Линейна регресия с една променлива

Доц. д-р Ивайло Пенев

Кат. „Компютърни науки и технологии“

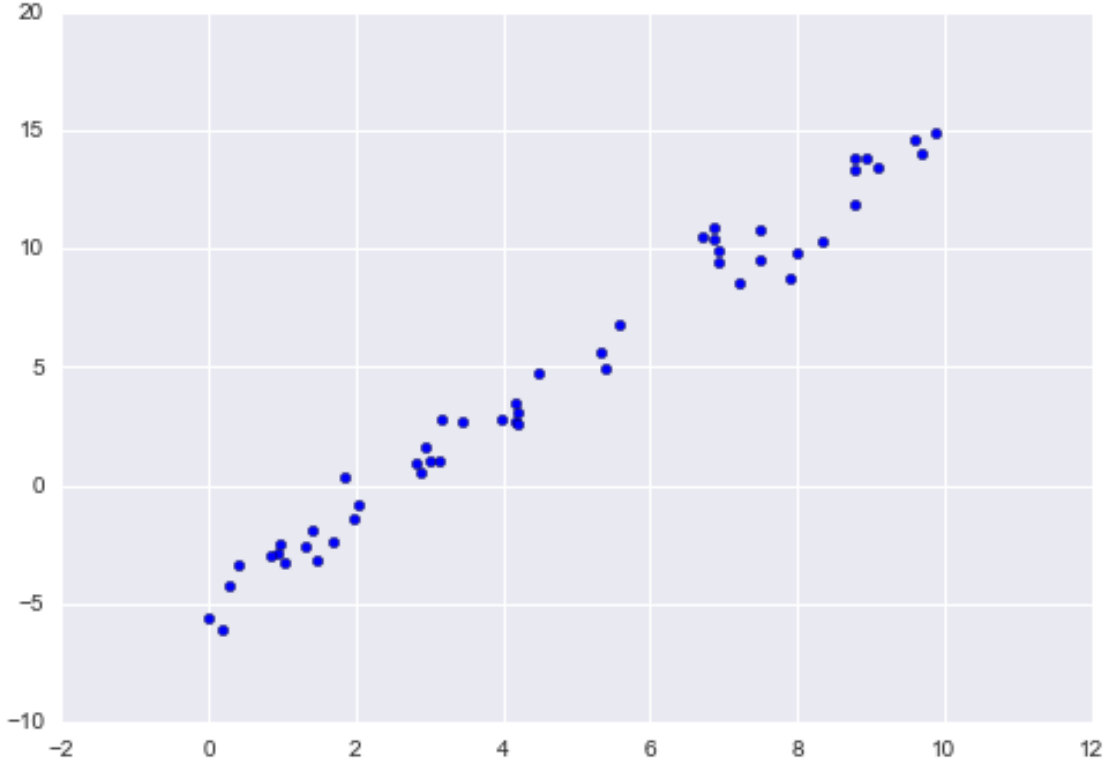
Пример 1

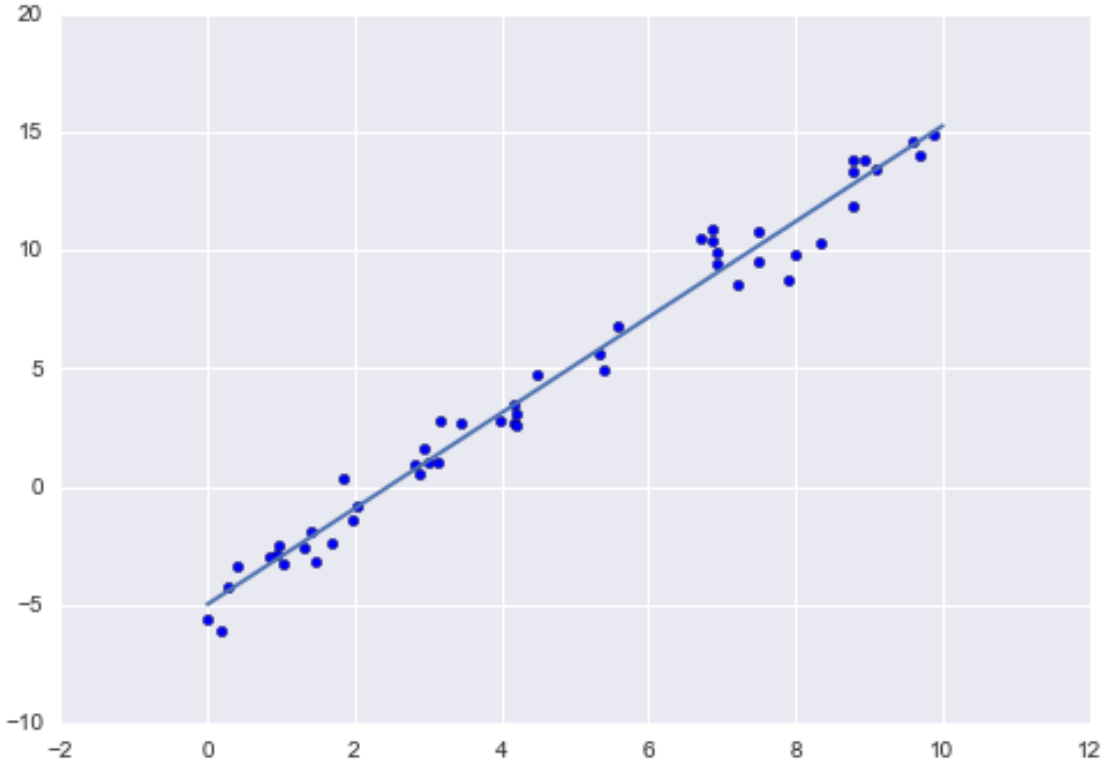
- Данни за цени на къщи

Площ (x)	Цена в $ (y)
2104	460 000
1416	232 000
1534	315 000
852	178 000
…	…

- m – брой обучителни данни (training examples)
  x – входна променлива (“input” variable/feature)
  y – изходна стойност (“output” variable/target)
  (x,y) – един обучителен пример (training example)
  $x^{(i)}$, $y^{(i)}$ – i-ти обучителен пример

Пример 2

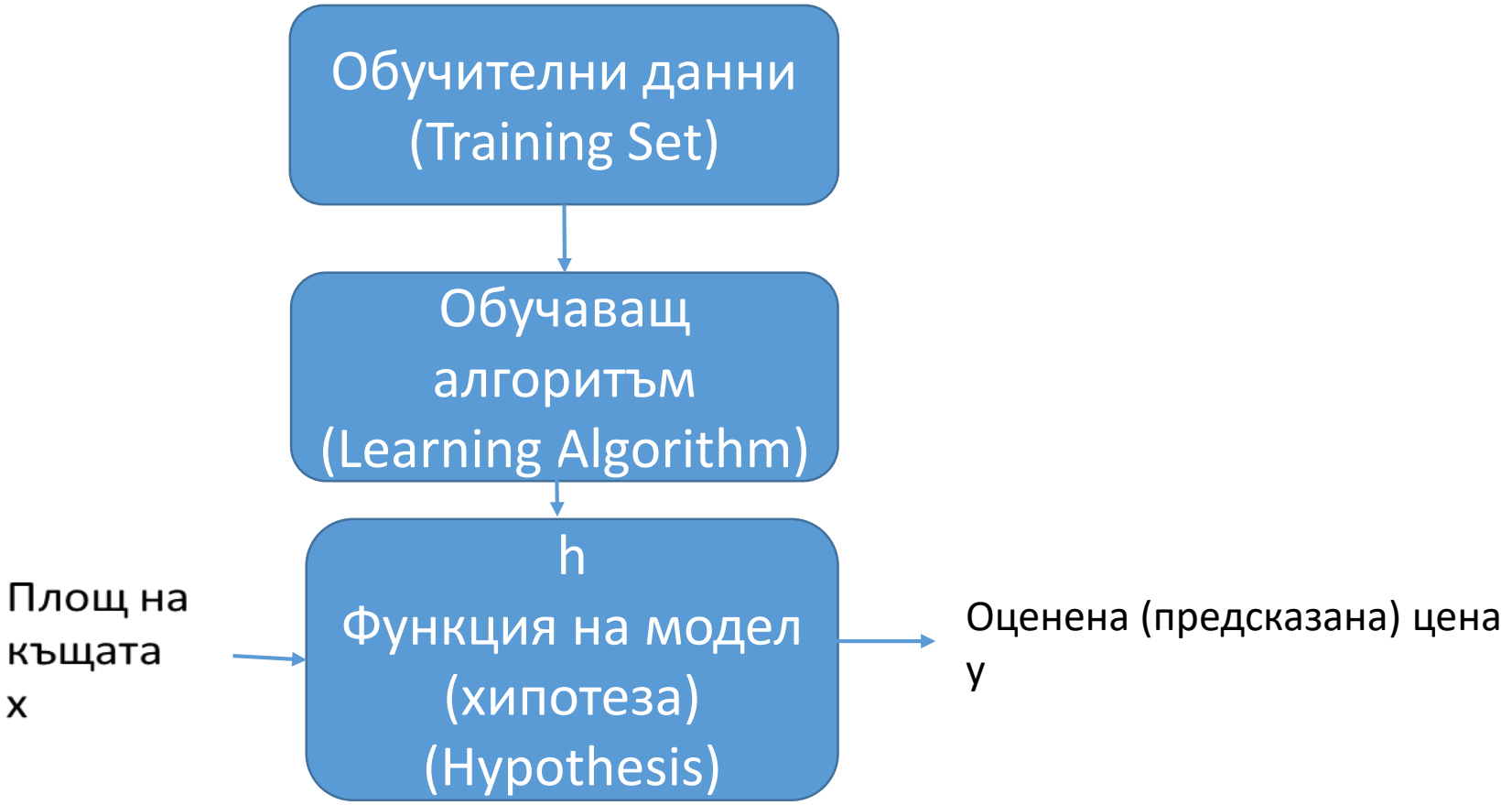




Регресионна задача

- В регресионната задача имаме входни променливи и изходни стойности (резултати), които отнасяме към **непрекъсната функция на резултатите.**
- При линейна регресия с една променлива (univariate linear regression) предсказваме **единствена изходна стойност y** за **единствена входна стойност x**.
- Линейната регресия е пример за надзиравано машинно обучение (supervised machine learning) – известно е как входните данни влияят върху резултатите

Модел на регресионна задача



Функция – модел - хипотеза (Hypothesis function)

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$$

- $\hat{y}$ - прогнозна изходна стойност (прогнозен резултат)
- x – входна стойност (параметър на функцията – параметър на модела)
- h_θ - функция на модела (хипотеза)
- θ_0, θ_1 - параметри (коефициенти)
- При линейната регресия моделът (хипотезата) представлява уравнение на права линия

Пример

- Дадени са обучаващи данни:

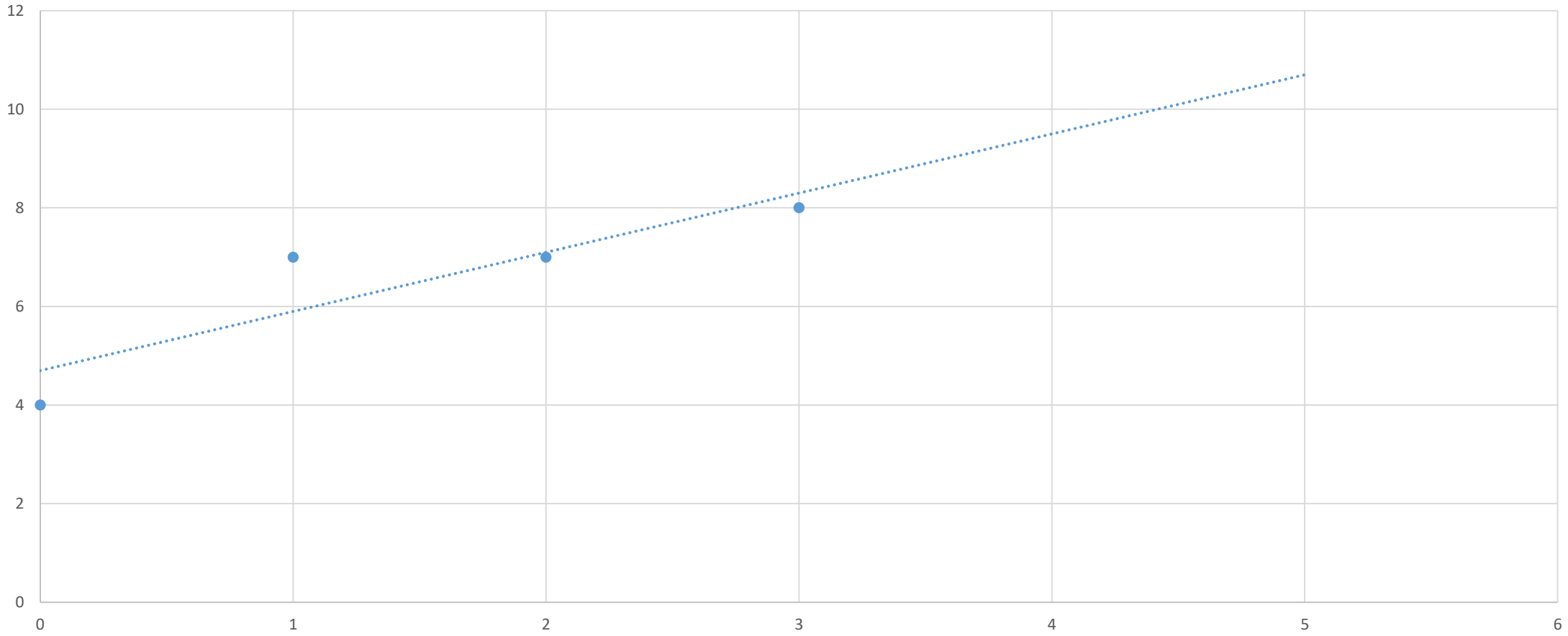
Входни данни x	Резултат y
0	4
1	7
2	7
3	8

- Правим предположение за параметрите на функцията, напр.

h_θ - $\theta_0 = 2, \theta_1 = 2$

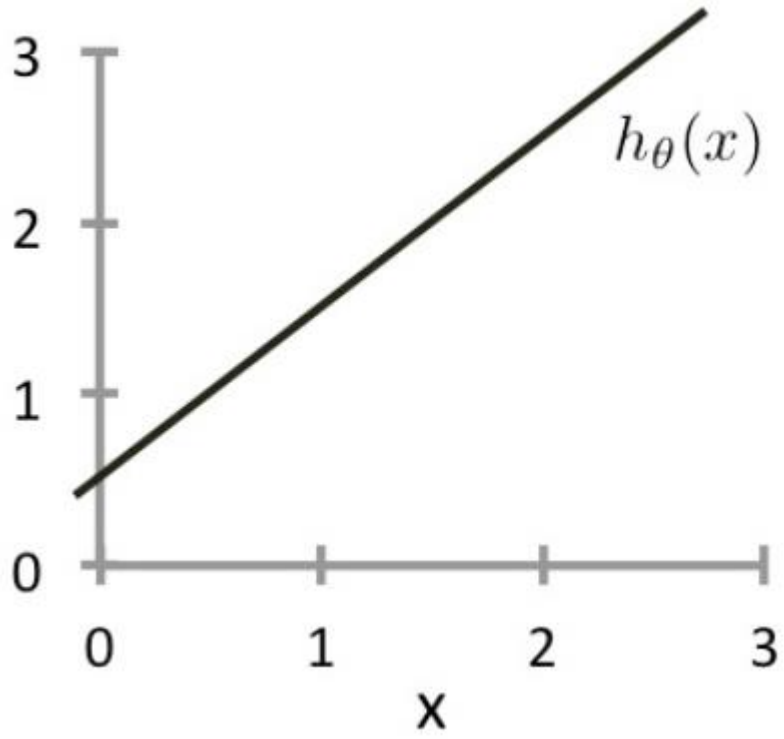
- Функцията придобива вида $h_\theta(x) = 2 + 2x$
- При x=1, y=4, т.е. разликата с обучаващите данни за x=1 е 3
- Търсим стойности на θ_0, θ_1, за които разликата $h_\theta - y$ е минимална

Графично представяне на функцията



Контролен въпрос

Дадена е следната графика за хипотеза $h_\theta(x) = \theta_0 + \theta_1 x$.
Определете стойностите на θ_0 и θ_1.



а) $\theta_0 = 0, \theta_1 = 1$

б) $\theta_0 = 0.5, \theta_1 = 1$

в) $\theta_0 = 1, \theta_1 = 0.5$

г) $\theta_0 = 1, \theta_1 = 1$

Ценова функция (Cost function)

- Използва се за измерване на точността на функцията на хипотезата

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x) - y_i)^2$$

- $J(\theta_0, \theta_1)$ - ценова ф-я с параметри θ_0, θ_1
- m – брой данни за обучение на ценовата ф-я
- $\hat{y}_i$ - предсказана стойност за резултата
- y_i - действителна стойност за резултата от данните за обучение
- Тази ф-я се нар. средна квадратична грешка (Mean Squared Error)

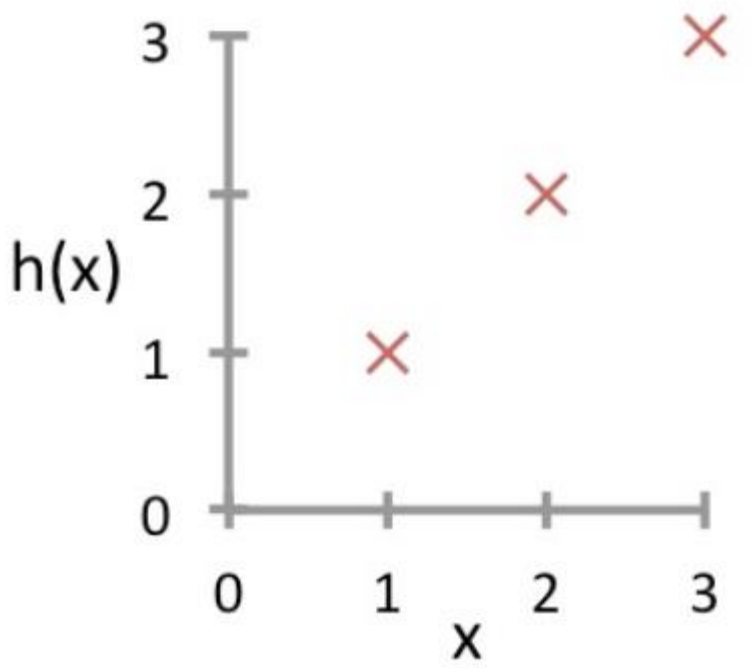
Примери за хипотеза и цена при $\theta_0 = 0$

- Дадени са 3 броя обучителни данни (m=3), изобразени на фигурата. Да се изчислят стойностите на цената:

а) При $\theta_1 = 1$ $J(\theta_1) = ?$

б) При $\theta_1 = 0.5$ $J(\theta_1) = ?$

в) При $\theta_1 = 0$ $J(\theta_1) = ?$



Цел на обучението

- Целта на линейната регресия е да намерим такива стойности на коефициентите θ_0, θ_1, за които ценовата ф-я $J(\theta_0, \theta_1)$ има минимална стойност. При такива стойности:
  - Разликите $\hat{y}_i - y_i$ са минимални
  - Правата линия минава най-близо до всички точки от данните за обучение
- В идеалния случай правата преминава през всички точки, т.е. $\hat{y}_i - y_i = 0$ за $\forall i$, т.е. $J(\theta_0, \theta_1) = 0$

Обобщение на линейна регресионна задача

- Функция на хипотезата (Hypothesis)

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- Параметри (Parameters)

$$\theta_0, \theta_1$$

- Функция на цената (Cost Function)

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x) - y_i)^2$$

- Цел

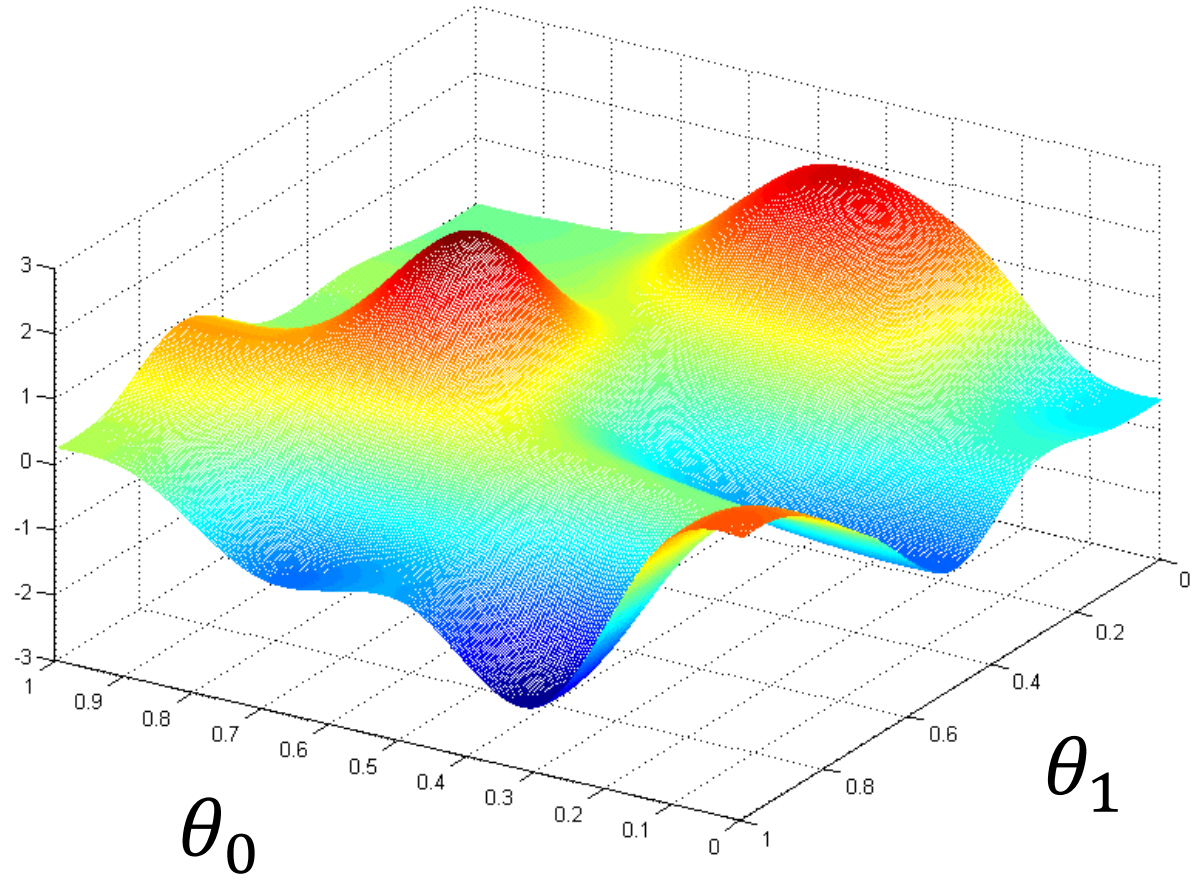
$$\underset{\theta_0, \theta_1}{minimize}\, J(\theta_0, \theta_1)$$

Обща идея на градиентното спускане

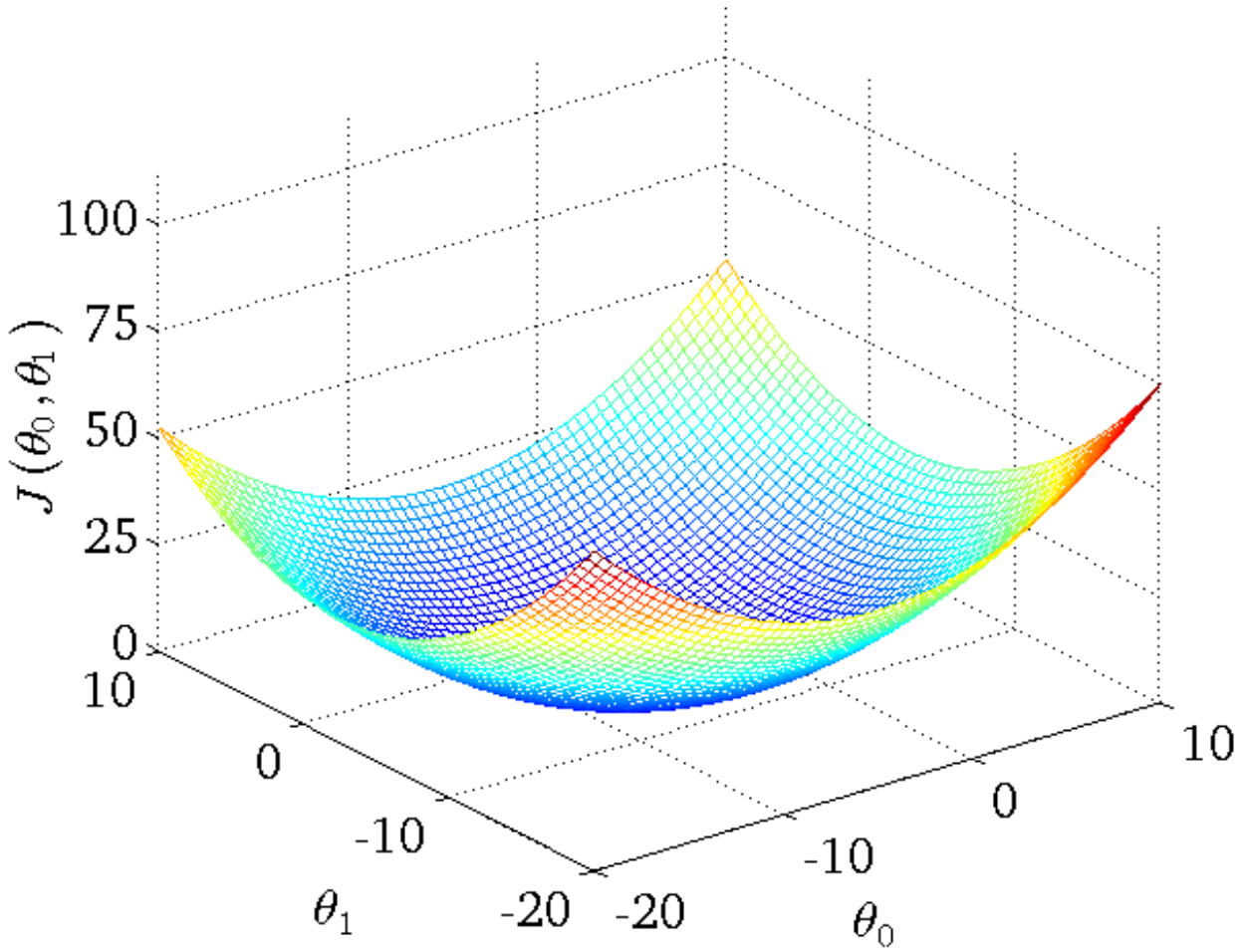
- Градиентното спускане (gradient descent) е алгоритъм за намиране минимум на функция (не само в линейната регресия)
- Имаме функция на цената $J(\theta_0, \theta_1)$
- Търсим минимум (минимална стойност) $min\ J(\theta_0, \theta_1)$
- Обща идея
  - Започваме с някакви начални стойности на параметрите θ_0, θ_1 (напр. $\theta_0 = 0, \theta_1 = 0$)
  - Повтаряме (итеративно) θ_0, θ_1 до достигане на $min\ J(\theta_0, \theta_1)$

Графично представяне на функцията на цената

$J(\theta_0, \theta_1)$



Графично представяне на функцията на цената



Математическа формулировка на градиентното спускане

- Целта на алгоритъма е чрез постепенна промяна на коефициентите θ_0, θ_1 (с много малки „стъпки“) да се намери минимум на ценовата функция $J(\theta_0, \theta_1)$
- Градиентното спускане има следният общ вид:

Повтаряй до достигане на минимум на ф- та $J(\theta_0, \theta_1)$

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ *за j=0 и j=1*

, където

$j = 0,1$- индекс на коефициента

α – скорост на обучението (learning rate)

Градиентно спускане при линейна регресия

- За линейна регресия градиентното спускане има следният вид:

Повтаряй до достигане на минимум на ф-та $J(\theta_0, \theta_1)$: {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} ((h_\theta(x_i) - y_i) x_i)$$

}

, където

m – брой данни за обучение на функцията

θ_0, θ_1 - коефициенти на функцията, които се променят едновременно

x_i, y_i - входни данни и изходни стойности от данните за обучение