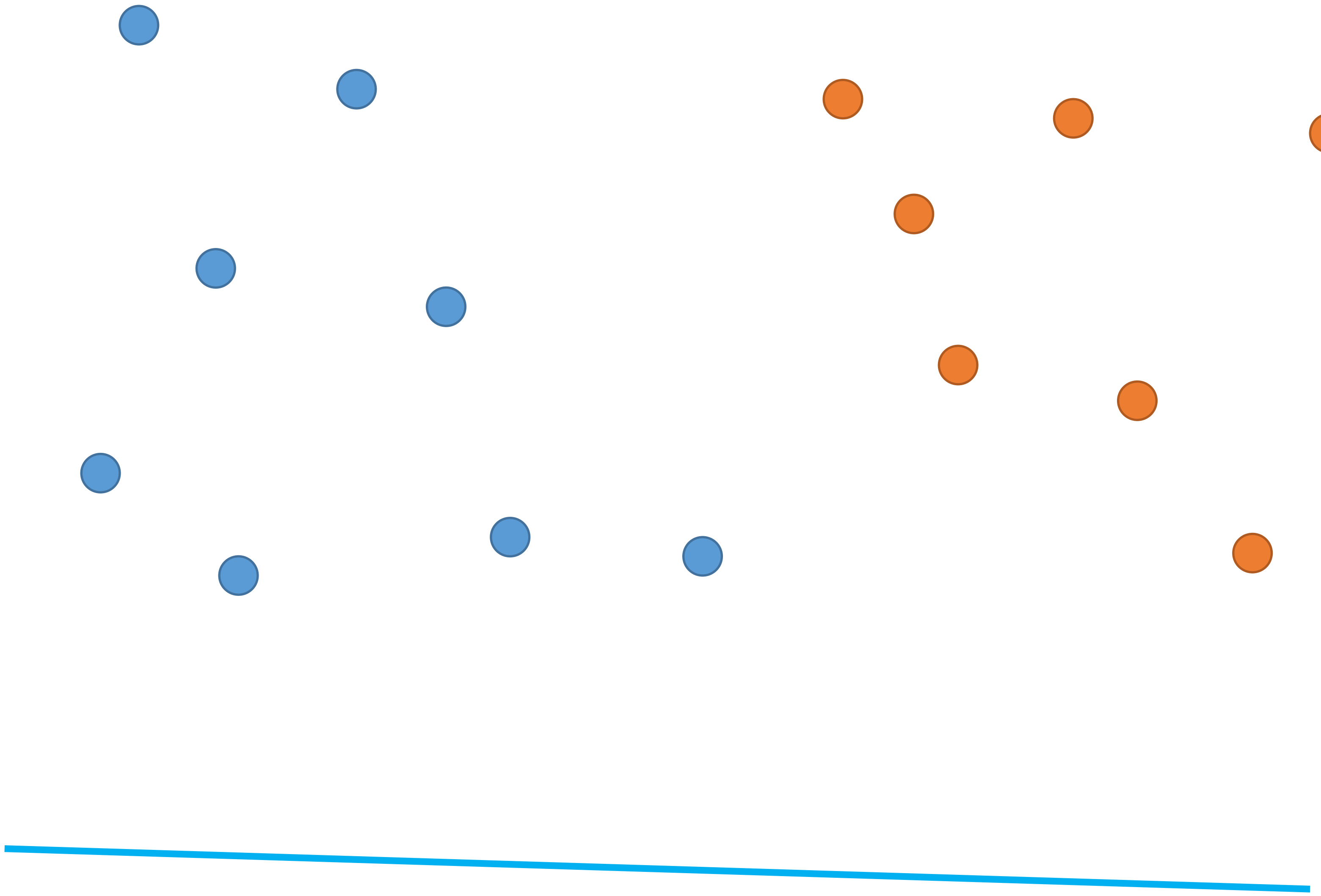
Какой классификатор лучше?



Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта

Отступ классификатора



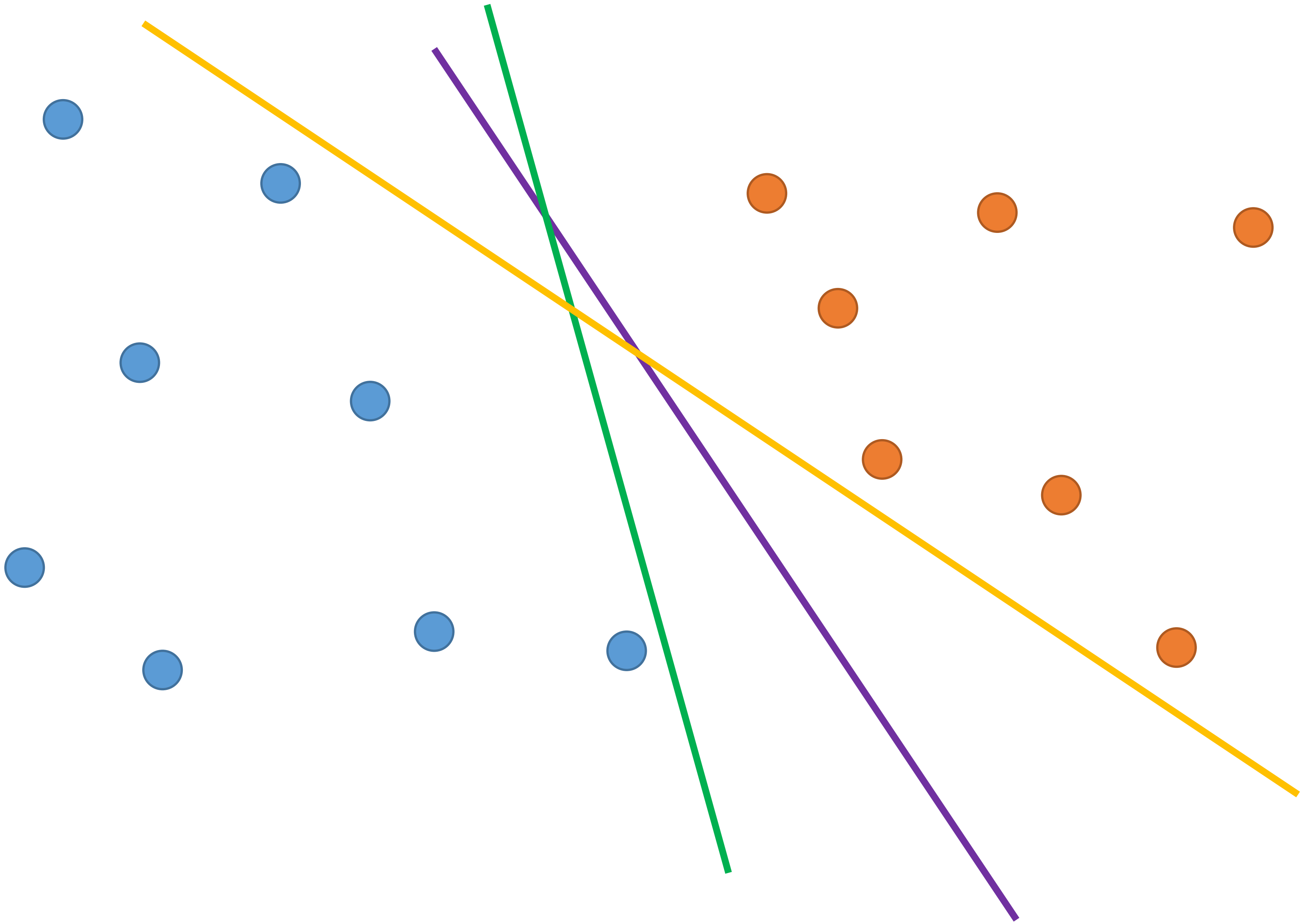
Отступ классификатора

- Будем максимизировать отступ классификатора — расстояние от гиперплоскости до ближайшего объекта
- При этом будет стараться сделать поменьше ошибок
- По сути, делаем как можно меньше предположений о модели, и верим, что это понизит вероятность переобучения

Простой случай

- Будем считать, что выборка линейно разделима
- Существует линейный классификатор, не допускающий ни одной ошибки

Линейно разделимый случай



Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

Отступ классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle + w_0 = 0$:

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

- Отступ классификатора:

$$\min_{i=1,\dots,\ell} \frac{|\langle w, x_i \rangle + w_0|}{\|w\|}$$

Небольшое предположение

- Линейный классификатор:

$$a(x) = \operatorname{sign}\left(\langle w, x_i \rangle + w_0\right)$$

- Если мы поделим w и w_0 на число $a > 0$, то выходы классификатора никак не поменяются:

$$a(x) = \operatorname{sign}\left(\frac{\langle w, x_i \rangle + w_0}{a}\right) = \operatorname{sign}\left(\langle w, x_i \rangle + w_0\right)$$

Небольшое предположение

- Поделим w и w_0 на $\min\limits_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| > 0$, после этого будет выполнено

$$\min_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| = 1$$

Отступ классификатора

- Расстояние от точки до гиперплоскости $\langle w, x \rangle + w_0 = 0$:

$$\frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

- Отступ классификатора:

$$\min_{i=1,\dots,\ell} \frac{|\langle w, x_i \rangle + w_0|}{\|w\|} = \frac{\min_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0|}{\|w\|} = \frac{1}{\|w\|}$$

Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

$$\frac{1}{\|w\|} \to \max_w$$

Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

$$\frac{1}{\|w\|} \to \max_w$$

- При условии, что $\min_{i=1,\dots,\ell} |\langle w, x_i \rangle + w_0| = 1$

Линейно разделимый случай

- **Требование 1**: $y_i(\langle w, x_i \rangle + w_0) > 0$ для всех $i = 1, \dots, \ell$
- **Требование 2**: максимальный отступ классификатора

$$\frac{1}{\|w\|} \to \max_w$$

- При условии, что $|\langle w, x_i \rangle + w_0| \geq 1$
- И мы минимизируем $\|w\|$ — тогда где-то модуль отступа будет равен 1

Метод опорных векторов (SVM)

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