Машинное обучение

Лекция 5
Линейная классификация

Андрей Нарцев
andrei.nartsev@gmail.com
anartsev@hse.ru

НИУ ВШЭ, 2024

Классификация

- $\mathbb{Y} = \{-1, +1\}$
- -1 — отрицательный класс
- $+1$ — положительный класс
- $a(x)$ должен возвращать одно из двух чисел

Линейная регрессия

$$a(x) = \boxed{w_0 + \sum_{j=1}^{d} w_j x_j}$$

Вещественное число!

Линейный классификатор

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

Линейный классификатор

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

Свободный коэффициент

Веса

Признаки

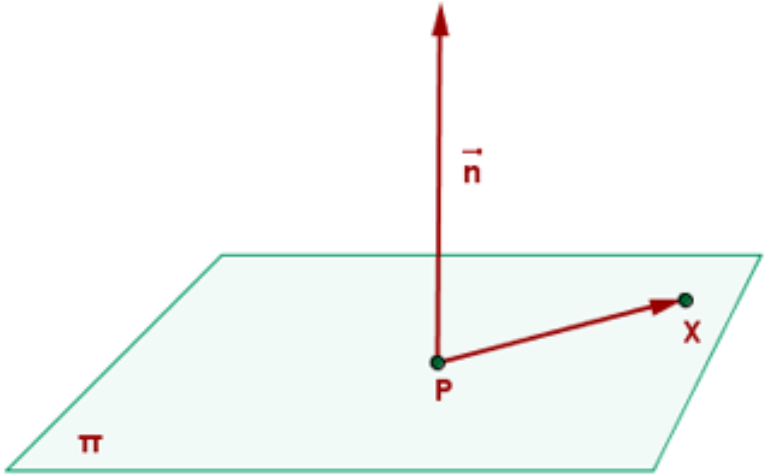
Линейный классификатор

- Будем считать, что есть единичный признак

$$a(x) = \operatorname{sign} \sum_{j=1}^{d} w_j x_j = \operatorname{sign} \langle w, x \rangle$$

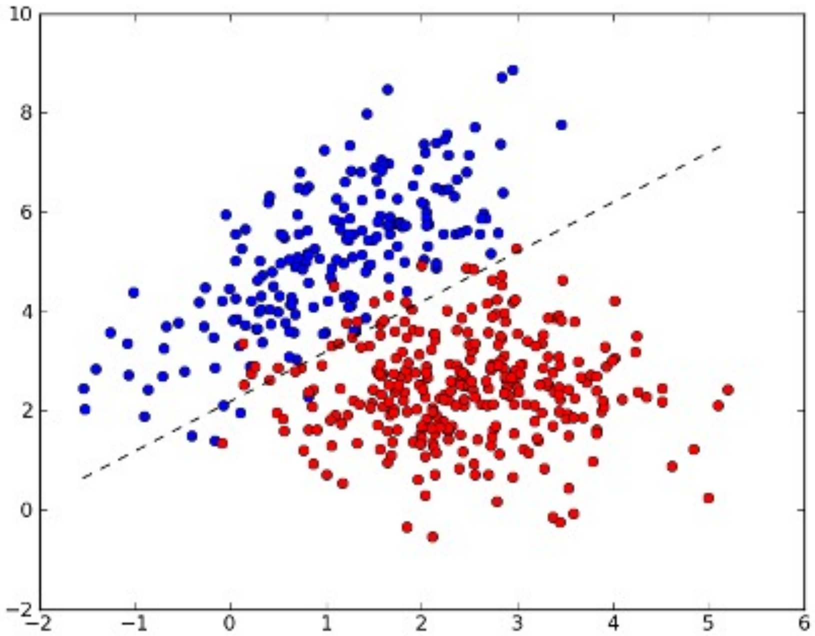
Геометрия линейного классификатора

Уравнение гиперплоскости: $\langle w, x \rangle = 0$



Геометрия линейного классификатора

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



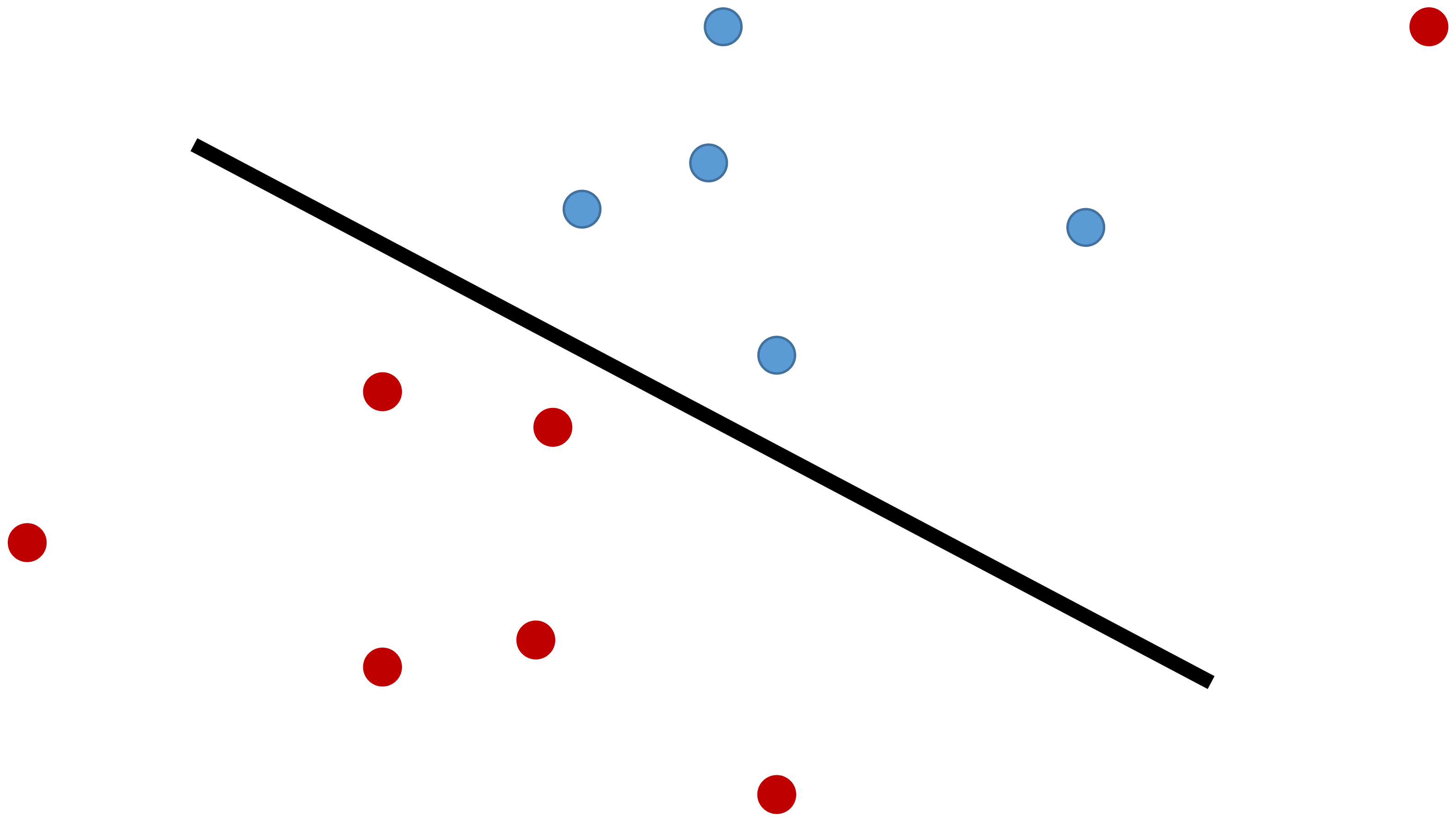
Геометрия линейного классификатора

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

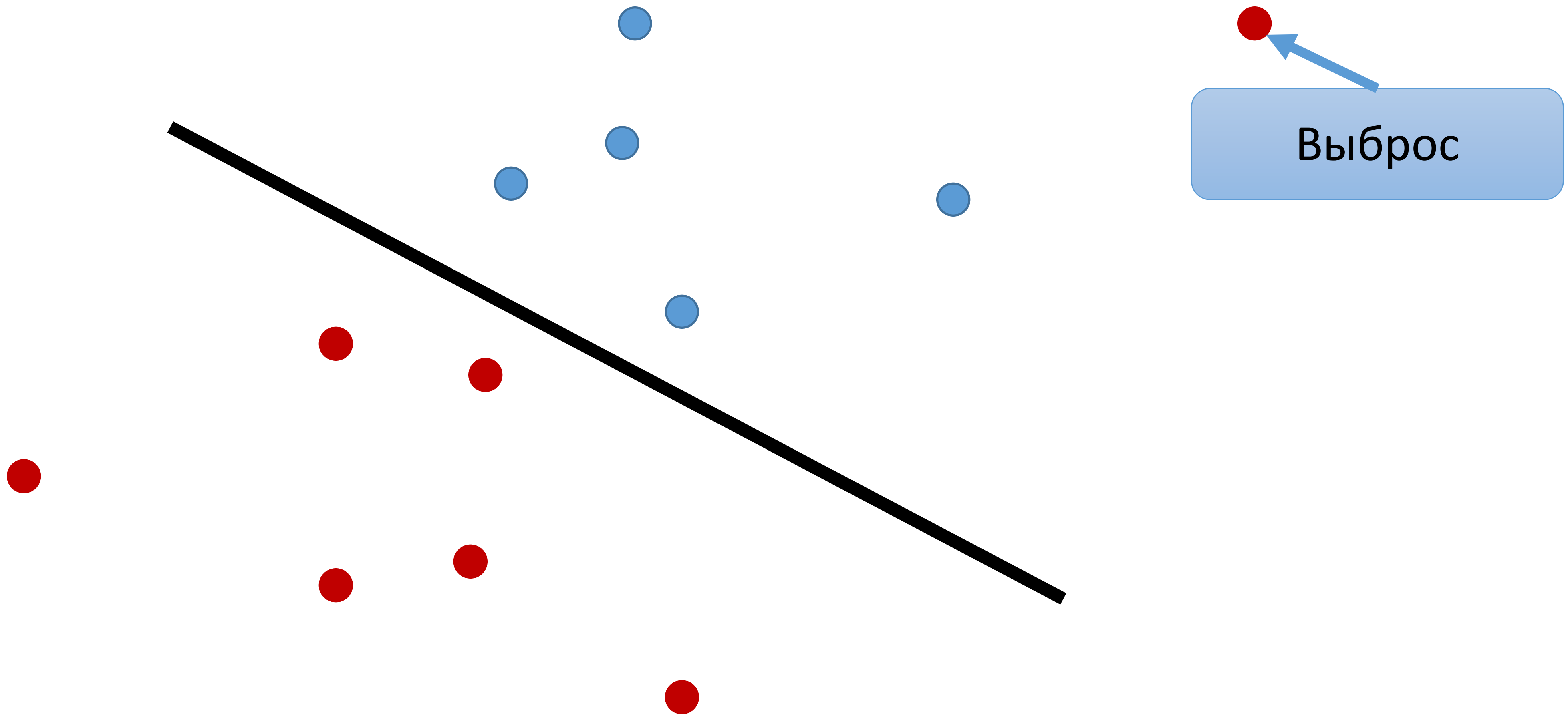
$$\frac{|\langle w, x \rangle|}{\|w\|}$$

- Чем больше $\langle w, x \rangle$, тем дальше объект от разделяющей гиперплоскости

Геометрия линейного классификатора

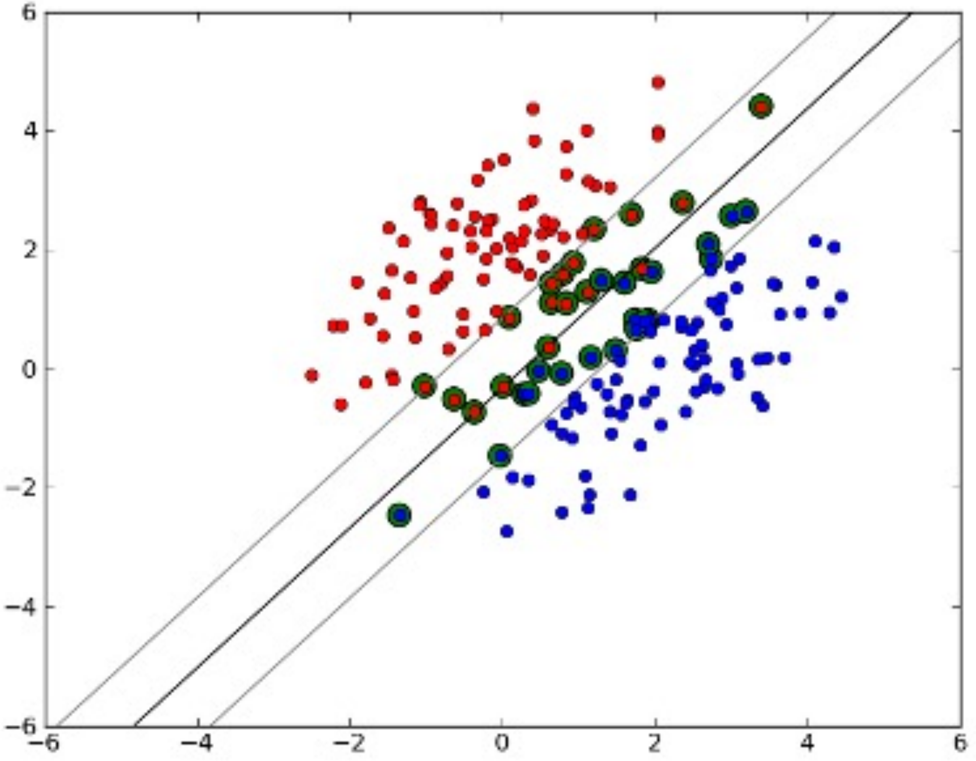


Геометрия линейного классификатора



Отступ

- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$ — классификатор дает верный ответ
- $M_i < 0$ — классификатор ошибается
- Чем дальше отступ от нуля, тем больше уверенности



Порог

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

- t — порог классификатора
- Можно подбирать для оптимизации функции потерь, отличной от использованной при обучении

Линейный классификатор

- Линейный классификатор разделяет два класса гиперплоскостью
- Чем больше отступ по модулю, тем дальше объект от гиперплоскости
- Знак отступа говорит о корректности предсказания

Обучение линейных классификаторов

Функция потерь в классификации

- Частый выбор — бинарная функция потерь

$$L(y, a) = [a \neq y]$$

- Функционал ошибки — доля ошибок (error rate)

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

- Нередко измеряют долю верных ответов (accuracy):

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

Доля ошибок для линейного классификатора

- Функционал ошибки:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\operatorname{sign}\,(\langle w, x_i \rangle) \neq y_i]$$

- Индикатор — недифференцируемая функция

Отступы для линейного классификатора

- Функционал ошибки:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\operatorname{sign}(\langle w, x_i \rangle) \neq y_i]$$

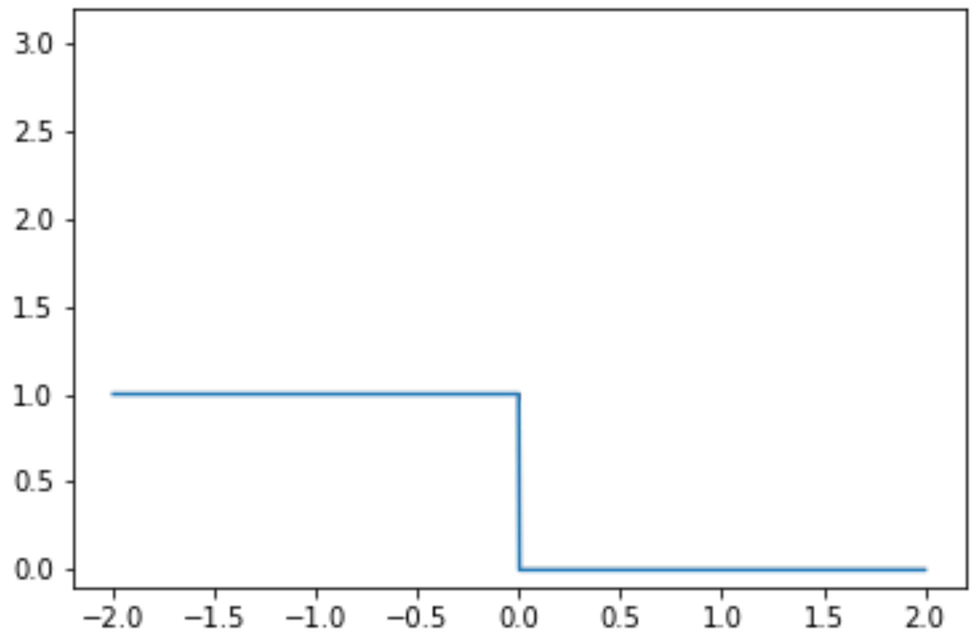
- Альтернативная запись:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0]$$

Отступы для линейного классификатора

$$L(M) = [M < 0]$$

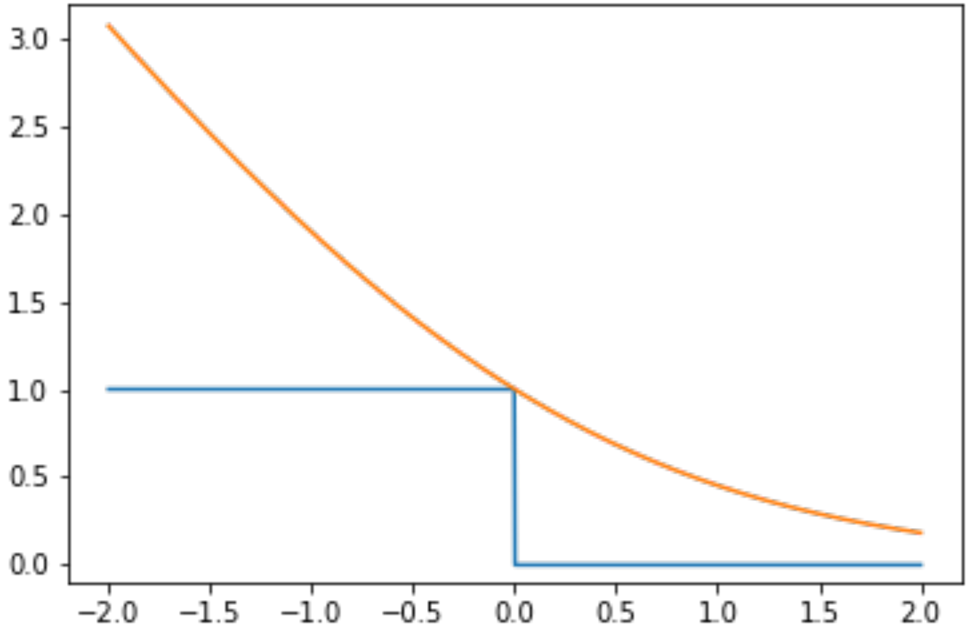
- Нельзя продифференцировать



Верхняя оценка

$$L(M) = [M < 0] \leq \tilde{L}\,(M)$$

- Оценим сверху дифференцируемой функцией



Верхняя оценка

$$0 \le \frac{1}{\ell}\sum_{i=1}^{\ell}[y_i\langle w, x_i\rangle < 0] \le \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i\langle w, x_i\rangle) \to \min_w$$

- Минимизируем верхнюю оценку
- Надеемся, что она прижмёт долю ошибок к нулю

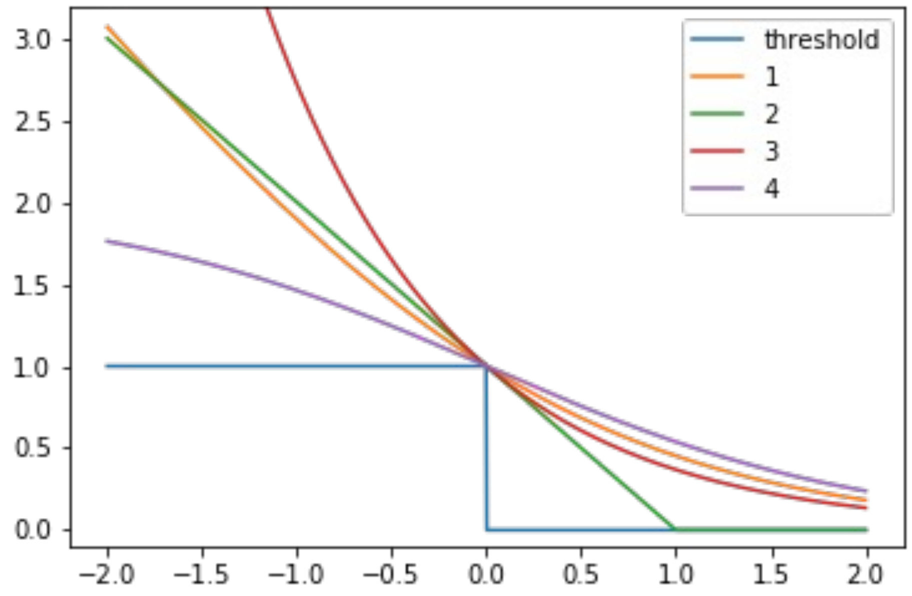
Примеры верхних оценок

1. $\tilde{L}(M) = \log(1 + e^{-M})$ — логистическая

2. $\tilde{L}(M) = \max(0, 1 - M)$ — кусочно-линейная

3. $\tilde{L}(M) = e^{-M}$ — экспоненциальная

4. $\tilde{L}(M) = \frac{2}{1+e^{M}}$ — сигмоидная



Пример обучения

- Выбираем логистическую функцию потерь:

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

- Вычисляем градиент:

$$\nabla_w \tilde{Q}(w, X) = -\frac{1}{\ell} \sum_{i=1}^{\ell} \frac{y_i x_i}{1 + \exp(y_i \langle w, x_i \rangle)}$$

Пример обучения

- Делаем градиентный спуск:

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