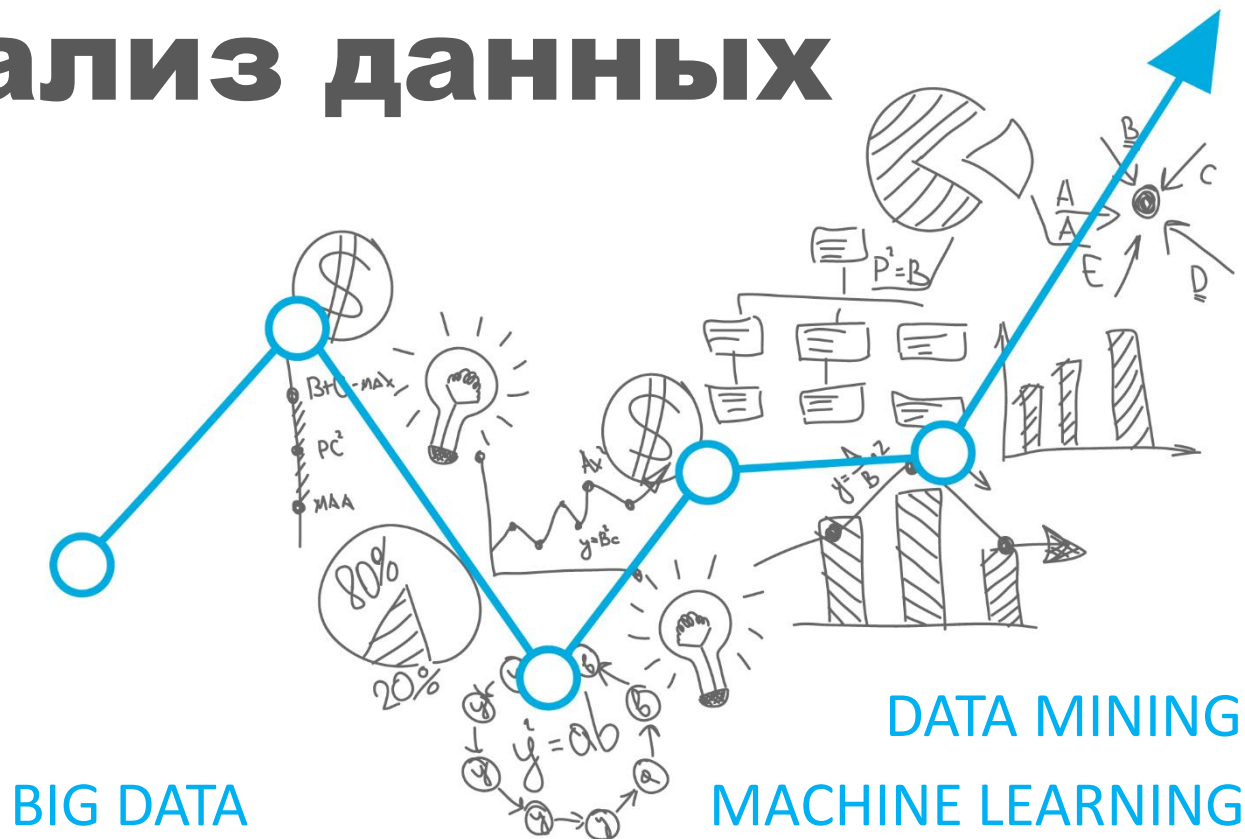
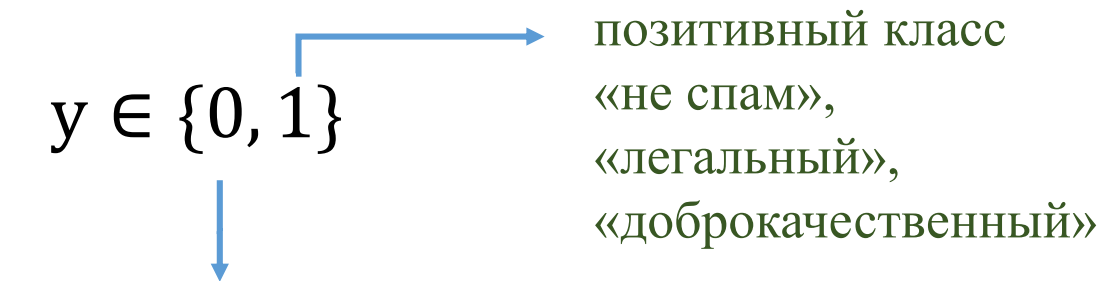


Интеллектуальный анализ данных



Лекция 4. Логистическая регрессия

- Электронная почта: **спам** / не спам
- Онлайн сайты: **вредоносный** / легальный
- Диагностика рака: **злокачественный** / доброкачественный



негативный класс
«спам»,
«вредоносный»,
«злокачественный»

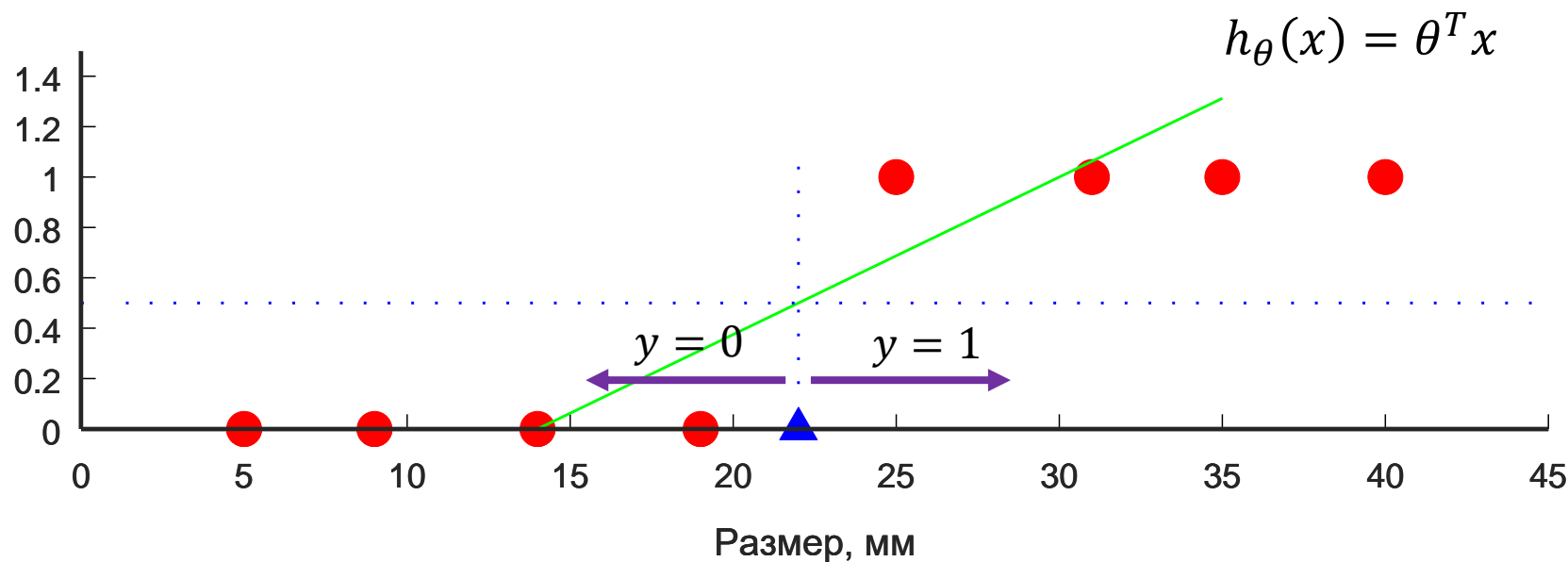
$$y \in \{A, B\} \quad y \in \{-, +\}$$

$y \in \{0, 1, 2, 3, 4\}$ –
многоклассовая классификация
(будет рассмотрена позднее)

Диагностика рака:

$y = 1$ – злокачественный

$y = 0$ – доброкачественный



Пороговое условие: $h_{\theta}(x) \geq 0.5 \rightarrow y = 1$

$h_{\theta}(x) < 0.5 \rightarrow y = 0$

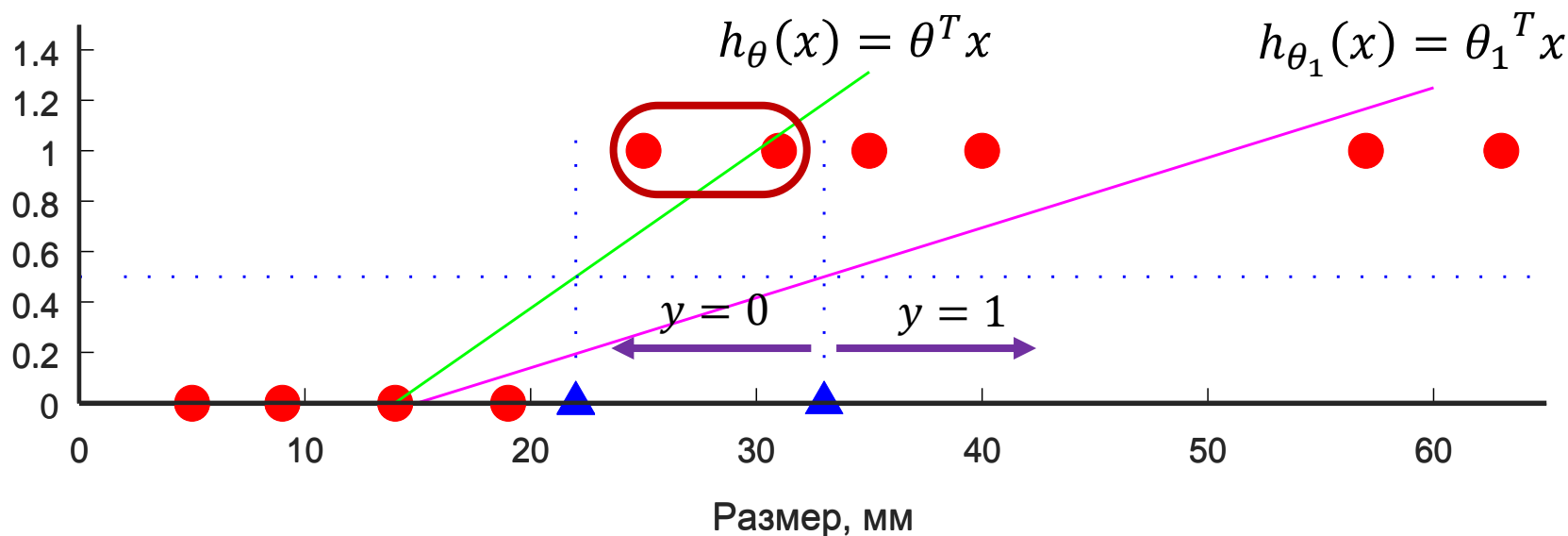
Можно ли использовать линейную регрессию для классификации?

В принципе, да, но не в таком виде!

Диагностика рака:

$y = 1$ – злокачественный

$y = 0$ – доброкачественный



Пороговое условие: $h_{\theta}(x) \geq 0.5 \rightarrow y = 1$

$h_{\theta}(x) < 0.5 \rightarrow y = 0$

$y \in \{0, 1\}$

$h_{\theta}(x) \in \mathbb{R}$

$h_{\theta}(x) < 0 ?$

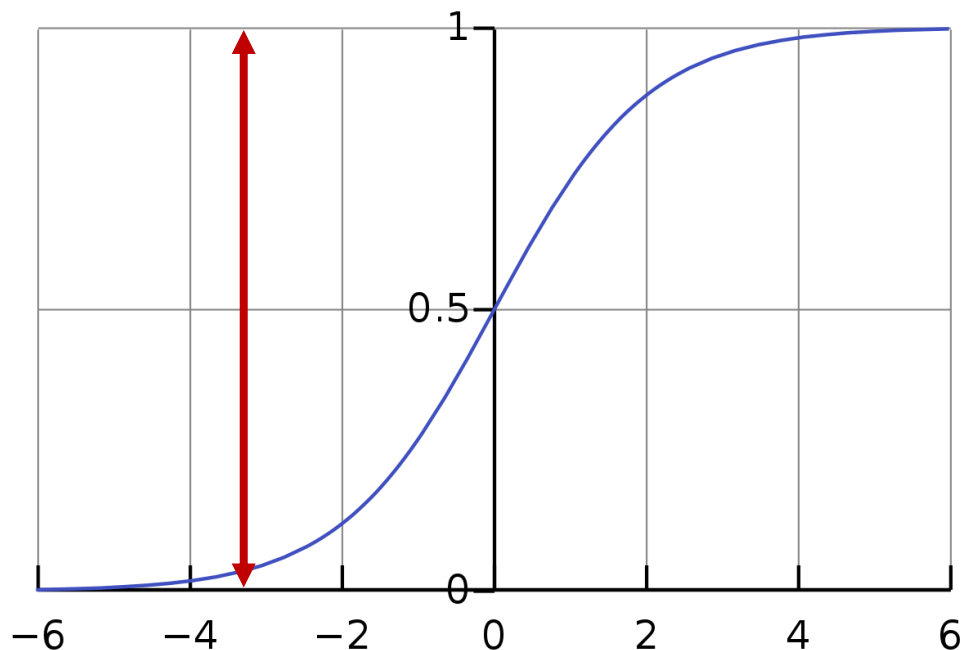
$h_{\theta}(x) > 1 ?$

$$0 \leq h_{\theta}(x) \leq 1$$

Как интерпретировать значения
 $h_{\theta}(x) > 0$ и $h_{\theta}(x) < 1$?

$$g(z) = \frac{1}{1 + e^{-z}}$$

сигмоида, сигмовидная функция,
логистическая функция



$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Вид гипотезы
логистической
регрессии

$h_{\theta}(x)$ – вероятность того, что $y = 1$ для заданного x

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{размер} \end{bmatrix} \quad \text{Например, } h_{\theta}(x) = 0.7$$

$h_{\theta}(x) = 0.7$ означает, что пациент (x) имеет шанс 70% наличия «злокачественной» ($y = 1$) болезни

$h_{\theta}(x) = P(y = 1|x; \theta)$ – вероятность того, что $y = 1$ для заданного x при параметрах θ

$$y \in \{0, 1\} \quad P(y = 1|x; \theta) + P(y = 0|x; \theta) = 1$$

Предположим, что имеется задача медицинской диагностики определения доброкачественной или злокачественной раковой опухоли. Для некоторого пациента, описанного вектором параметров x , логистическая регрессия выдала значение $h_{\theta}(x) = P(y = 1|x; \theta) = 0.7$, в связи с чем мы оцениваем шанс 70%, что опухоль является злокачественной.

Какова должна быть оценка, что опухоль является доброкачественной, то есть $P(y = 0|x; \theta)$?

☒ $P(y = 0|x; \theta) = 0.3$

☐ $P(y = 0|x; \theta) = 0.7$

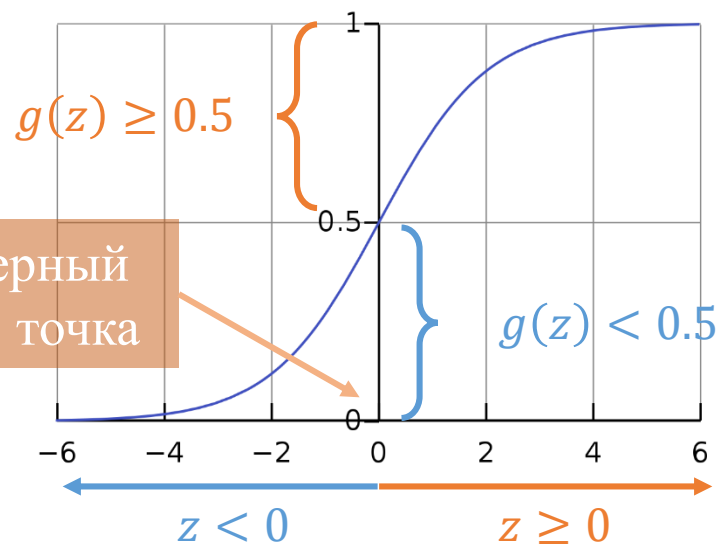
☐ $P(y = 0|x; \theta) = 0.7^2$

☐ $P(y = 0|x; \theta) = 0.3 \times 0.7$

$$h_{\theta}(x) = g(z) = P(y = 1|x; \theta)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

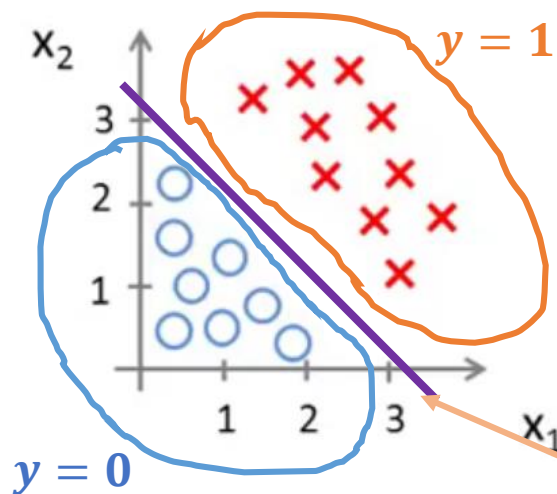
Одномерный
случай: точка



$$y = 1 : \quad h_{\theta}(x) \geq 0.5 \quad z \geq 0 \quad \theta^T x \geq 0$$

$$y = 0 : \quad h_{\theta}(x) < 0.5 \quad z < 0 \quad \theta^T x < 0$$

$$\left. \begin{array}{l} h_{\theta}(x) = 0.5 \\ \theta^T x = 0 \end{array} \right\} \text{Уравнение } \mathbf{границы решения} \text{ (разделяющей кривой, разделяющей плоскости, разделяющей гиперплоскости, } \textit{разделяющей гиперповерхности})$$



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

решение
(результат алгоритма
градиентного спуска)

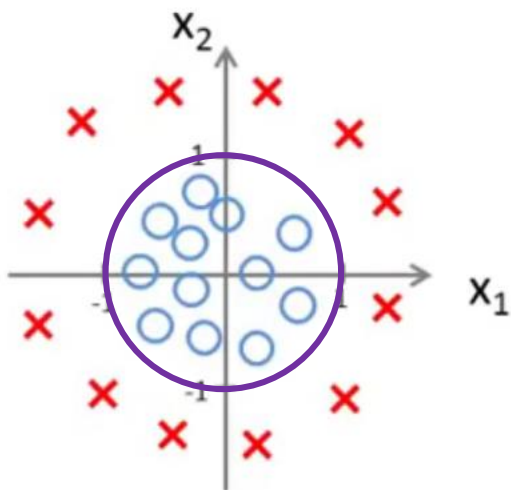
Двумерный
случай: линия

$$y = 1 : \quad -3 + x_1 + x_2 \geq 0$$

$$\theta^T x = 0$$

$$-3 + x_1 + x_2 = 0$$

$$x_1 + x_2 = 3$$



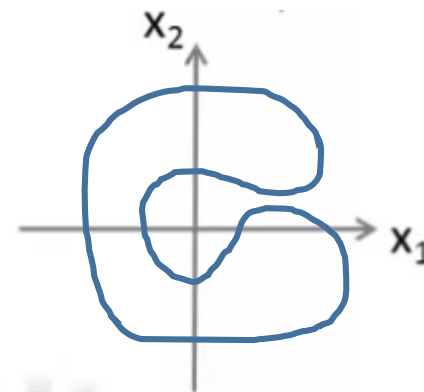
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

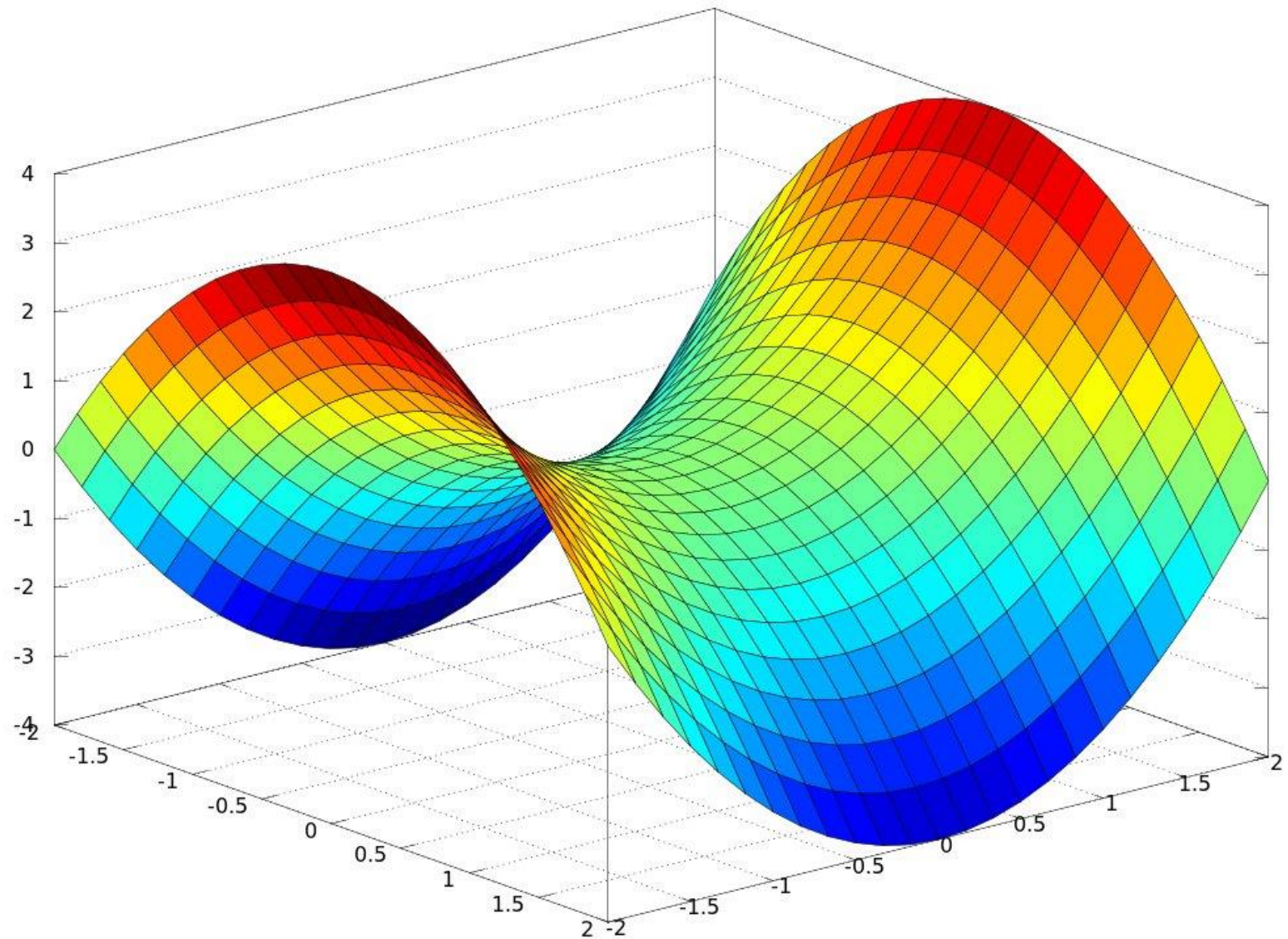
$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

$$y = 1: -1 + x_1^2 + x_2^2 \geq 0$$

$$x_1^2 + x_2^2 = 1$$

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \theta_6 x_1^2 x_2 + \dots)$$

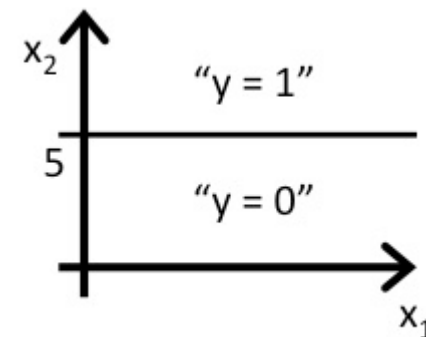
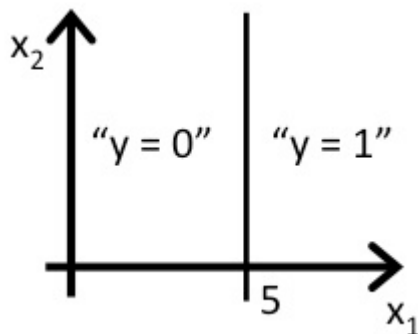
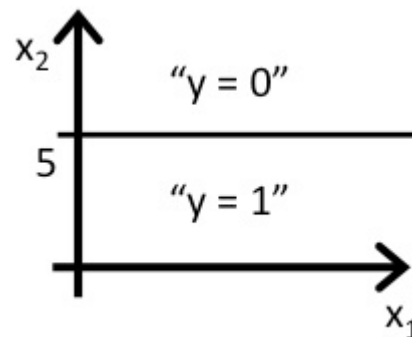
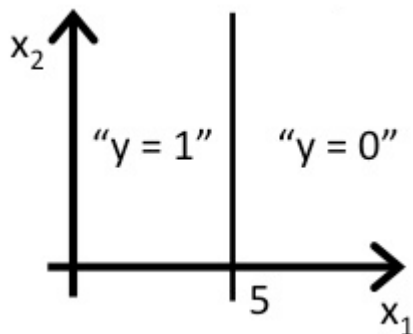




Поверхность в трехмерном пространстве

Рассмотрим логистическую регрессию с двумя параметрами x_1 и x_2 . Предположим, что $\theta_0 = 5$, $\theta_1 = -1$, $\theta_2 = 0$. Таким образом $h_\theta(x) = g(5 - x_1)$.

Что является границей решения для $h_\theta(x)$?



Обучающая выборка из m элементов:

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

Объекты:

$$x \in \mathbb{R}^{n+1} = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} \quad x_0 = 1$$

Ответы:

$$y \in \{0, 1\}$$

(два класса)

Гипотеза:

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Решение:

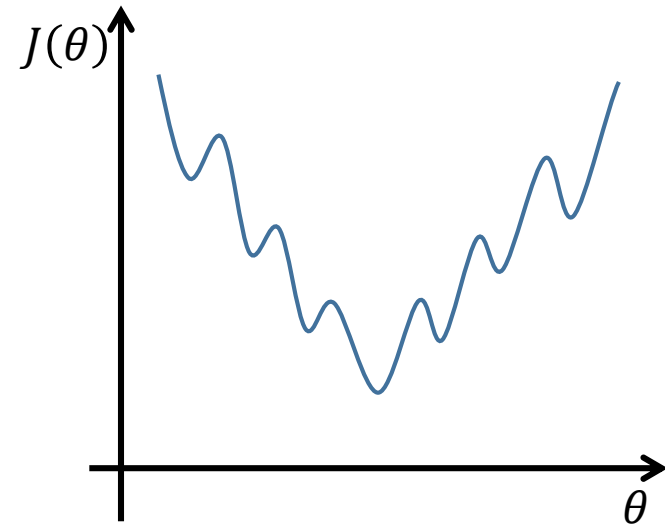
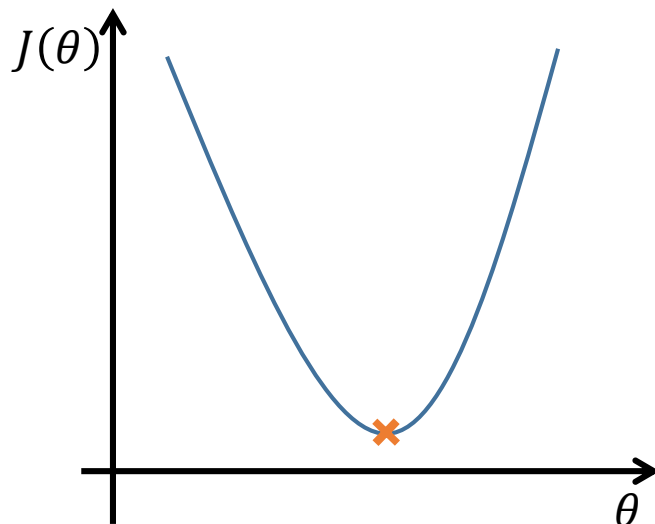
θ ?

(метод градиентного спуска)

Для линейной регрессии: $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

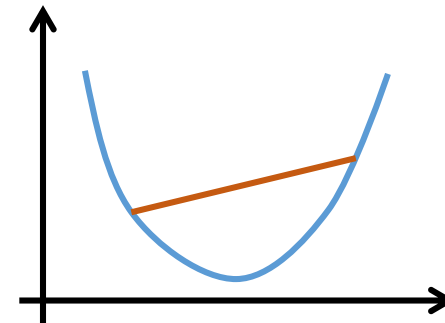
~~$$h_{\theta}(x) = \theta^T x$$~~

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

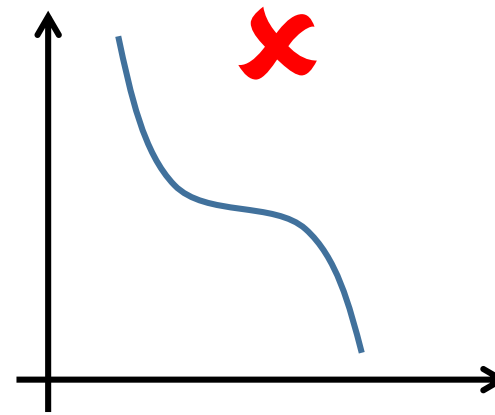
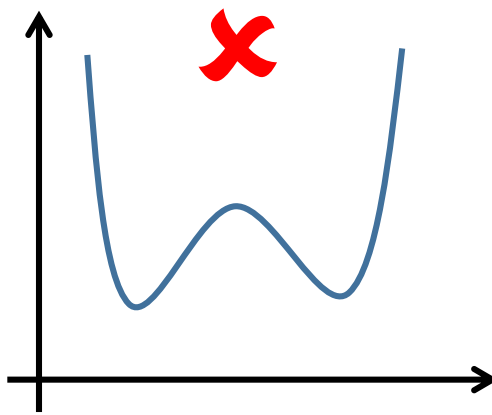
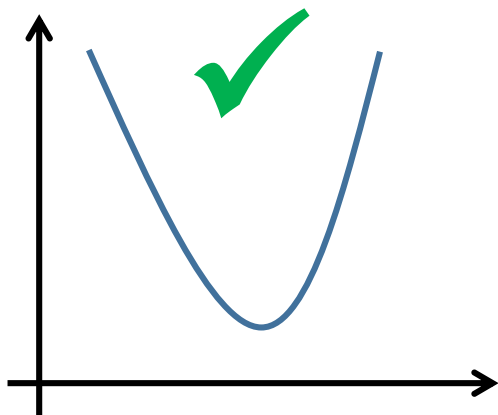


В чем проблема с этой функцией?

Функция выпуклая, когда отрезок между любыми двумя точками графика функции в векторном пространстве лежит не ниже соответствующей дуги графика.
Может быть выпуклой вверх или вниз.

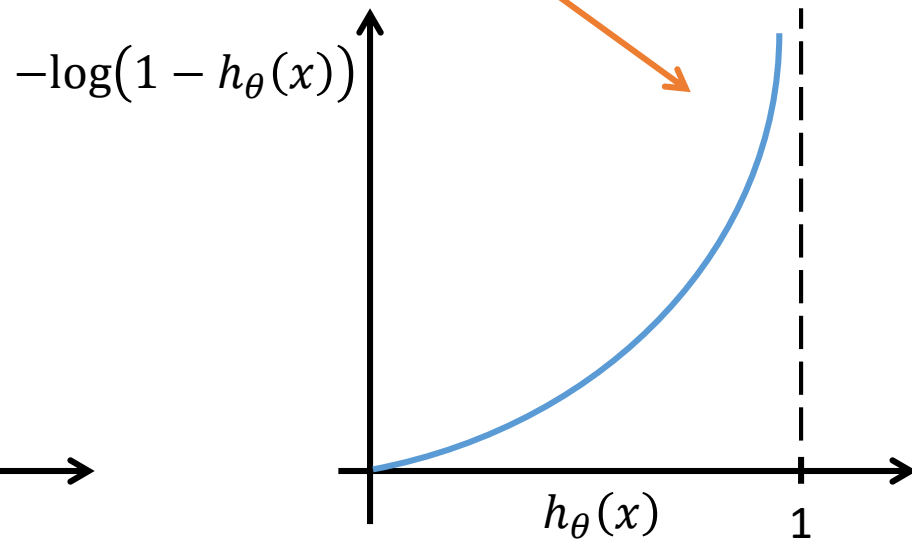
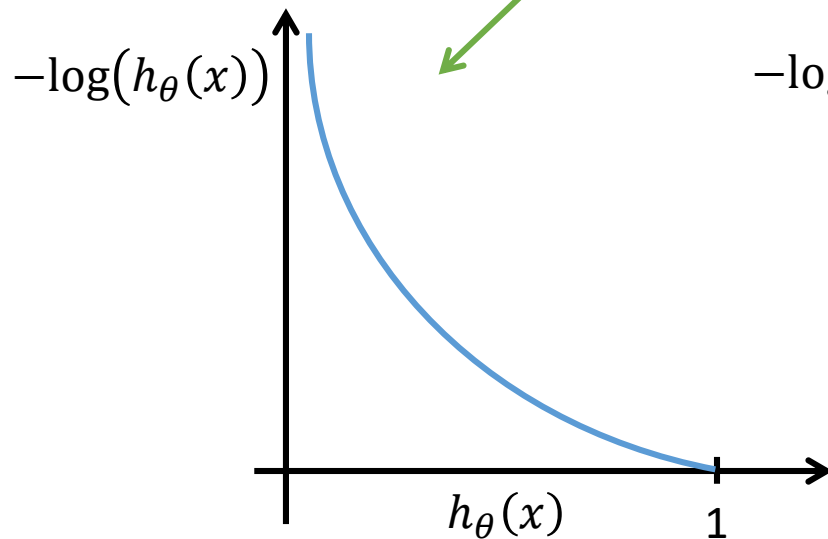


Какая из представленных ниже кривых выпуклая?



Для логистической регрессии: $J(\theta) = \frac{1}{m} \sum_{i=1}^m C(h_{\theta}(x^{(i)}), y^{(i)})$

$$C(h_{\theta}(x^{(i)}), y^{(i)}) = \begin{cases} -\log(h_{\theta}(x^{(i)})), & \text{если } y = 1 \\ -\log(1 - h_{\theta}(x^{(i)})), & \text{если } y = 0 \end{cases}$$

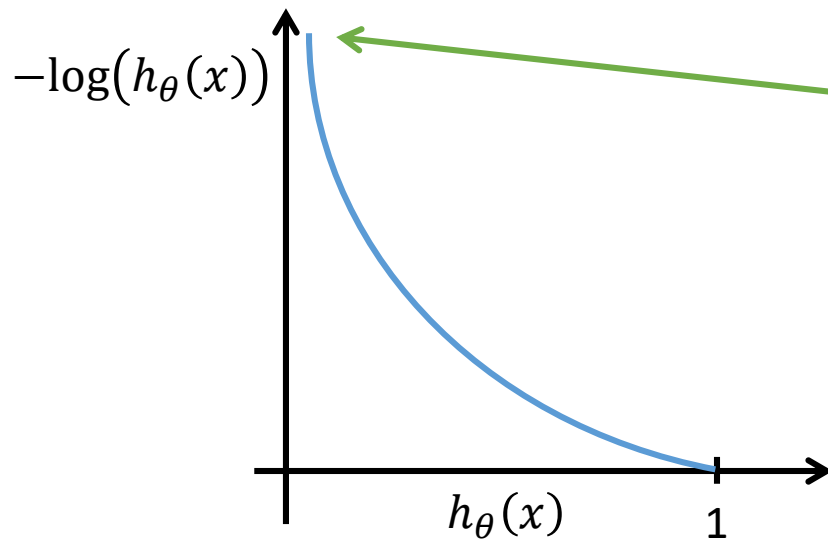


Задача медицинской диагностики рака:

$y^{(i)} = 1$ – злокачественная

$y^{(i)} = 0$ – доброкачественная

$$C(h_{\theta}(x^{(i)}), y^{(i)}) = \begin{cases} -\log(h_{\theta}(x^{(i)})), & \text{если } y = 1 \\ -\log(1 - h_{\theta}(x^{(i)})), & \text{если } y = 0 \end{cases}$$



Случай:

$$y^{(i)} = 1$$

$$h_{\theta}(x^{(i)}) = 0$$

Ошибка
предсказания!

Большой штраф за ошибку

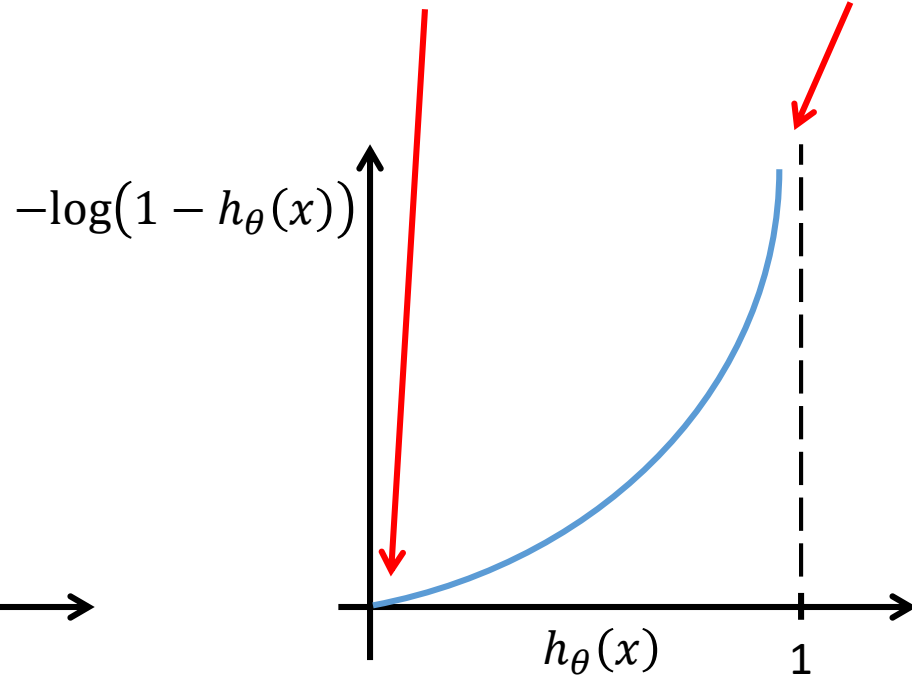
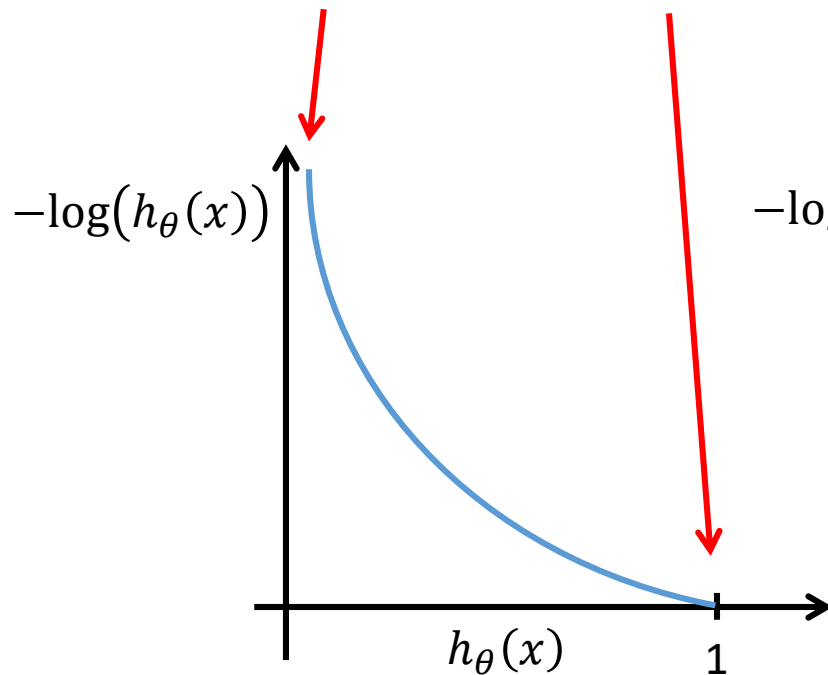
$$C(h_{\theta}(x^{(i)}), y^{(i)}) = \begin{cases} -\log(h_{\theta}(x^{(i)})), & \text{если } y = 1 \\ -\log(1 - h_{\theta}(x^{(i)})), & \text{если } y = 0 \end{cases}$$

$y^{(i)} = 1$
 $h_{\theta}(x^{(i)}) = 0$
 Большой штраф

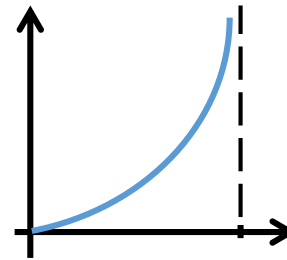
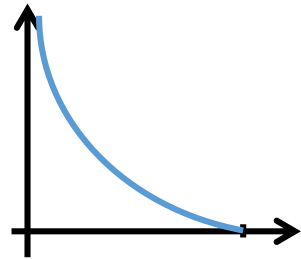
$y^{(i)} = 1$
 $h_{\theta}(x^{(i)}) = 1$
 Штраф = 0

$y^{(i)} = 0$
 $h_{\theta}(x^{(i)}) = 0$
 Штраф = 0

$y^{(i)} = 0$
 $h_{\theta}(x^{(i)}) = 1$
 Большой штраф



$$C(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)), & \text{если } y = 1 \\ -\log(1 - h_{\theta}(x)), & \text{если } y = 0 \end{cases}$$



Какие из следующих утверждений верны?

- ☒ Если $h_{\theta}(x) = y$, то $C(h_{\theta}(x), y) = 0$ (для $y = 0$ и $y = 1$)
- ☒ Если $y = 0$ и $h_{\theta}(x) \rightarrow 1$, то $C(h_{\theta}(x), y) \rightarrow \infty$
- ☒ Если $y = 0$ и $h_{\theta}(x) \rightarrow 0$, то $C(h_{\theta}(x), y) \rightarrow 0$
- ☒ Если $h_{\theta}(x) = 0.5$, то $C(h_{\theta}(x), y) > 0$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m C(h_{\theta}(x^{(i)}), y^{(i)})$$

$$C(h_{\theta}(x^{(i)}), y^{(i)}) = \begin{cases} -\log(h_{\theta}(x^{(i)})), & \text{если } y^{(i)} = 1 \\ -\log(1 - h_{\theta}(x^{(i)})), & \text{если } y^{(i)} = 0 \end{cases}$$

$$C(h_{\theta}(x^{(i)}), y^{(i)}) = -y^{(i)} \log(h_{\theta}(x^{(i)})) - \underbrace{(1 - y^{(i)})}_{y^{(i)} = 1} \log(1 - h_{\theta}(x^{(i)}))$$

0

$$y^{(i)} = 0$$

$$C(h_{\theta}(x^{(i)}), y^{(i)}) = \underbrace{-y^{(i)} \log(h_{\theta}(x^{(i)}))}_0 - (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m C(h_{\theta}(x^{(i)}), y^{(i)})$$

$$C(h_{\theta}(x^{(i)}), y^{(i)}) = -y^{(i)} \log(h_{\theta}(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))$$

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right]$$

Как найти θ , при которых $J(\theta)$ принимает минимальное значение?

Алгоритм градиентного спуска!

$$\frac{\partial}{\partial \theta_j} J(\theta) \quad j = 0, \dots, n \quad \theta \in \mathbb{R}^{n+1} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$$

Найти $n + 1$ частных производных: $\frac{\partial}{\partial \theta_j} J(\theta) \quad j = 0, \dots, n$

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right]$$

$$h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

$$\theta \in \mathbb{R}^{n+1} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$$

$$x \in \mathbb{R}^{n+1} = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix}$$

Частные
производные
по всем
параметрам
 $\theta_0, \dots, \theta_n$

$$\frac{\partial}{\partial \theta_0} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\frac{\partial}{\partial \theta_1} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_1^{(i)}$$

...

$$\frac{\partial}{\partial \theta_n} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_n^{(i)}$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}, \quad j = 0, \dots, n$$

Очень похоже на
линейную регрессию

Но!
$$h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

Шаг градиентного спуска:

$$\theta := \theta - \alpha \nabla J(\theta)$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{bmatrix}$$

$$\nabla J(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_0} J(\theta) \\ \frac{\partial}{\partial \theta_1} J(\theta) \\ \dots \\ \frac{\partial}{\partial \theta_n} J(\theta) \end{bmatrix}$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \quad j = 0, \dots, n$$

Какое из следующих выражений ему эквивалентно?

☒ $\theta := \theta - \alpha \frac{1}{m} \sum_{i=1}^m [(h_{\theta}(x^{(i)}) - y^{(i)}) x^{(i)}]$

☐ $\theta := \theta - \alpha \frac{1}{m} [\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})] x^{(i)}$

☐ $\theta := \theta - \alpha \frac{1}{m} x^{(i)} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$

☐ Все три верны

Задача сортировки электронных писем

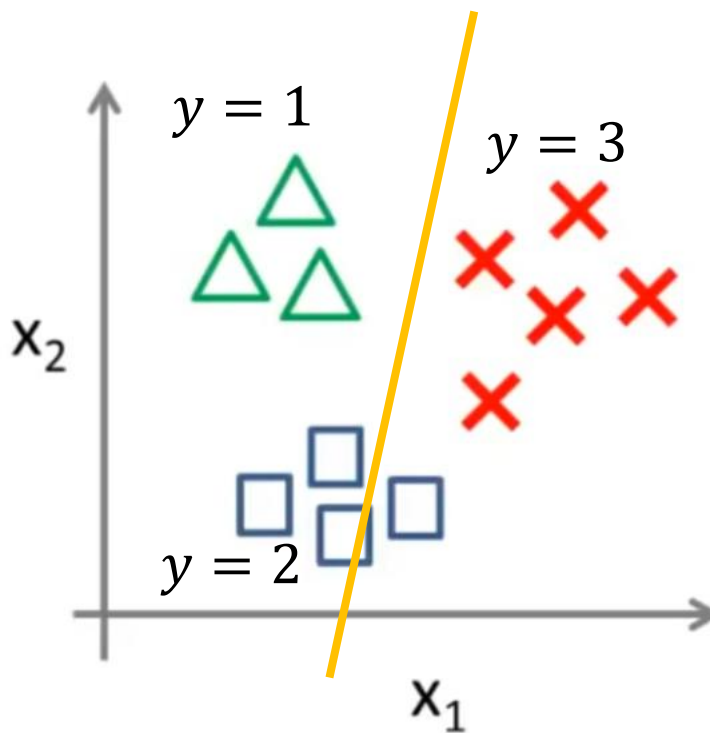
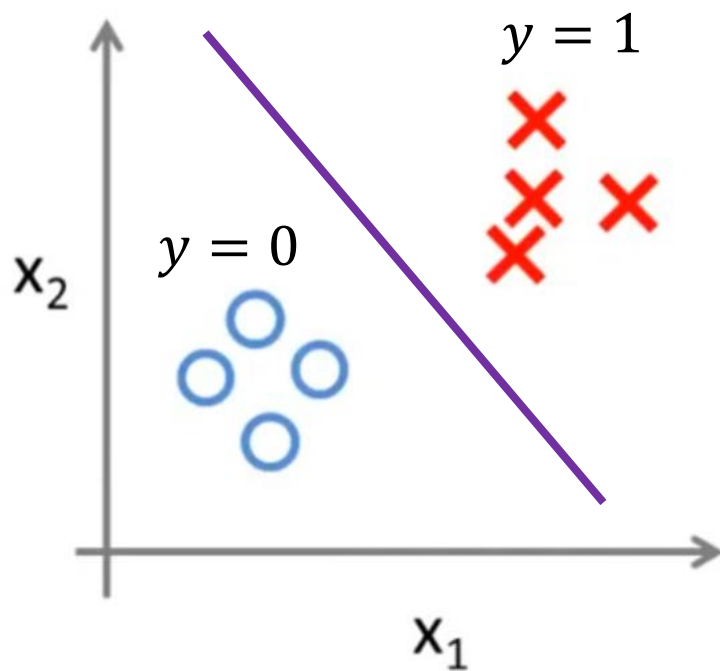
Дано t входящих писем, нужно присвоить им ярлыки:

Работа	Семья	Друзья	Развлечения	Реклама
$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$

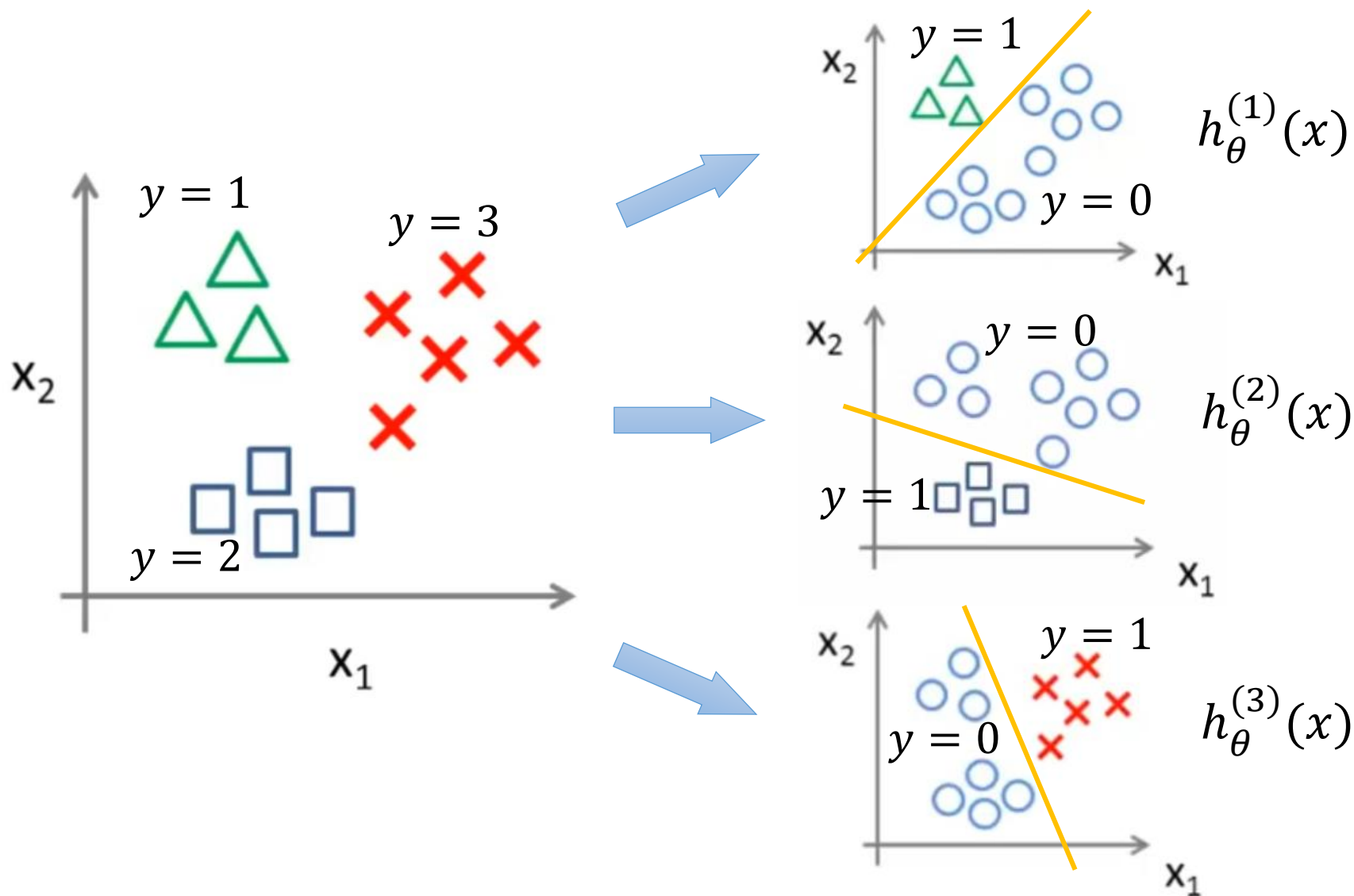
Задача прогноза облачности на завтра

Ясно	Переменная	Пасмурно
$y = 0$	$y = 1$	$y = 2$

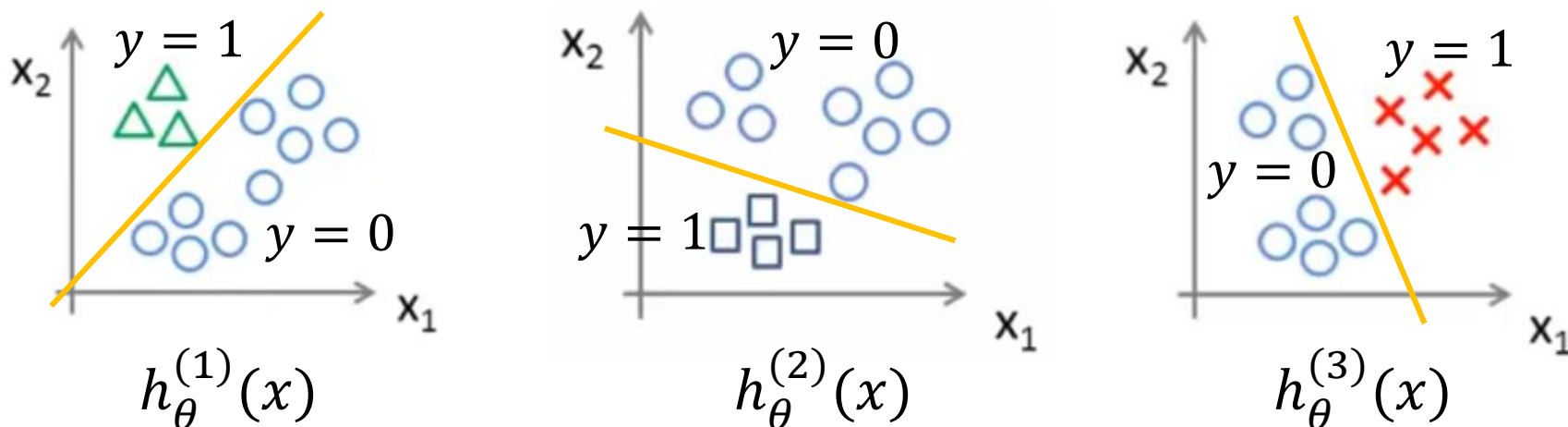
Граница решения???



Принцип «один-против-всеx»



Принцип «один-против-всех»



Дано x . Как определить, к какому классу он относится?

$h_{\theta}^{(k)}(x) = P(y = k|x; \theta)$ – вероятность того, что $y = k$ для заданного x при параметрах θ
 $0 \leq h_{\theta}^{(k)}(x) \leq 1$

$$H_{\theta}(x) = \underset{k}{\operatorname{argmax}} h_{\theta}^{(k)}(x)$$

Функция argmax возвращает k того $h_{\theta}^{(k)}(x)$, значение которого максимально

В задаче классификации на k классов (т.е. $y \in \{1, \dots, k\}$) сколько различных классификаторов методом логистической регрессии необходимо обучить?

☐ $k - 1$

☒ k

☐ $k + 1$

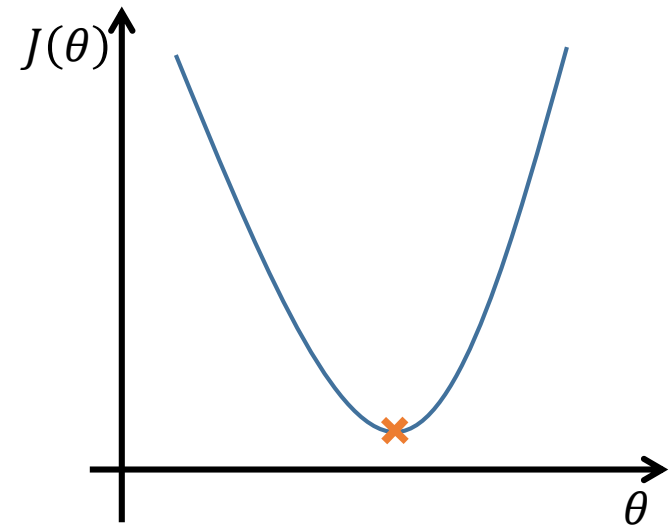
☐ k^2

☐ $\log_2 k$

Оптимизация – математическая задача нахождения минимума функции

Определить такие параметры θ , при которых $J(\theta)$ принимает наименьшее значение

Найти наилучшую гипотезу $h_{\theta}(x)$



Градиентный спуск – простой алгоритм оптимизации

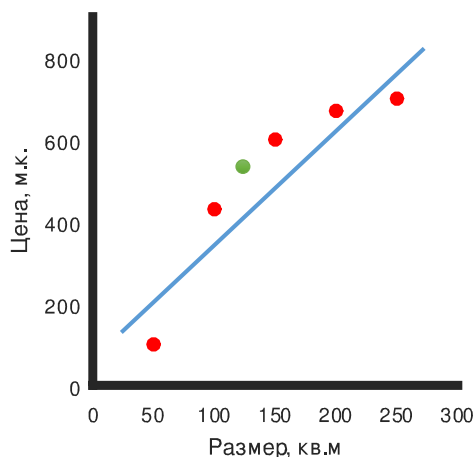
- Легко реализуется
- Понятен
- Позволяет реализацию методами линейной алгебры
- Требуется настройки (α, ε)
- Может долго сходиться
- Не находит глобальный минимум

Методы первого порядка (используют первую производную, градиент)

- Метод сопряженных градиентов Флетчера – Ривса
- Стохастический градиентный спуск (SGD)
- Адаптивный стохастический градиентный спуск (Adagrad, Adam, RMSProp)

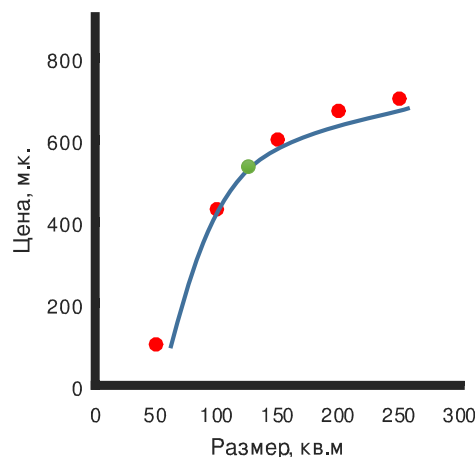
Методы второго порядка (используют также вторую производную, гессиан)

- Алгоритм Бroyдена – Флетчера – Гольдфарба – Шанно (BFGS, L-BFGS)
- Метод Ньютона (его тензорные варианты – Shampoo)
- Алгоритм адаптивной оценки Гессиана (AdaHessian)

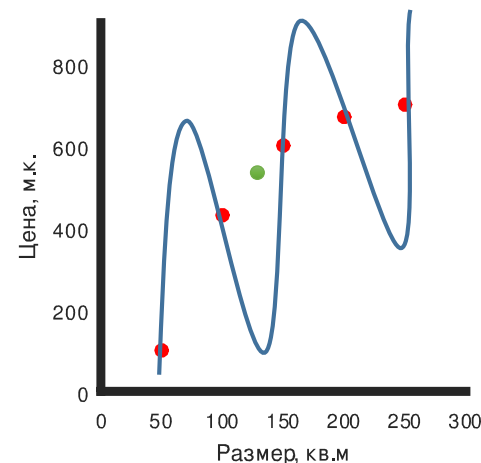


$$h_{\theta}^{(1)}(x) = \theta_0 + \theta_1 x$$

**Недообученная
модель**



$$h_{\theta}^{(2)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



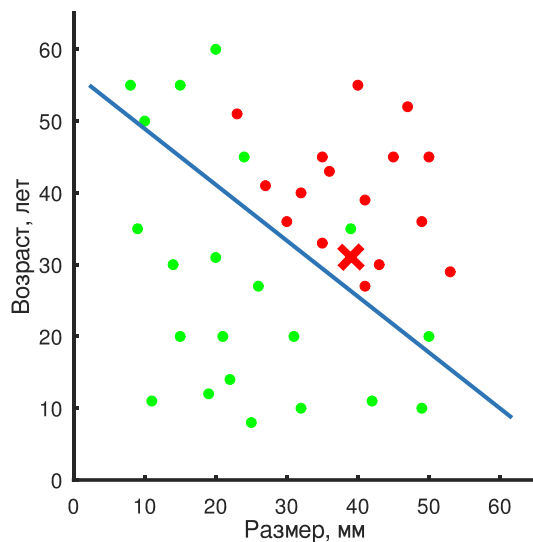
$$h_{\theta}^{(3)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_4 x^4$$

**Переобученная
модель**

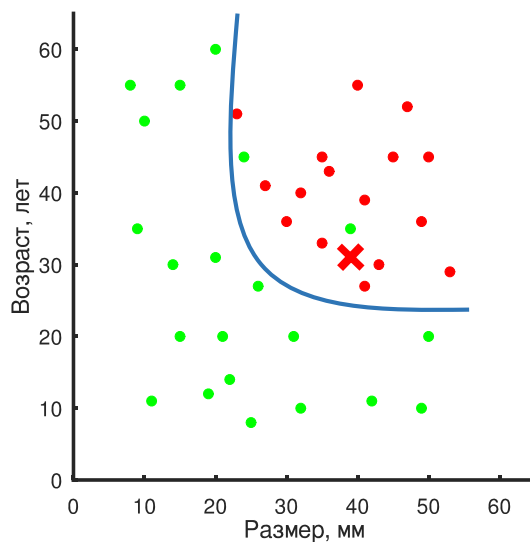
$$J^{(1)}(\theta) > J^{(2)}(\theta) > J^{(3)}(\theta)$$

Модель имеет хорошую **обобщающую способность**, если хорошо работает и на обучающем наборе и на новых данных

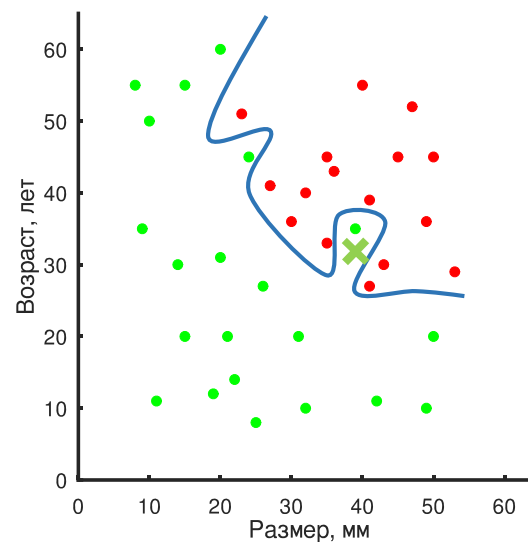
Хорошо на обучающем наборе и плохо на новых данных — плохая обобщающая способность.



$$h_{\theta}^{(1)}(x) = \theta_0 + \theta_1 x$$



$$h_{\theta}^{(2)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



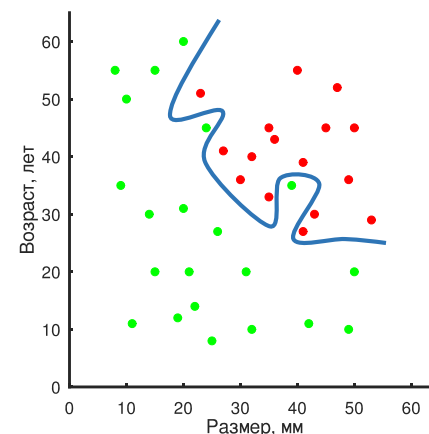
$$h_{\theta}^{(3)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_4 x^4$$

$$J^{(1)}(\theta) > J^{(2)}(\theta) > J^{(3)}(\theta)$$

Недообученная
модель

Переобученная
модель

Рассмотрим задачу медицинской диагностики (задача классификации). Если гипотеза $h_{\theta}(x)$ переобучается на наборе данных, то это означает, что ...
(выбрать один правильный ответ)



- ☒ гипотеза делает точные предсказания на обучающем наборе данных и хорошо обобщается, делает точные предсказания для новых, ранее не встречаемых примерах
- ☒ гипотеза делает плохие предсказания на обучающем наборе данных, но хорошо обобщается
- ☒ гипотеза делает точные предсказания на обучающем наборе данных, но плохо обобщается
- ☒ гипотеза делает плохие предсказания на обучающем наборе данных и плохо обобщается

Причины переобучения

- Ошибки конструирования модели (слишком сложная функция)
- Много исходных параметров ($n \gg m$)

x_1 — площадь дома

x_2 — количество комнат

x_3 — количество этажей

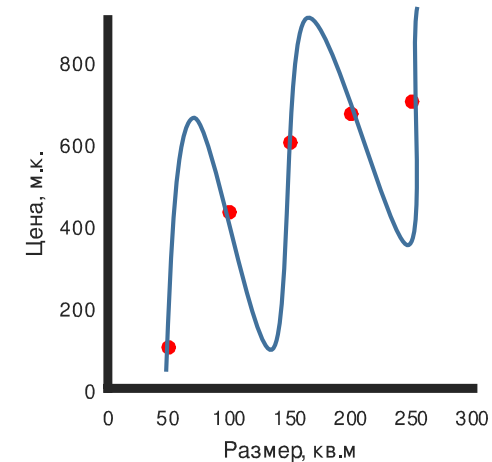
x_4 — возраст дома

...





x_{98} — расстояние до магазина

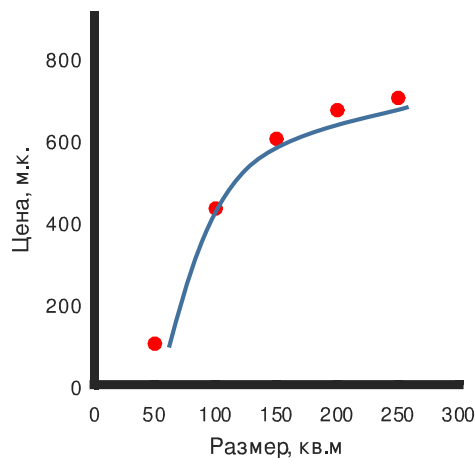
x_{99} — средний доход жителей по соседству

...

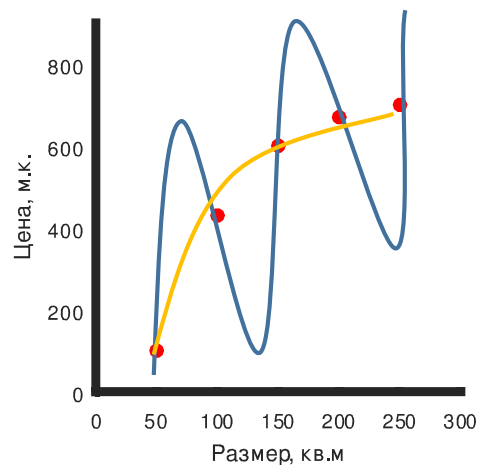


$$h_{\theta}^{(3)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_4 x^4$$

1. Использовать внешние критерии оценки $J(\theta)$  Будем изучать позже
2. Уменьшить размер вектора θ : 
 - вручную выбрать признаки, которые оставить в модели
 - использовать методы отбора признаков  Будем изучать позже
3. Использовать регуляризацию
 - сохранить все признаки в исходных данных, но уменьшить для некоторых из них степень влияния на результат
 - хорошо работает, когда имеется множество признаков, все из которых имеют некоторое влияние на результат  Рассмотрим подробнее



$$h_{\theta}^{(2)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



$$h_{\theta}^{(3)}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4}$$

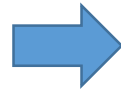
Добавим штраф на параметры θ_3 и θ_4 :

$$J(\theta) = \left[\frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \underline{1000 \cdot \theta_3^2} + \underline{1000 \cdot \theta_4^2} \right] \rightarrow \min$$

$$\theta_3 \approx 0$$

$$\theta_4 \approx 0$$

Небольшие значения
некоторых $\theta_0, \theta_1, \dots, \theta_n$



- Упрощают гипотезу
- Делают ее менее склонной к переобучению

Регуляризованная функция стоимости:

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Параметр
регуляризации

регуляризатор

$j = 1, \dots, n$

θ_0 не участвует в
регуляризации

θ_0 можно включать или не включать в
регуляризацию, результат почти не
меняется. Не будем включать θ_0

Решая задачу регуляризованной линейной регрессии градиентным спуском, мы должны получить значения θ такие, чтобы найти минимум функции стоимости:

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

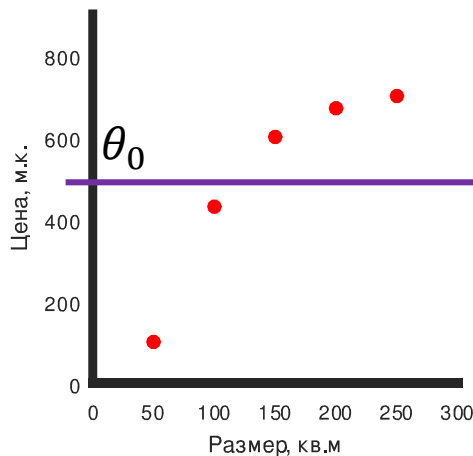
Что если задать значение λ очень большим (например, $\lambda = 10^{10}$) ?

- ☒ Алгоритм обрабатывает правильно, большое значение λ не повредит ему
- ☒ Устранить переобучение не получится, но на обучающем наборе данных результат будет достаточно точным
- ☒ Получим недообученную модель
- ☒ Градиентный спуск не сможет завершить работу

Решая задачу регуляризованной линейной регрессии градиентным спуском, мы должны получить значения θ такие, чтобы найти минимум функции стоимости:

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Что если задать значение λ очень большим (например, $\lambda = 10^{10}$) ?



$$h_{\theta}(x) = \theta_0 + \cancel{\theta_1 x} + \cancel{\theta_2 x^2} + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4}$$

\downarrow \downarrow \downarrow \downarrow
 0 0 0 0

$$h_{\theta}(x) = \theta_0$$

**Недообученная
модель!**

Выбор значения λ должен быть обоснован
(рассматривается в дальнейших лекциях)

Функция стоимости:
$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

Частные производные:

$$\frac{\partial}{\partial \theta_0} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \left[\frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \right] + \frac{\lambda}{m} \theta_j$$

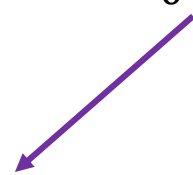
$$j = 1, \dots, n$$

Цикл шага алгоритма градиентного спуска:

повторять {
 для $j = 0, \dots, n$:
 $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \dots, \theta_n)$
} пока $d\theta_j > \varepsilon, j = 0, \dots, n$

θ_0 не регуляризуется

повторять {
 $\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})$
 для $j = 1, \dots, n$:
 $\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m}\right) - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$
} пока $d\theta_j > \varepsilon, j = 0, \dots, n$



Решаем задачу регуляризованной линейной регрессии градиентным спуском на обучающем наборе размером $m > 0$, используя некоторый небольшой шаг $\alpha > 0$ и некоторый заданный параметр регуляризации $\lambda > 0$. Для вычисления следующего значения $\theta_1, \dots, \theta_n$ используется выражение:

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m} \right) - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Какое из следующих утверждений является верным?

☒ $1 - \alpha \frac{\lambda}{m} > 1$

☒ $1 - \alpha \frac{\lambda}{m} = 1$

☒ $1 - \alpha \frac{\lambda}{m} < 1$

☒ Никакое из перечисленных

Регуляризация нормального уравнения

$$X = \begin{bmatrix} x_0^{(1)} & \dots & x_n^{(1)} \\ \dots & \dots & \dots \\ x_0^{(m)} & \dots & x_n^{(m)} \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}$$

$$Y = \begin{bmatrix} y^{(1)} \\ \dots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^m$$

$\min_{\theta} J(\theta) ?$

$$\theta = (X^T X)^{-1} X^T y$$

не регуляризованная форма
нормального уравнения

Регуляризованная форма нормального уравнения

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} \right)^{-1} X^T y$$

Если размер обучающей выборки мал: $m \ll n$

$$\theta = \underline{(X^T X)^{-1} X^T y}$$

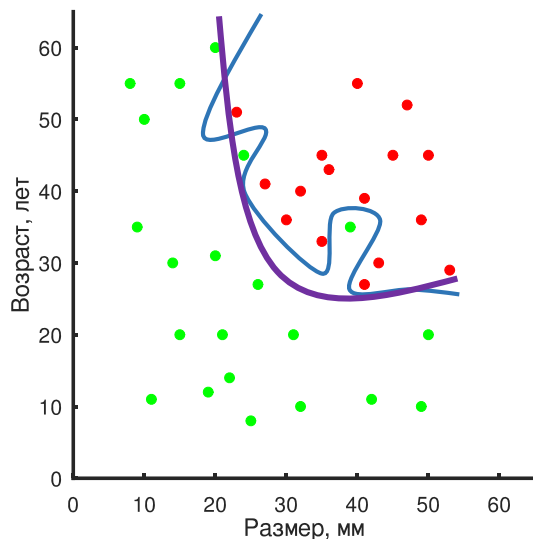
не обратимая
(вырожденная) матрица

$$\theta = \left(\underline{X^T X + \lambda \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}} \right)^{-1} X^T y$$

обратимая матрица
при условии $\lambda > 0$

Регуляризация позволяет получить решение нормального уравнения во многих случаях

При больших значениях n нахождение обратной матрицы по прежнему очень долгий процесс




$$h_{\theta}(x) = g(\theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_4 x^4)$$

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right] + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$$

Не дает параметрам $\theta_1, \dots, \theta_n$ расти, уменьшая склонность модели к переобучению
(θ_0 у нас не участвует в регуляризации)

Цикл шага алгоритма градиентного спуска:

повторять {
 $\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$ 
 для $j = 1, \dots, n$:
 $\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m}\right) - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$
} пока $d\theta_j > \varepsilon, j = 0, \dots, n$

θ_0 не регуляризуется

Точно также, как и для линейной регрессии, но ...

~~$$h_{\theta}(x^{(i)}) = \theta^T x^{(i)}$$~~

$$h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

Используя логистическую регрессию, какой способ контролировать, что градиентный спуск работает корректно, является правильным из представленных?

- ☒ Отобразить график $-\frac{1}{m} [\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))]$ от номера итерации и убедиться в том, что функция уменьшается
- ☒ Отобразить график $-\frac{1}{m} [\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$ от номера итерации и убедиться в том, что функция уменьшается
- ☒ Отобразить график $-\frac{1}{m} [\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))]$ $-\frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$ от номера итерации и убедиться в том, что функция уменьшается
- ☒ Отобразить график значения $\sum_{j=1}^n \theta_j^2$ от номера итерации и убедиться в том, что функция уменьшается