Занятие 4 Градиентный спуск. Линейные модели классификации.

Елена Кантонистова

elena.kantonistova@yandex.ru

ВШЭ, 2021

ПЛАН ЗАНЯТИЯ

- Градиентный спуск
- Линейные классификаторы
- Логистическая регрессия
- Практика

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ (НАПОМИНАНИЕ)

Обучающая выборка:

пусть $\boldsymbol{x}$ – объект ($x_1, x_2, \dots, x_l$ - его признаки), а y – ответ на объекте (произвольное число).

Модель линейной регрессии:

$$a(\boldsymbol{x}, w) = w_0 + w_1 x_1 + w_2 x_2 + .. = \sum_{i=1}^{l} w_j x_j$$

ГРАДИЕНТНЫЙ СПУСК

ЛИНЕЙНАЯ РЕГРЕССИЯ

Обучение линейной регрессии - минимизация среднеквадратичной ошибки:

$$\mathrm{Q}(a, X) = \frac{1}{l} \sum_{i=1}^{l} (a(x_i) - y_i)^2 \rightarrow \min_{w}$$

ЛИНЕЙНАЯ РЕГРЕССИЯ

Обучение линейной регрессии - минимизация среднеквадратичной ошибки:

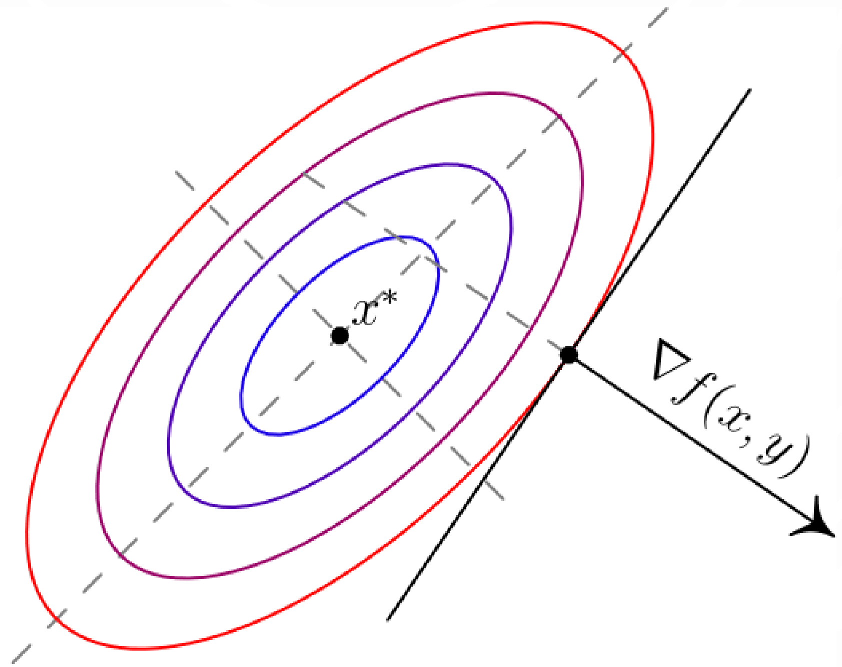
$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 \to \min_{w}$$

Обучение других алгоритмов машинного обучения – это тоже минимизация какой-то (может быть, очень сложной) функции потерь!

ТЕОРЕМА О ГРАДИЕНТЕ

Теорема. Градиент – это вектор, в направлении которого функция быстрее всего растёт.

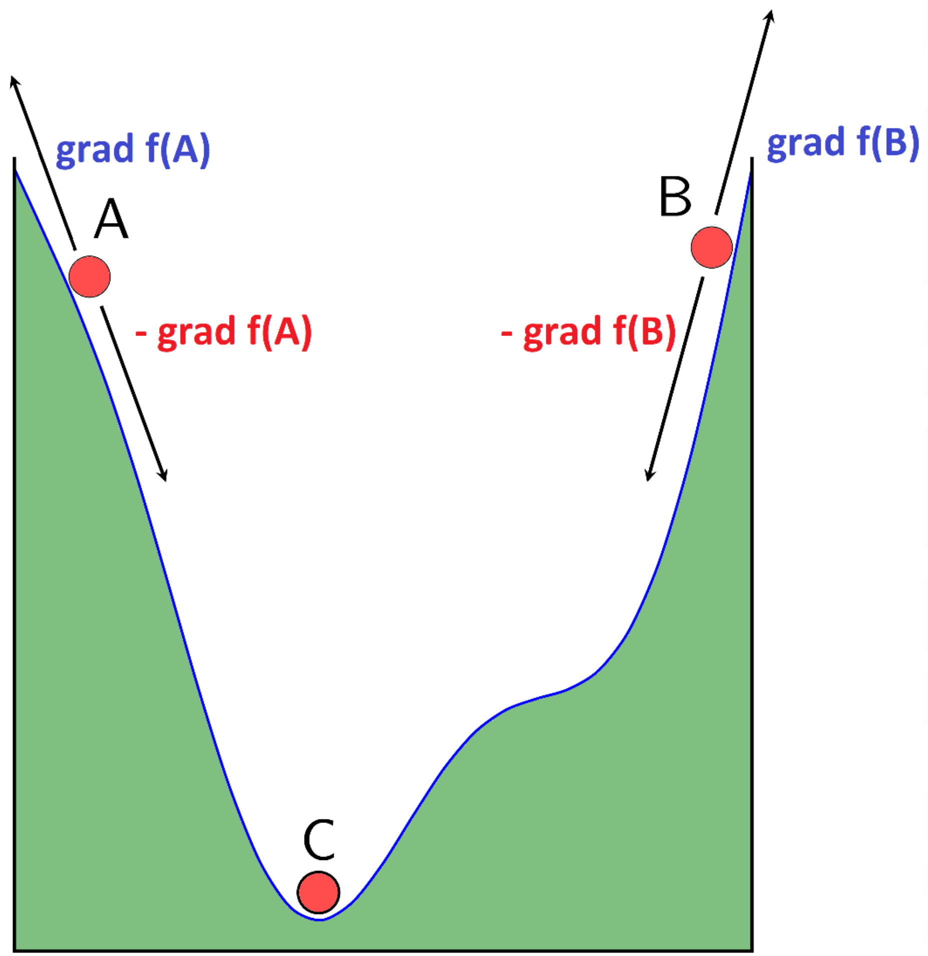
Антиградиент (вектор, противоположный градиенту) – вектор, в направлении которого функция быстрее всего убывает.



$$f(x,y) = \frac{3}{2}(x-y)^2 + \frac{1}{3}(x+y)^2$$

ТЕОРЕМА О ГРАДИЕНТЕ

Аантиградиент (вектор, противоположный градиенту) – вектор, в направлении которого функция быстрее всего убывает.

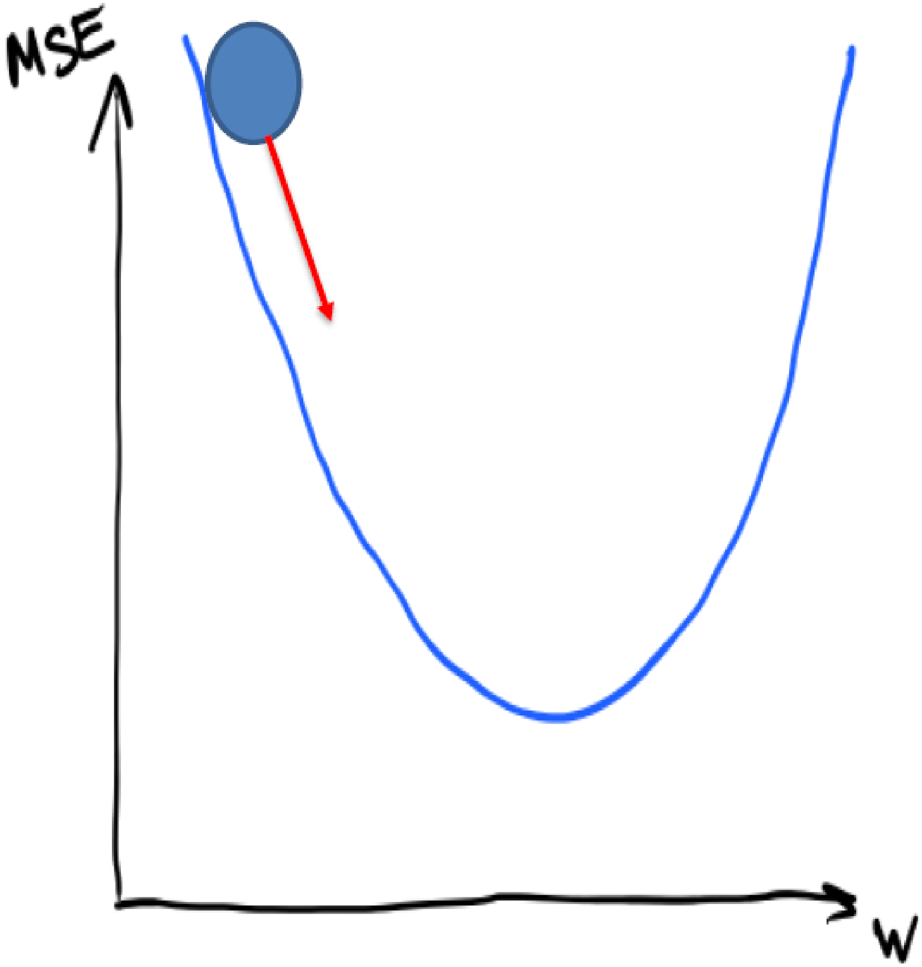


МЕТОД ГРАДИЕНТНОГО СПУСКА

- Наша задача при обучении модели **–** найти такие веса $\boldsymbol{w}$**,** на которых достигается **минимум функции ошибки**.

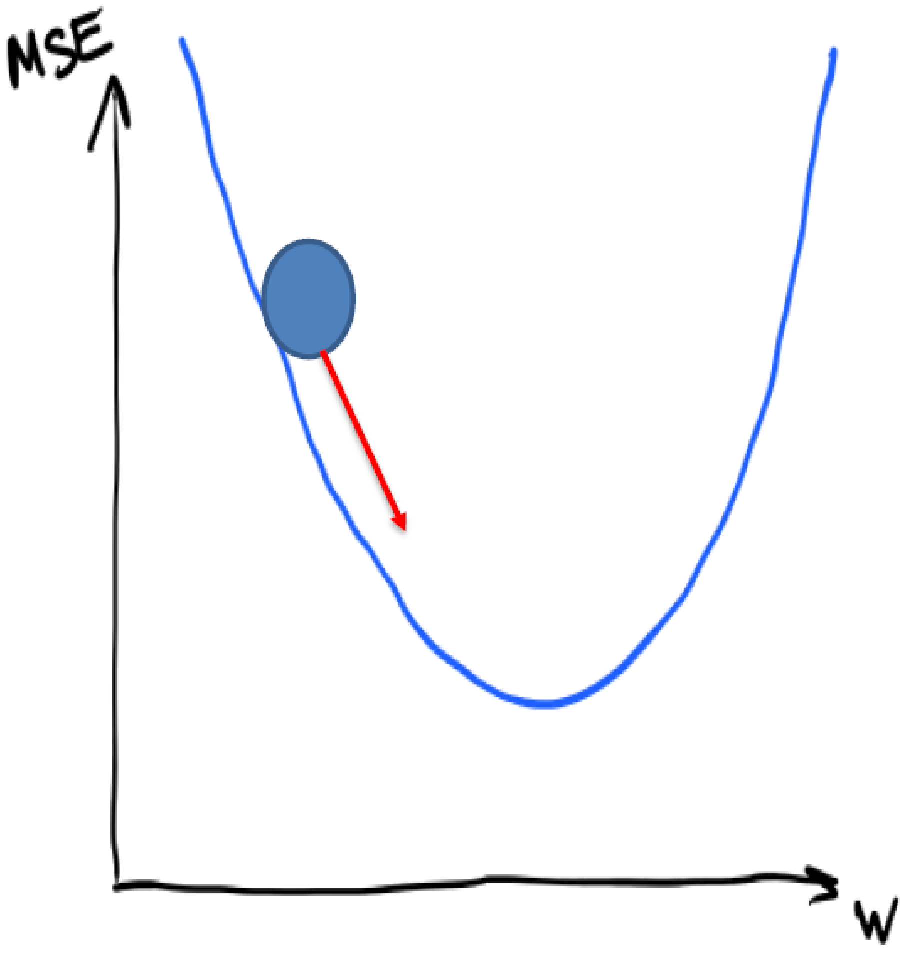
МЕТОД ГРАДИЕНТНОГО СПУСКА

На каждом шаге (на каждой итерации метода) движемся в сторону антиградиента функции потерь!



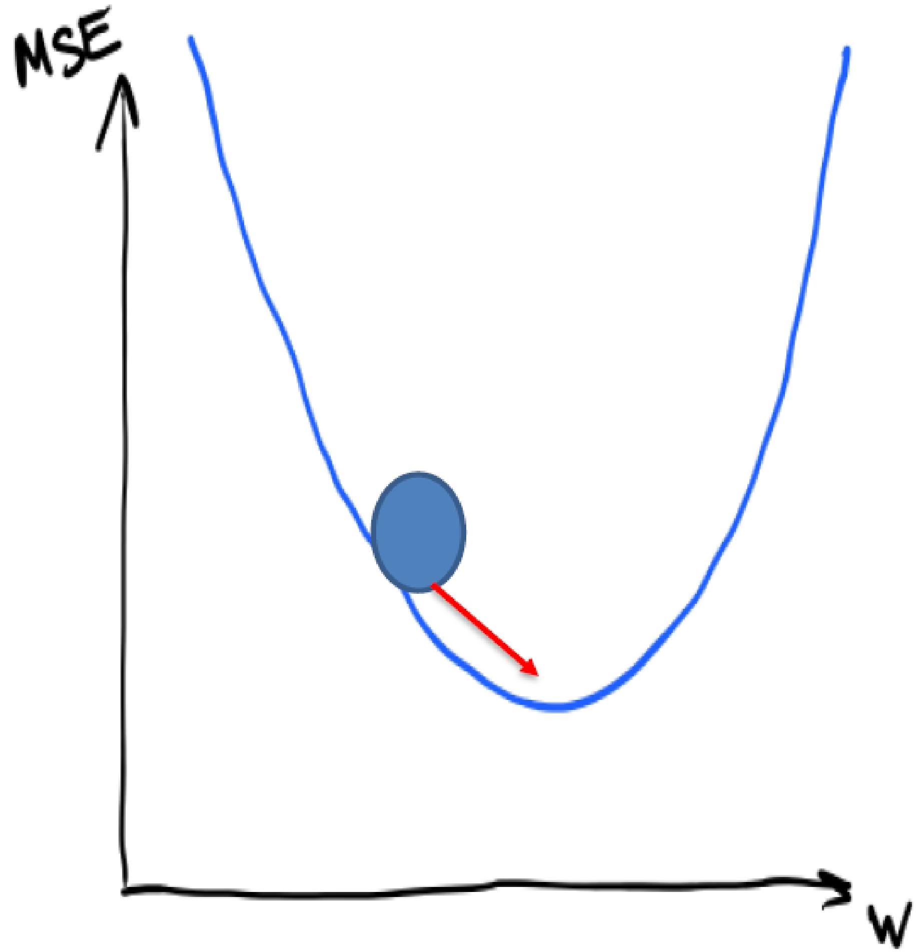
МЕТОД ГРАДИЕНТНОГО СПУСКА

На каждом шаге (на каждой итерации метода) движемся в сторону антиградиента функции потерь!



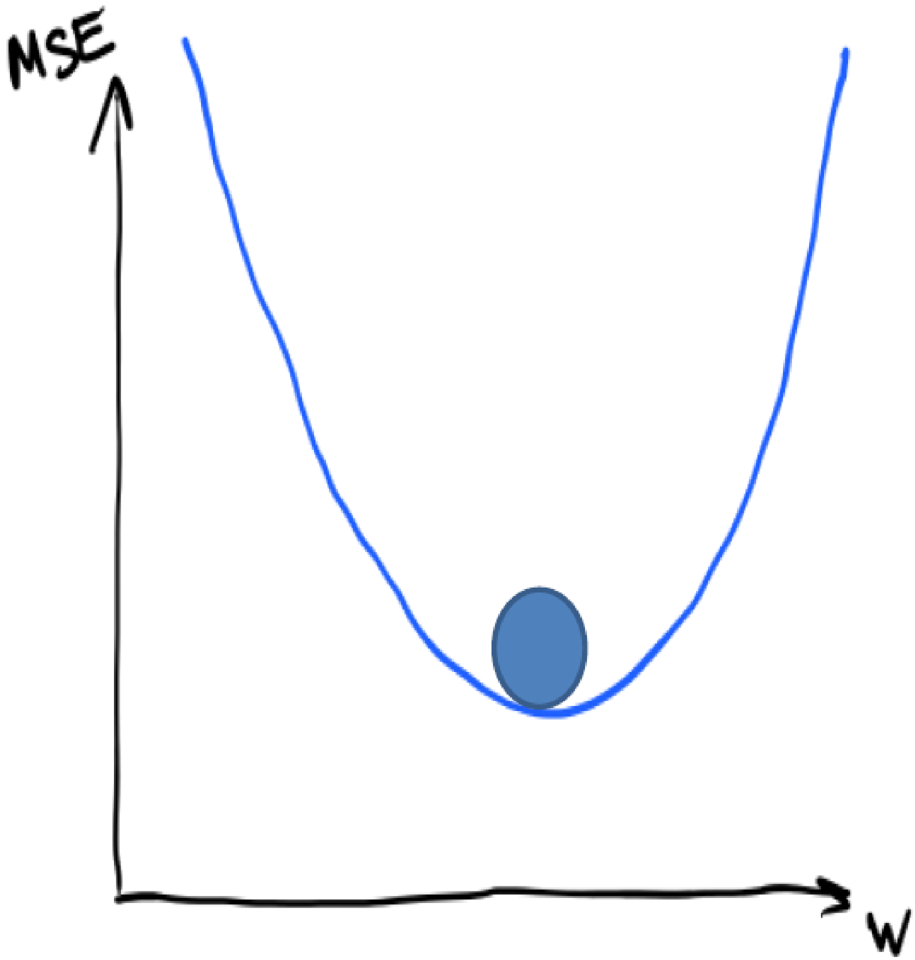
МЕТОД ГРАДИЕНТНОГО СПУСКА

На каждом шаге (на каждой итерации метода) движемся в сторону антиградиента функции потерь!



МЕТОД ГРАДИЕНТНОГО СПУСКА

На каждом шаге (на каждой итерации метода) движемся в сторону антиградиента функции потерь!



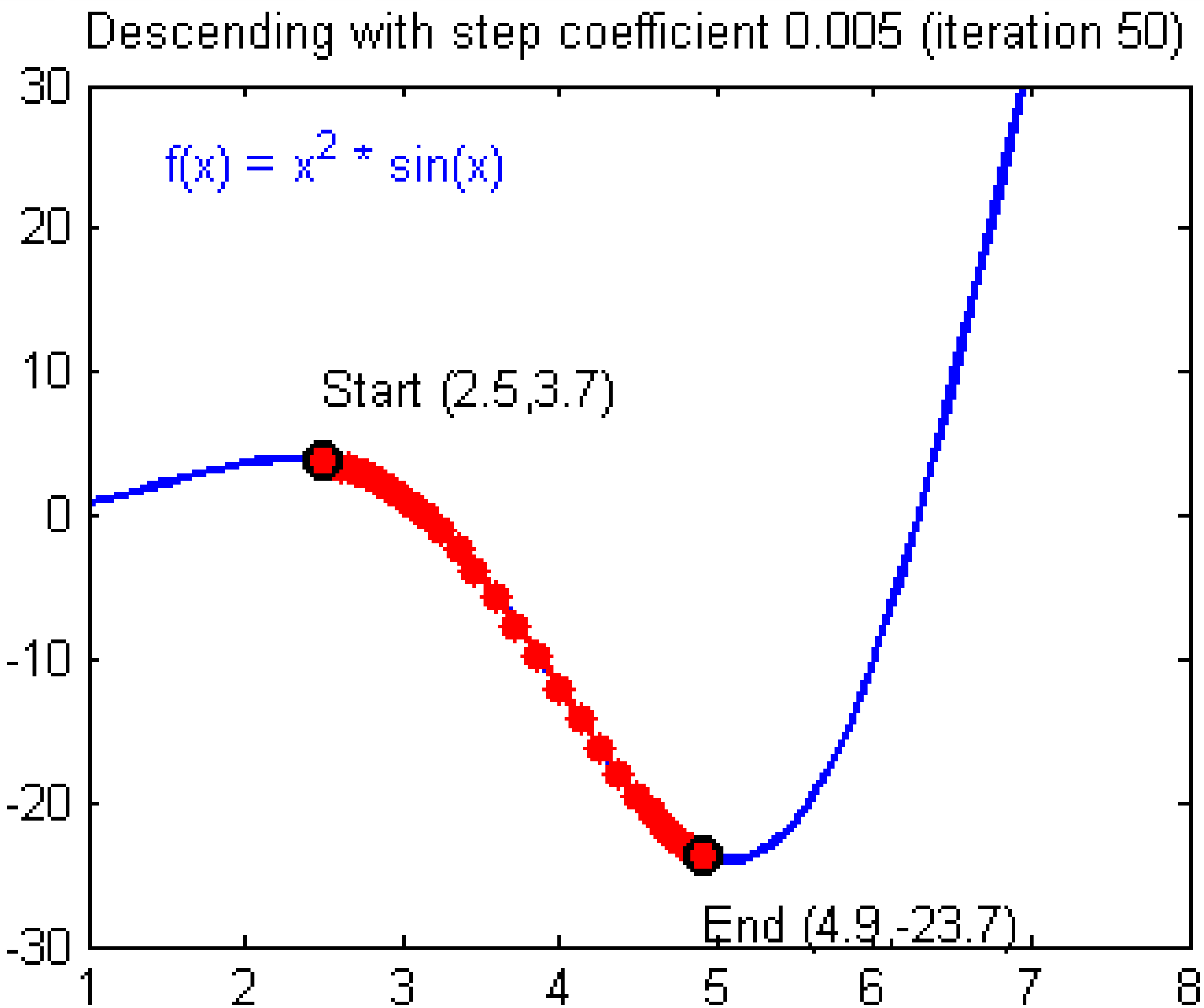
Метод градиентного спуска

Формулу для обновления весов можно записать в векторном виде:

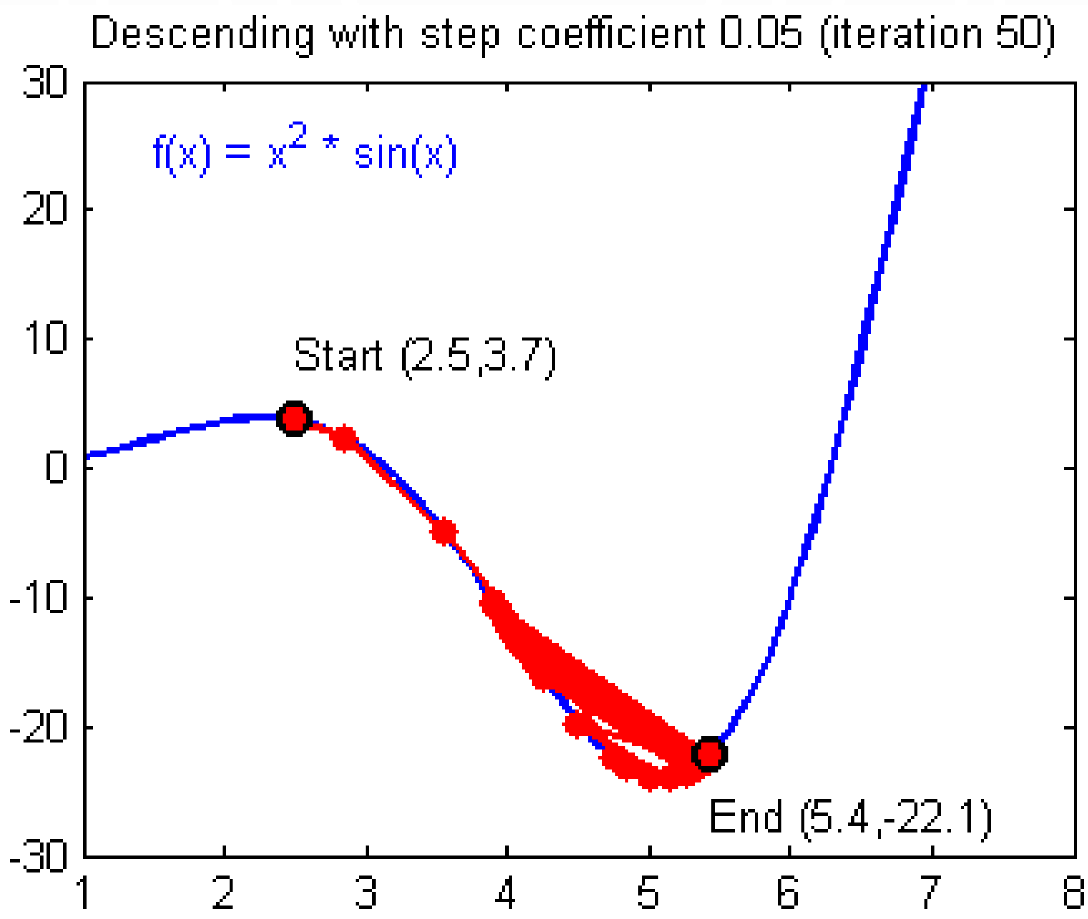
- Инициализируем веса $\boldsymbol{w}^{(0)}$.
- На каждом следующем шаге обновляем веса по формуле:

$$\boldsymbol{w}^{(k)} = \boldsymbol{w}^{(k-1)} - \eta \nabla Q\left(\boldsymbol{w}^{(k-1)}\right)$$

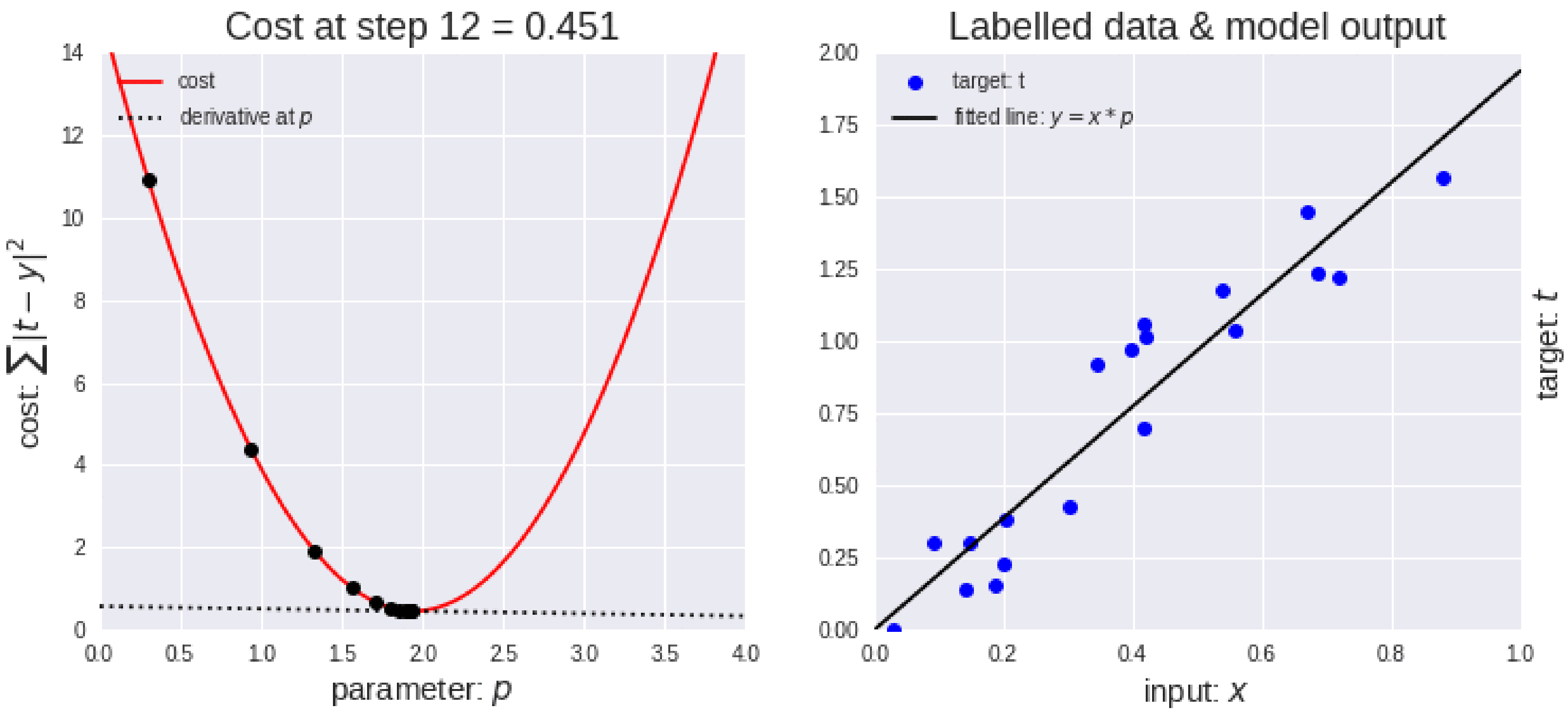
ГРАДИЕНТНЫЙ СПУСК



ПРОБЛЕМА ВЫБОРА ГРАДИЕНТНОГО ШАГА



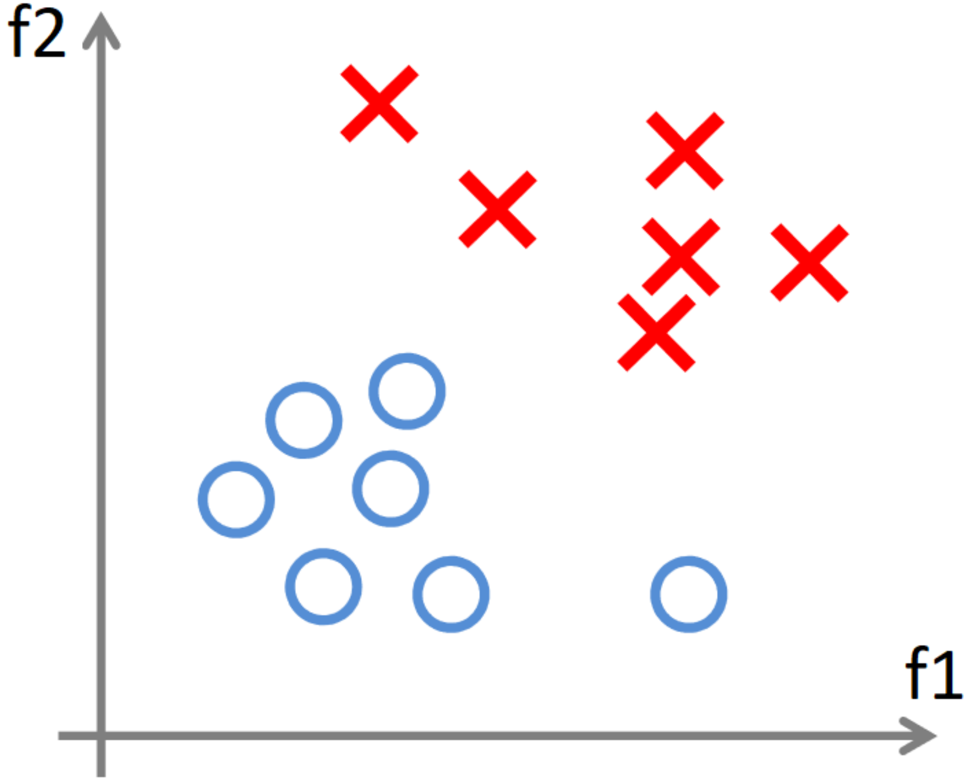
ДЕМО: ЛИНЕЙНАЯ РЕГРЕССИЯ



ЛИНЕЙНАЯ КЛАССИФИКАЦИЯ

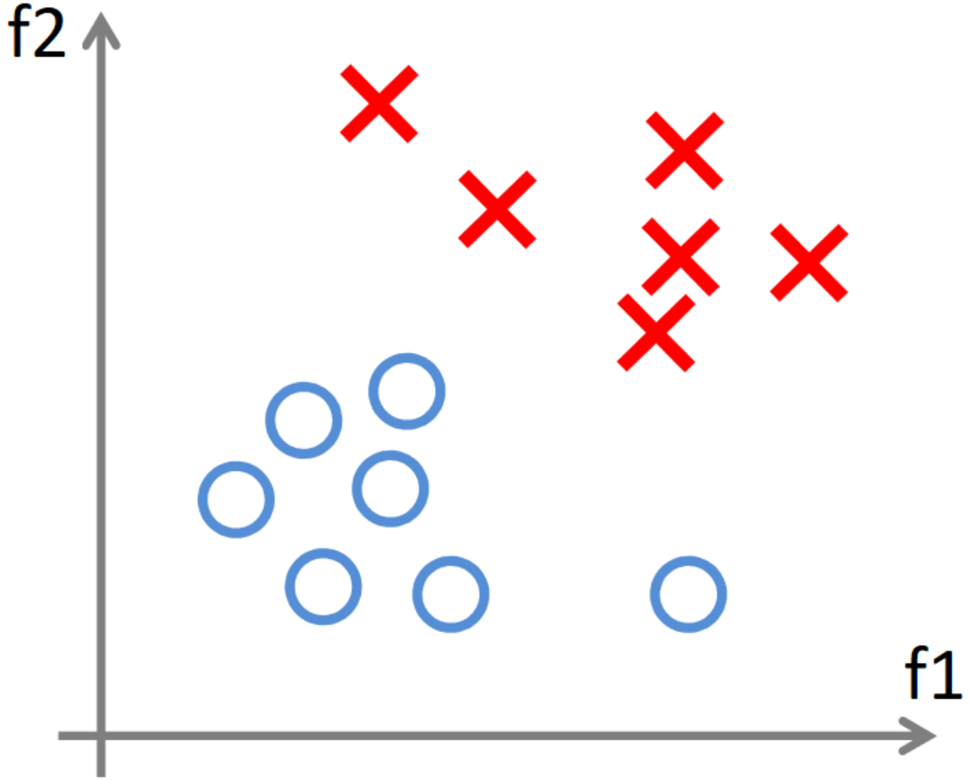
БИНАРНАЯ КЛАССИФИКАЦИЯ

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



БИНАРНАЯ КЛАССИФИКАЦИЯ

$y_1, y_2, \dots, y_n$ - ответы (***+1 или -1***).



Как выглядит модель линейного классификатора: $\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{w}) = ?$

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

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

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

$$\boldsymbol{a}(\boldsymbol{x},\boldsymbol{w}) = \boldsymbol{sign}\Big(\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{l}} \boldsymbol{w}_{\boldsymbol{j}}\boldsymbol{x}_{\boldsymbol{j}}\Big)$$

- Если сумма > 0, то $sign$(суммы) $= +1$, то есть объект отнесён к положительному классу
- если сумма < 0, то $sign$(суммы) $= -1$, то есть объект отнесён к отрицательному классу

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

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

- Если сумма > 0, то $sign$(суммы) $= +1$, то есть объект отнесён к положительному классу
- если сумма < 0, то $sign$(суммы) $= -1$, то есть объект отнесён к отрицательному классу
- значит, $\sum_{j=1}^{l} w_j x_j = 0$ – *уравнение разделяющей границы* между классами. *Это уравнение плоскости* (или прямой в двумерном случае), поэтому *классификатор является линейным*.

БИНАРНАЯ КЛАССИФИКАЦИЯ

Модель линейного классификатора:

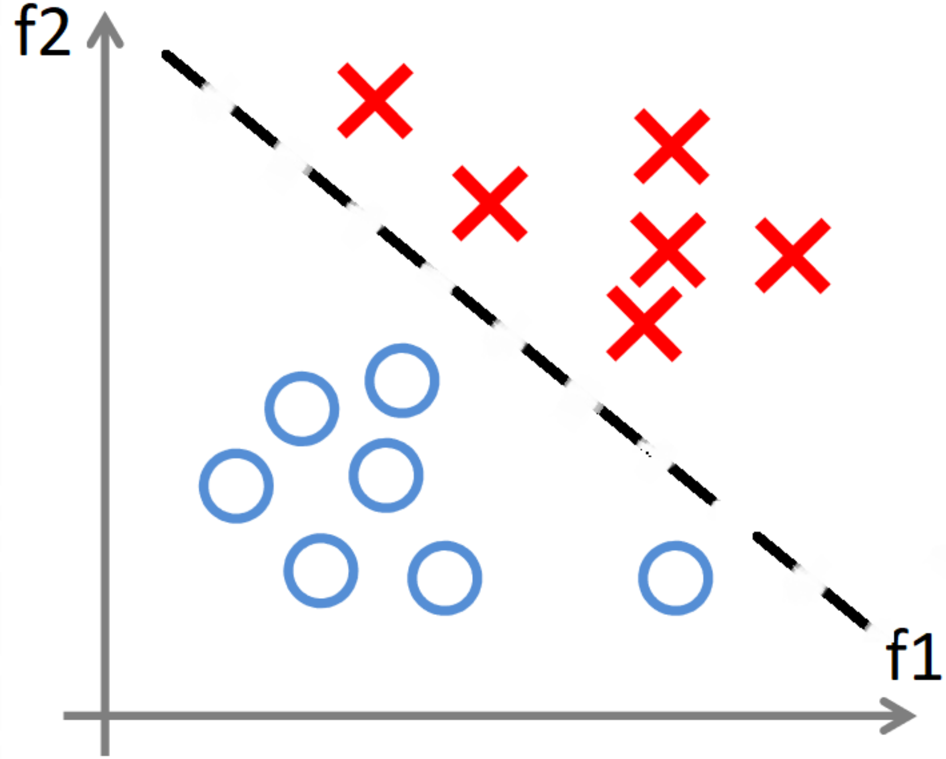
$$a(x,w) = sign\left(\sum_{j=1}^{l} w_j x_j\right)$$

Уравнение

$\sum_{j=1}^{l} w_j x_j = 0$

– уравнение плоскости

(или прямой).



Обучение классификатора

Как обучить линейный классификатор?

ОБУЧЕНИЕ КЛАССИФИКАТОРА

- Обучение - минимизация доли ошибок классификатора:

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

где $[a(\boldsymbol{x_i}) \neq y_i] = 1$, если предсказание на объекте неверное, то есть $a(\boldsymbol{x_i}) \neq y_i$, и 0 иначе.

ОБУЧЕНИЕ КЛАССИФИКАТОРА

- Обучение - минимизация доли ошибок классификатора:

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

где $[a(\boldsymbol{x_i}) \neq y_i] = 1$, если предсказание на объекте неверное, то есть $a(\boldsymbol{x_i}) \neq y_i$, и 0 иначе.

Для обучения классификатора можно использовать и другие функции потерь. Дальше рассмотрим несколько классических моделей классификации.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Хотим предсказывать не только классы, но и вероятности классов.

- Линейная регрессия: $a(x, w) = (x, w)$
- Логистическая регрессия: $\boldsymbol{a(x, w) = \sigma((x, w))}$,

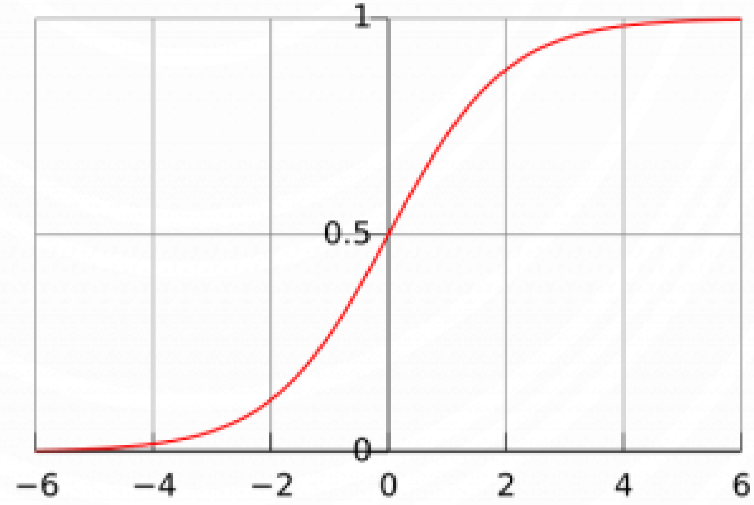
где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоида (логистическая функция)

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

- Линейная регрессия: $a(x, w) = (x, w) \in \mathbb{R}$
- Логистическая регрессия: $\boldsymbol{a(x, w) = \sigma((x, w))}$,

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоида (логистическая функция),

$\sigma(z) \in (0; 1)$.



Логистическая регрессия: $\boldsymbol{a(x, w) = \frac{1}{1+e^{-(x,w)}}}$

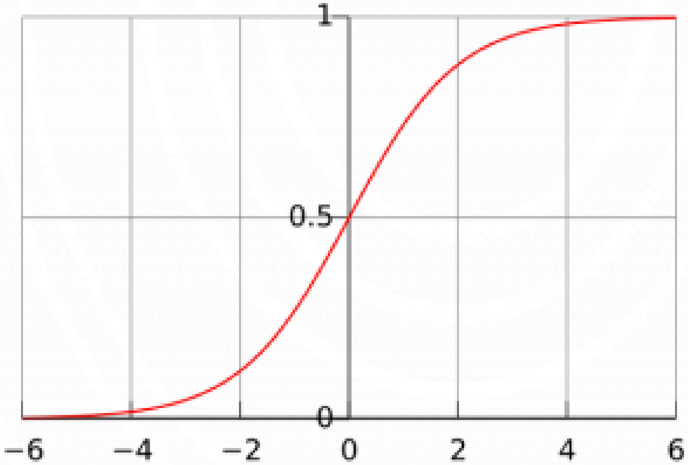
ВЕРОЯТНОСТНЫЙ СМЫСЛ

- $a(x,w)$ – вероятность того, что $y = +1$ на объекте x, то есть

$$\boldsymbol{a(x,w) = P(y = +1|x;w)}$$

РАЗДЕЛЯЮЩАЯ ГРАНИЦА

Предсказываем $y = +1$, если $a(x, w) \geq 0.5$.



$a(x, w) = g((x, w)) \geq 0.5$, если $(x, w) \geq 0$.

Получаем, что

- $y = +1$ при $(x, w) \geq 0$
- $y = -1$ при $(x, w) < 0$,

т.е. $(x, w) = 0$ – разделяющая гиперплоскость.

Логистическая регрессия

Логистическая регрессия - это линейный классификатор!

ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Если взять квадратичную функцию потерь

$L(a, y) = (a - y)^2$,

то возникнут проблемы:

- $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{1+e^{-(x,w)}} - y\right)^2$ - не выпуклая функция (*можем не попасть в глобальный минимум при оптимизации*)

ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Если взять квадратичную функцию потерь

$L(a, y) = (a - y)^2$,

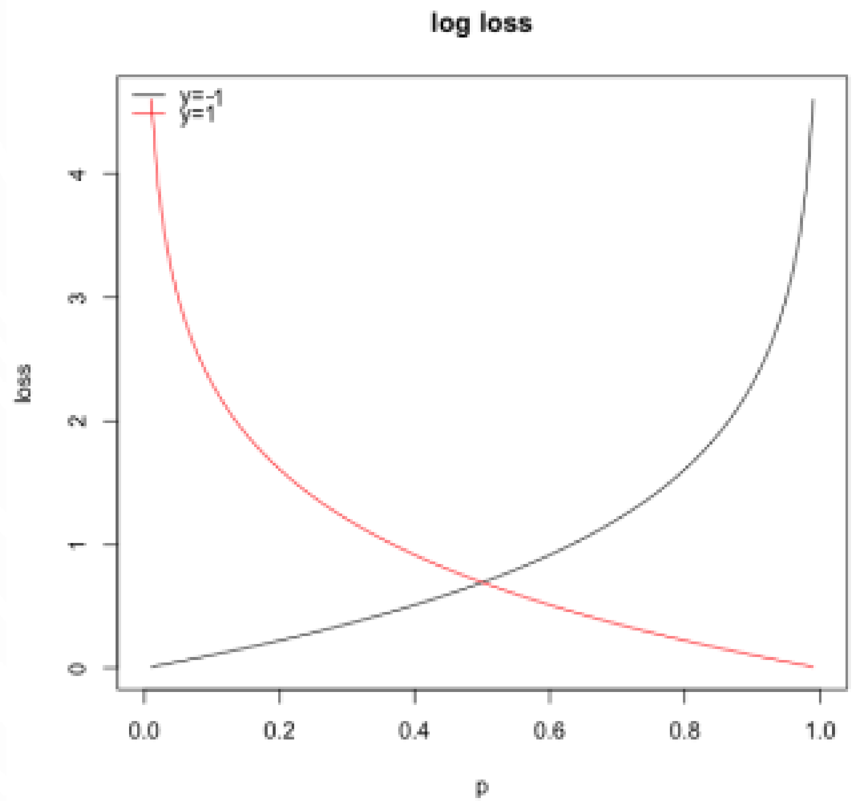
то возникнут проблемы:

- $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{1}{1+e^{-(x,w)}} - y\right)^2$ - не выпуклая функция (*можем не попасть в глобальный минимум при оптимизации*)
- *На совсем неправильном предсказании маленький штраф* (пусть предсказали вероятность 0% на объекте класса $y = +1$, тогда штраф всего $(1 - 0)^2 = 1$)

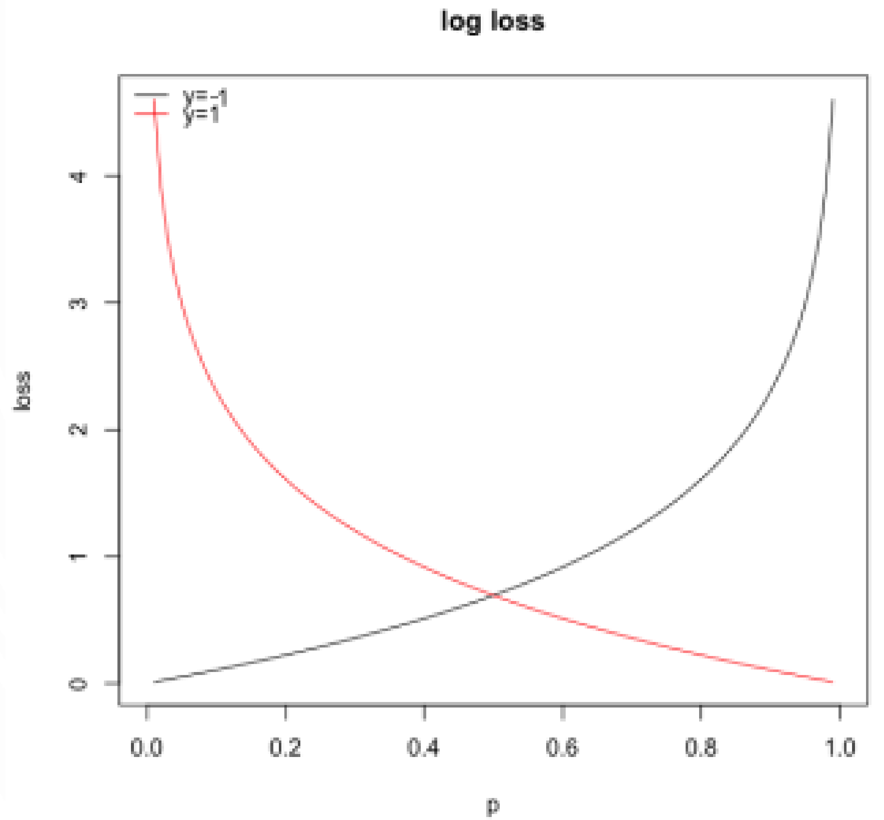
ФУНКЦИЯ ПОТЕРЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Возьмем логистическую функцию потерь (log-loss):

$$Q(w) = -\sum_{i=1}^{l}\left([y_i = +1] \cdot \log\big(a(x_i, w)\big) + [y_i = -1] \cdot \log\big(1 - a(x_i, w)\big)\right)$$



ЛОГИСТИЧЕСКАЯ ФУНКЦИЯ ПОТЕРЬ



- если $a(x,w) = 1$ и $y = +1$, то штраф $L(a,y) = 0$
- если $a(x,w) \to 0$, а $y = +1$, то штраф $L(a,y) \to +\infty$