Learning Systems (DT8008)

Basics and Prerequisites

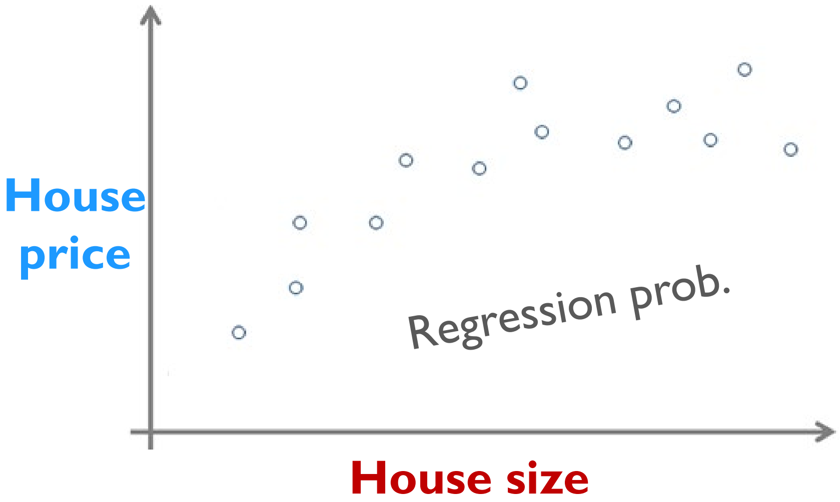
Terminology, definitions and review of some notions

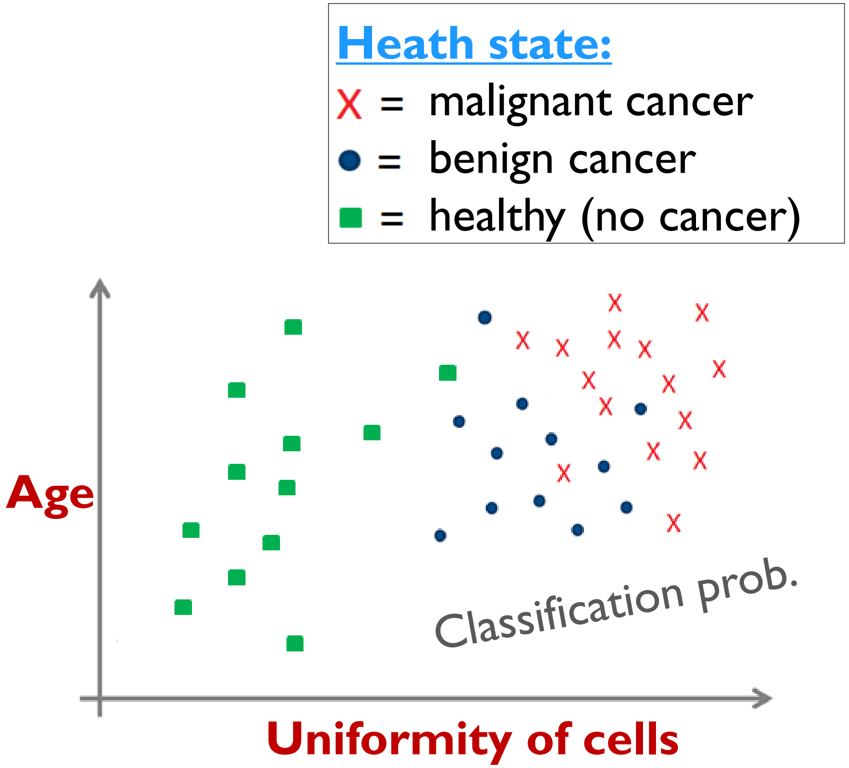
Dr. Mohamed-Rafik Bouguelia
mohbou@hh.se

Halmstad University

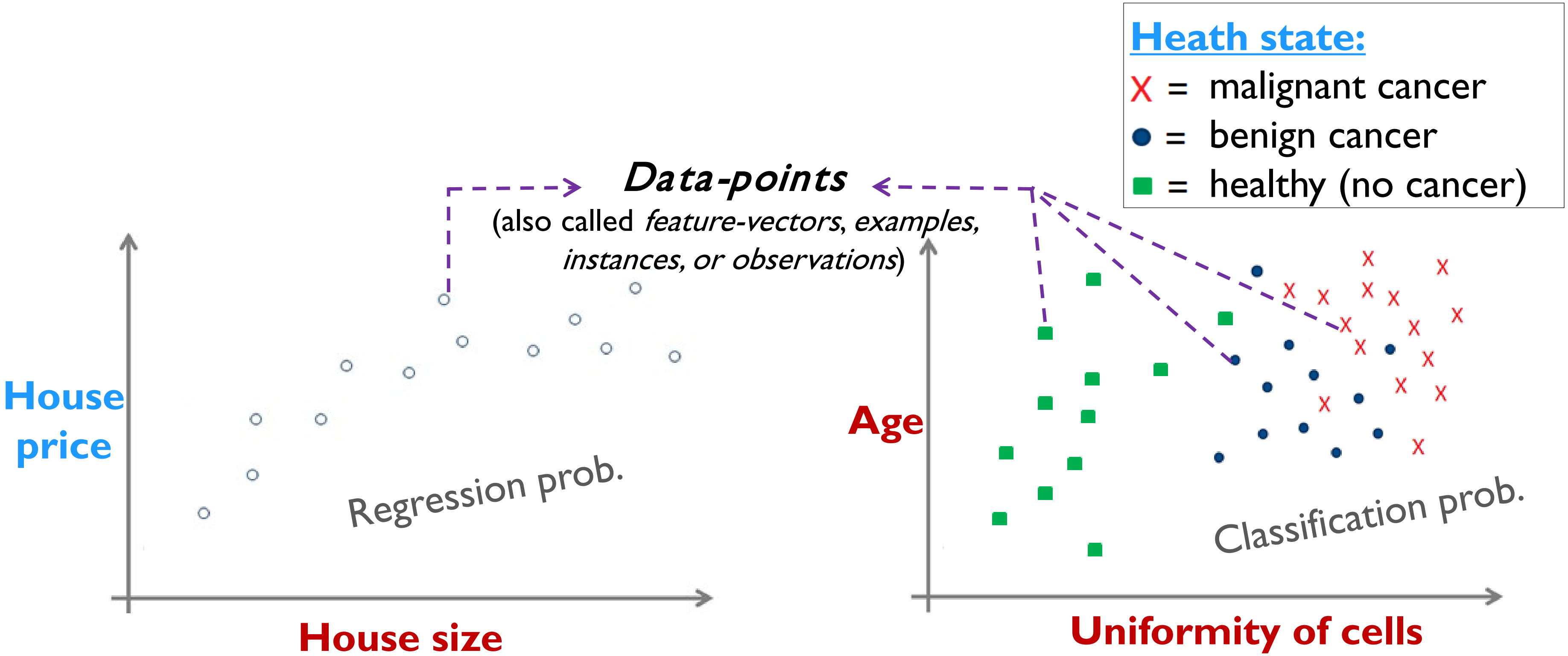
Dataset representation

Dataset representation



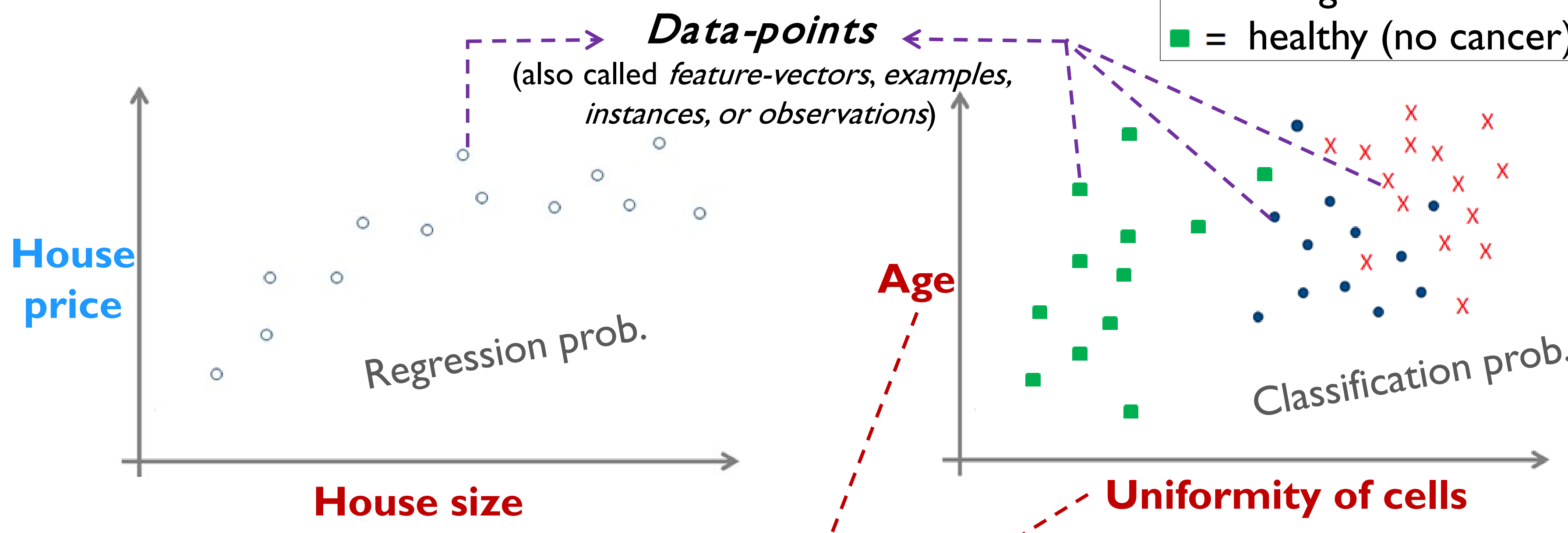


Dataset representation



Dataset representation

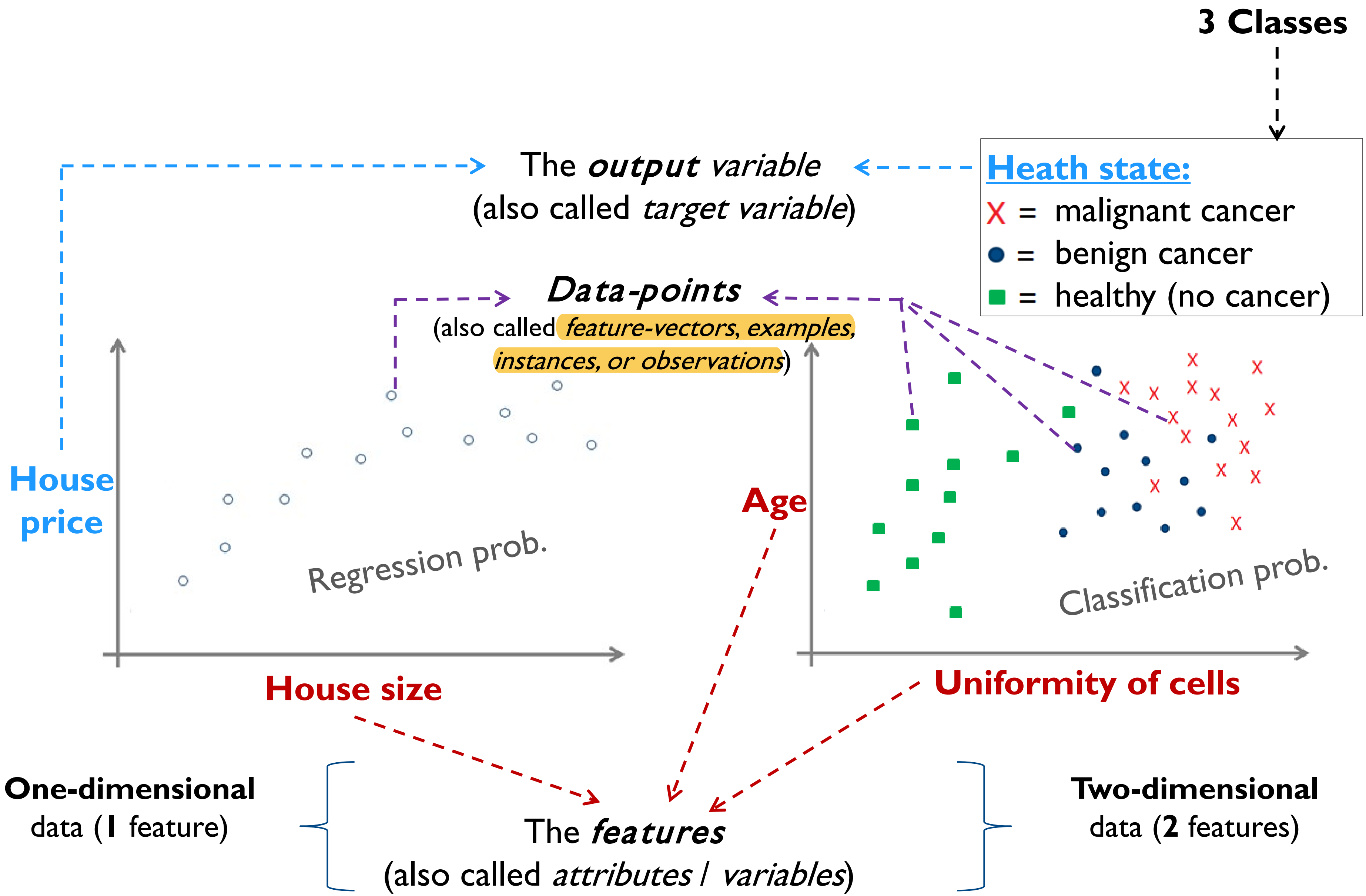
Heath state:
X = malignant cancer
● = benign cancer
■ = healthy (no cancer)



One-dimensional data (**1** feature)

The ***features***
(also called *attributes* / *variables*)

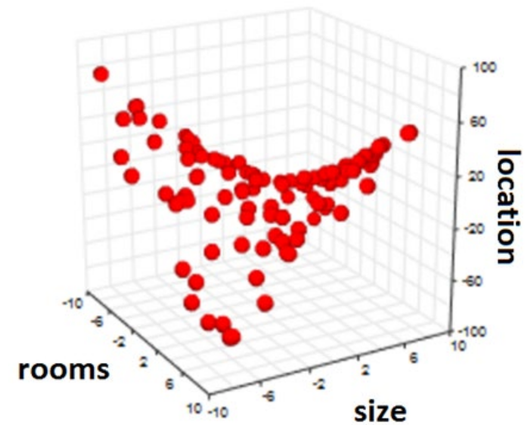
Two-dimensional data (**2** features)


3 Classes
The output variable
(also called target variable)
Heath state:
X = malignant cancer
● = benign cancer
■ = healthy (no cancer)
Data-points
(also called feature-vectors, examples, instances, or observations)
House price
Regression prob.
Age
Classification prob.
House size
Uniformity of cells
One-dimensional data (1 feature)
The features
(also called attributes / variables)
Two-dimensional data (2 features)

Dataset representation - notations

- Assume we have a set of n houses $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$

- Each house $x^{(i)}$ is characterized by:
  1. its size
  2. its number of rooms
  3. its location (distance from the city center)



- This is a 3-dimensional data (we have $d = 3$ features). So, each data-point $x^{(i)} \in R^3$ is represented as a feature-vector:
  - $x^{(1)} = < 80, 3, 4 >$
  - $x^{(2)} = < 20, 2, 3 >$
  - $\ldots$

- Let $x_j^{(i)}$ be the j^{th} feature value of the i^{th} house. It's a scalar value:
  - $x_1^{(1)} = 80$ $\quad x_2^{(1)} = 3$ $\quad x_3^{(1)} = 4$
  - $x_1^{(2)} = 20$ $\quad x_2^{(2)} = 2$ $\quad x_3^{(2)} = 3$
  - $\ldots$

$$\begin{array}{cc} \begin{array}{ccc} \text{size} & \text{rooms} & \text{location} \end{array} & \\ \left[\begin{array}{ccc} 80, & 3, & 4 \\ 20, & 2, & 3 \\ \cdots & & \\ \cdots & & \\ 47, & 2, & 7 \end{array}\right] & \begin{array}{l} \rightarrow x^{(1)} \\ \rightarrow x^{(2)} \\ \cdots \\ \cdots \\ \rightarrow x^{(n)} \end{array} \end{array}$$

➔ The data is represented as a matrix of n rows and d columns (here $d = 3$ features)

Dataset representation - notations

- We want to train a supervised ML algorithm to predict the price of new houses.

- We need first to prepare a **training dataset** which consists of:
  - The input data $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$
  - The real price (output) $y^{(i)}$ associated to each training data-point $x^{(i)}$
    - NOTE: These real prices are given to <u>teach</u> (or supervise) the algorithm, so that it <u>learns</u> (or models) the relation between "*the features that characterizes the input data*", and the "*desired output*" (price).

- The i^{th} house has a price $y^{(i)}$ (a scalar value) and is characterized by a feature-vector $x^{(i)}$. So, our **training dataset** is: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$

- Can, also be represented as matrix $\mathbf{X}$ and a vector of prices $\mathbf{y}$

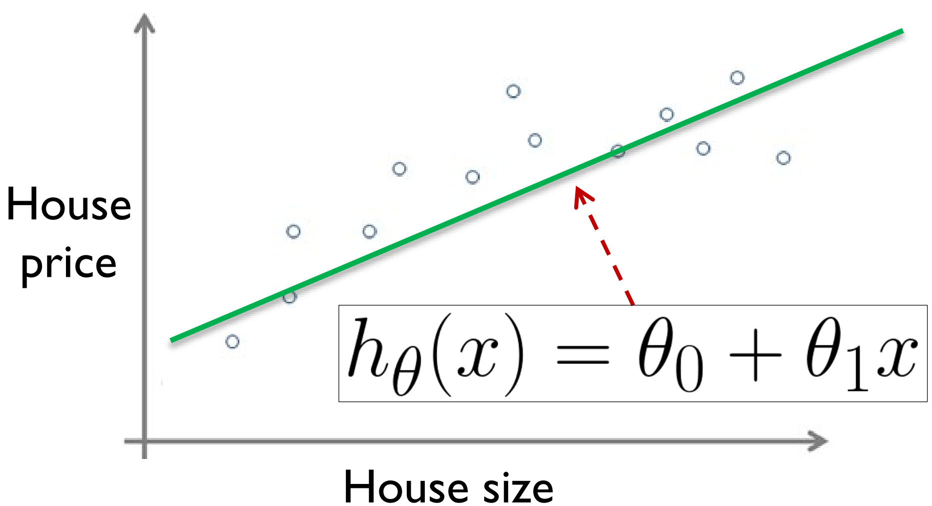
$$\mathbf{X} = \begin{array}{c} \begin{array}{cccc} \text{size} & \text{rooms} & \cdots & \text{location} \end{array} \\ \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \cdots & & & \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} \end{array} \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix}$$

Model representation

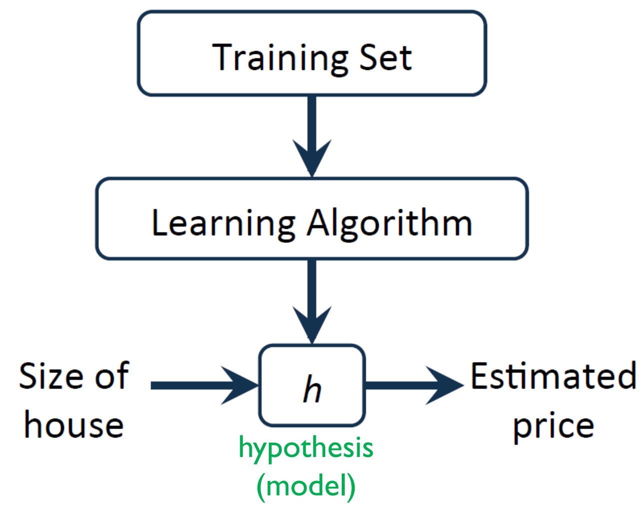
Model representation

- The model (to be learned) is a function h (called hypothesis).
- The model has parameters $\theta_0, \theta_1, \ldots$
- $\theta = <\theta_0, \theta_1, \ldots>$ is the vector of parameters, so the model is denoted as h_θ
- Learning (or training) means finding the optimal parameters on a given dataset.

- In this example, as we have one feature (house size), the input x is a scalar value (or just a one-dimensional vector).
- $h_\theta(x)$ is the *predicted* price for the input x using the model h_θ

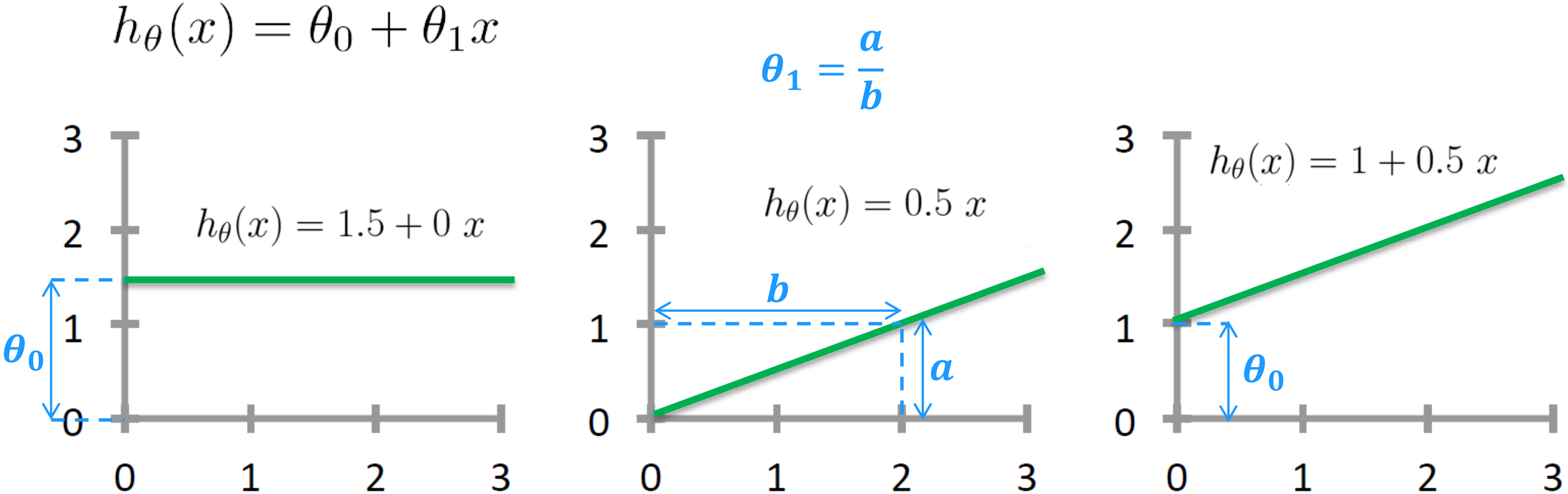


$$h_\theta(x) = \theta_0 + \theta_1 x$$



Model representation

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



$\theta_0 = 1.5$
$\theta_1 = 0$

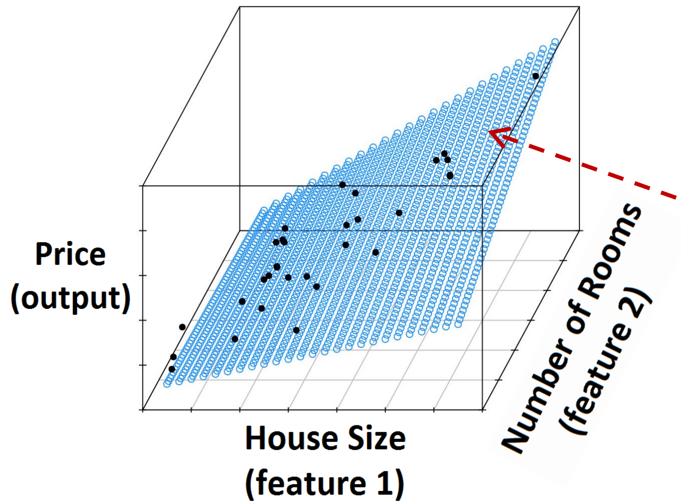
$\theta_0 = 0$
$\theta_1 = 0.5$

$\theta_0 = 1$
$\theta_1 = 0.5$

How to choose θ_0 and θ_1 ➔ We will see this in the next lecture.

Model representation

- The model (to be learned) is a function h (called hypothesis).
- The model has parameters $\theta_0, \theta_1, \ldots$
- $\theta = < \theta_0, \theta_1, \ldots >$ is the vector of parameters, so the model is denoted as h_θ
- Learning (or training) means finding the optimal parameters on a given dataset.



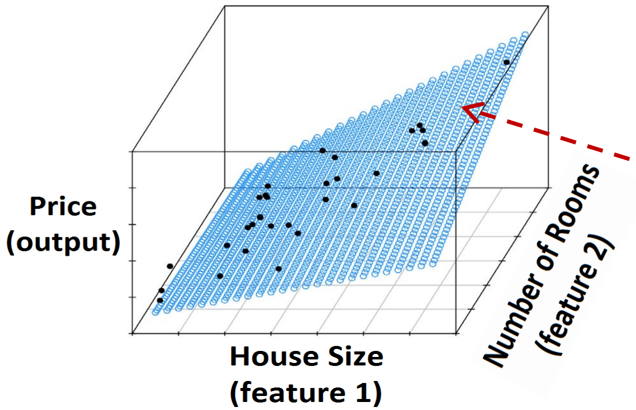
$$\boxed{h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2}$$

$$x = < x_1, x_2 >$$

- In this example, as we have two features (house size, number of rooms), the input $\boldsymbol{x} = < \boldsymbol{x_1}, \boldsymbol{x_2} >$ is a two-dimensional vector.
- $h_\theta(x)$ is the *predicted* price for the input x using the model h_θ

Model representation

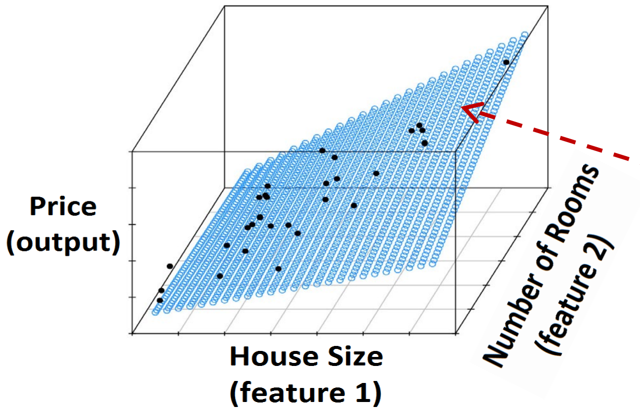
- How would you write the equation in a more compact format (using vectors) ?



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

Model representation

- How would you write the equation in a more compact format (using vectors) ?



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

- Just redefine x as: $x = < 1, x_1, x_2 >$ including 1 at the beginning.
- We have $\theta = < \theta_0 , \theta_1, \theta_2 >$
- So: $\boldsymbol{h_\theta(x) = \theta^T x} = x^T \theta = x.\theta = \theta.x$ ➔ dot product between two vectors.

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix}}_{\boldsymbol{x^T}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}}_{\boldsymbol{\theta}} = \underbrace{\theta_0 + \theta_1 x_1 + \theta_2 x_2}_{\boldsymbol{h_\theta(x)}}$$

Error of a model

- Given a dataset: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$
- The error $E(\theta)$ of a model h_θ is on this dataset is:

$$E(\theta) = \sum_{i=1}^{n} \left[\underbrace{h_\theta(x^{(i)})}_{\text{The predicted output for the data-point } x^{(i)}} - \underbrace{y^{(i)}}_{\text{The true output for } x^{(i)}} \right]^2$$

The predicted output for the data-point $x^{(i)}$

e.g. the predicted price of the i^{th} house.

The true output for $x^{(i)}$

e.g. the true price of the i^{th} house

NOTE: The *error function* is also sometimes called "*cost function*" or "*loss function*".

Notations to remember

- $x^{(i)} \in R^d$ the i^{th} **data-point** (or feature-vector). It is a d-dimensional vector.
- $x_j^{(i)} \in R$ the value of the $\boldsymbol{j^{th}}$ **feature** (or attribute, or variable) in the data-point $x^{(i)}$.
- $y^{(i)}$ the value of the **output** variable (or target variable), for the i^{th} data-point.
  $y^{(i)} \in R$ in regression, and $y^i \in N$ in classification.
- $\mathbf{X} \in R^{n \times d}$ a **dataset** represented as a matrix of n lines and d columns.
- $\theta \in R^p$ a vector representing the model **parameters**. It has p parameters.
  Sometimes also called **weights** vector.
- h_θ a **model** (hypothesis function) with parameters $\theta = < \theta_0 , \theta_1, \dots >$.
- $h_\theta(x)$ the output *predicted* by the model h_θ for the data-point x.
- $E(\theta)$ the **error** (or cost, or loss) of a model h_θ, computed on some dataset.

HALMSTAD UNIVERSITY

Some notions of Linear Algebra

Matrices and Vectors

$$A = \begin{bmatrix} 1402 & 191 \\ 1371 & 821 \\ 949 & 1437 \\ 147 & 1448 \end{bmatrix}$$

$$A \in \mathbb{R}^{4\times 2}$$

The matrix A has a dimension of 4×2

$A_{ij} =$ "i, j entry" in the i^{th} row, j^{th} column.

- A vector is simply an $n \times 1$ matrix

$$u = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

$$u \in \mathbb{R}^4$$

u_i is the i^{th} element of u

Matrix addition

$$\begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 0.5 \\ 2 & 5 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0.5 \\ 4 & 10 \\ 3 & 2 \end{bmatrix}$$

$$\cancel{\begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 0.5 \\ 2 & 5 \end{bmatrix}} = Error$$

Scalar Multiplication

$$3 \times \begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 6 & 15 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 5 \\ 3 & 1 \end{bmatrix} \times 3$$

$$\begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} / 4 = \frac{1}{4} \begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{3}{2} & \frac{3}{4} \end{bmatrix}$$

Combination of operands

$$3 \times \begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} - \begin{bmatrix} 4 & 0 \\ 6 & 3 \end{bmatrix} /3$$

Matrix Vector multiplication

$$\underbrace{\begin{bmatrix}1 & 3\\4 & 0\\2 & 1\end{bmatrix}}_{3\times 2}\underbrace{\begin{bmatrix}1\\5\end{bmatrix}}_{2\times 1} = \begin{bmatrix}1\times 1+3\times 5\\4\times 1+0\times 5\\2\times 1+1\times 5\end{bmatrix} = \underbrace{\begin{bmatrix}16\\4\\7\end{bmatrix}}_{3\times 2}$$

$$\begin{bmatrix}2 & 1\end{bmatrix}\begin{bmatrix}1\\5\end{bmatrix} = \begin{bmatrix}2\\1\end{bmatrix}\cdot\begin{bmatrix}1\\5\end{bmatrix}$$

Just a dot product between two vectors

Matrix Vector multiplication

- Example: to predict the outputs of all data-points in a dataset using a linear model h_θ, just multiply the dataset matrix by the vector of parameters θ

$$\begin{array}{cc} & \begin{array}{ccc} & \text{size} & \text{rooms} \end{array} \\ \begin{array}{c} x^{(1)} \rightarrow \\ x^{(2)} \rightarrow \\ \cdots \\ \cdots \\ x^{(n)} \rightarrow \end{array} & \underbrace{\begin{bmatrix} 1 & 80 & 3 \\ 1 & 20 & 2 \\ \cdots & & \\ \cdots & & \\ 1 & 47 & 2 \end{bmatrix}}_{\boldsymbol{n \times 3}} \end{array} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}}_{\boldsymbol{3 \times 1}} = \begin{bmatrix} \theta_0 + \theta_1 \times 80 + \theta_2 \times 3 \\ \theta_0 + \theta_1 \times 20 + \theta_2 \times 2 \\ \cdots \\ \cdots \\ \theta_0 + \theta_1 \times 47 + \theta_2 \times 2 \end{bmatrix} = \underbrace{\begin{bmatrix} h_\theta(x^{(1)}) \\ h_\theta(x^{(1)}) \\ \cdots \\ \cdots \\ h_\theta(x^{(n)}) \end{bmatrix}}_{\boldsymbol{n \times 1}}$$

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

Matrix Matrix multiplication

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 0 & 1 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 11 & 10 \\ 9 & 14 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 5 \end{bmatrix} = \begin{bmatrix} 11 \\ 9 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 10 \\ 14 \end{bmatrix}$$

Matrix Matrix multiplication

- Example: to predict the outputs of all data-points in a dataset using several linear models (h_θ, g_θ, f_θ) just multiply the dataset matrix by a matrix that contains on each column the parameters of one model.

$$\begin{cases} h_\theta(x) = -40 + 0.25\ x_1 + 3\ x_2 \\ g_\theta(x) = 200 + 0.1\ x_1 + 5\ x_2 \\ f_\theta(x) = -150 + 0.4\ x_1 - 1\ x_2 \end{cases}$$

Predictions of h

Predictions of g

Predictions of f

$$\underbrace{\overbrace{\begin{bmatrix} 1 & 80 & 3 \\ 1 & 20 & 2 \\ & \dots & \\ & \dots & \\ 1 & 47 & 2 \end{bmatrix}}^{\text{Dataset matrix}}}_{n \times 3} \underbrace{\overbrace{\begin{bmatrix} -40 & 200 & -2 \\ 0.25 & 0.1 & 0.4 \\ 30 & 5 & -1 \end{bmatrix}}^{\text{Each column is the parameters of one model}}}_{3 \times 3} = \underbrace{\begin{bmatrix} 70 & 223 & 27 \\ 25 & 212 & 4 \\ & \dots & \\ & \dots & \\ 31.7 & 214.7 & 14.8 \end{bmatrix}}_{n \times 3}$$

Matrix multiplication properties

- **Matrix multiplication is not commutative**

Let A and B be matrices. Then in general, $A \times B \neq B \times A$. (not commutative.)

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \end{bmatrix}$$

- **Matrix multiplication is associative**

$$A \times B \times C.$$

$$\left.\begin{array}{l} \text{Let } D = B \times C. \quad \text{Compute } A \times D \\ \text{Let } E = A \times B. \quad \text{Compute } E \times C \end{array}\right\} \text{Same result}$$

Identity matrix, inverse, and transpose

- **Identity matrix**

Denoted I

Examples of identity matrices:

$$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$
2 x 2

$$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$$
3 x 3

$$\begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 0 & 0\\0 & 0 & 1 & 0\\0 & 0 & 0 & 1\end{bmatrix}$$
4 x 4

For any matrix A,

$$A \quad I = I \quad A = A$$

- **Inverse of a matrix**

If A is an $n \times n$ matrix, and if it has an inverse, then:

$$AA^{-1} = A^{-1}A = I$$

$$\begin{bmatrix}3 & 4\\2 & 16\end{bmatrix}\begin{bmatrix}0.4 & -0.1\\-0.05 & 0.075\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} = I$$

- **Transpose**

$$A = \begin{bmatrix}1 & 2 & 0\\3 & 5 & 9\end{bmatrix}$$

$$A^T = \begin{bmatrix}1 & 3\\2 & 5\\0 & 9\end{bmatrix}$$

Norm of a vector

The 2-norm of a vector $x \in \mathbb{R}^d$ is: $\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}$

Example:

$$x = \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix}$$

The 2-norm (or l_2 norm, or Euclidian norm) of the vector is:

$$\|x\|_2 = \|x\| = \sqrt{3^2 + 1^2 + 5^2}$$

More generally:

The p-norm of a vector $x \in \mathbb{R}^d$ is: $\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}$

Euclidian distance:

The Euclidian distance between two vectors x and z, is the Euclidian norm of their difference:

$$\|x - z\| = \sqrt{\sum_{i=1}^{d} (x_i - z_i)^2}$$

Norm of a vector

$$||u||^2 = u^T u = \sum_j {u_j}^2$$

Example:

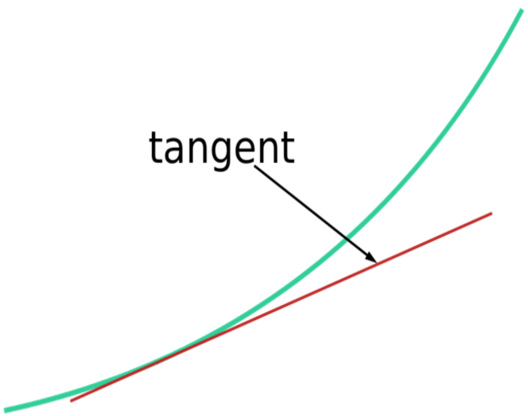
$$u = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}$$

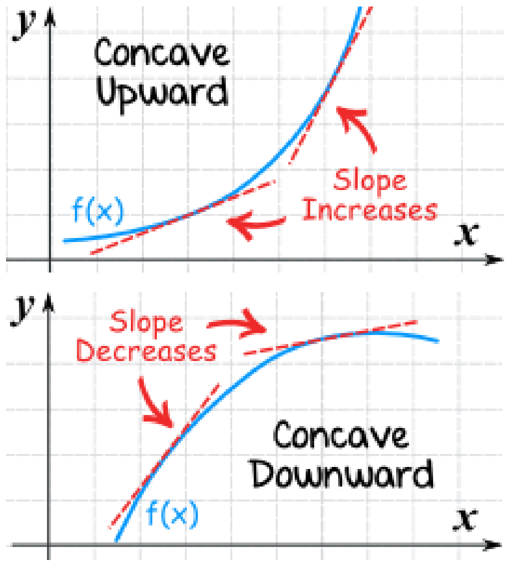
$$||u||^2 = [3\ \ 2\ \ 5] \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} = 3^2 + 2^2 + 5^2 = 38$$

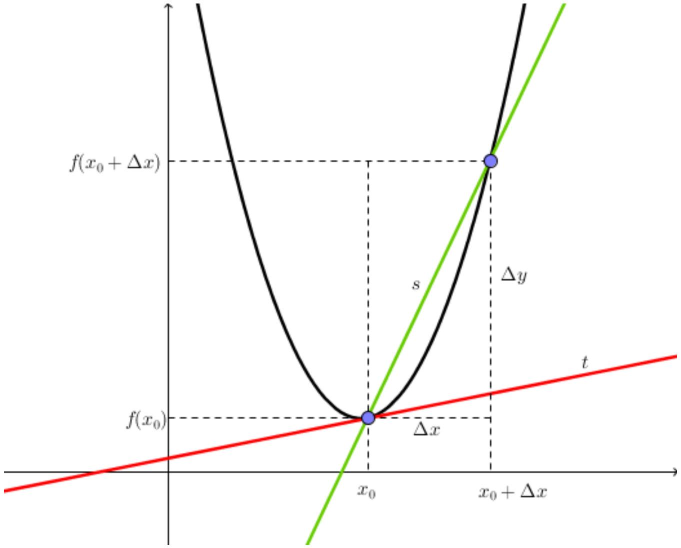
Derivatives

Definition of a derivative

$$f'(x) = \frac{\mathrm{d}f(x)}{\mathrm{d}x} = \lim_{\Delta x \to 0} \left(\frac{f(x + \Delta x) - f(x)}{\Delta x} \right)$$







Derivatives – Time saving rules

- *Sum Rule*:

$$\frac{\mathrm{d}}{\mathrm{d}x}(f(x)+g(x)) = \frac{\mathrm{d}}{\mathrm{d}x}(f(x)) + \frac{\mathrm{d}}{\mathrm{d}x}(g(x))$$

- *Power Rule*:

$$\begin{aligned} \text{Given} \quad & f(x) = ax^b, \\ \text{then} \quad & f'(x) = abx^{(b-1)} \end{aligned}$$

- *Product Rule*:

$$\begin{aligned} \text{Given} \quad & A(x) = f(x)g(x), \\ \text{then} \quad & A'(x) = f'(x)g(x) + f(x)g'(x) \end{aligned}$$

- *Chain Rule*:

$$\begin{aligned} \text{Given} \quad & h = h(p) \quad \text{and} \quad p = p(m), \\ \text{then} \quad & \frac{\mathrm{d}h}{\mathrm{d}m} = \frac{\mathrm{d}h}{\mathrm{d}p} \times \frac{\mathrm{d}p}{\mathrm{d}m} \end{aligned}$$

Question:
Compute the derivative of the error function E with respect to each parameter of the linear model $h_\theta(x) = \theta_0 + \theta_1 x$

$$E(\theta) = \sum_{i=1}^{n} \left[h_\theta(x^{(i)}) - y^{(i)} \right]^2$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{x}\right) &= -\frac{1}{x^2} \\ \frac{\mathrm{d}}{\mathrm{d}x}(\sin(x)) &= \cos(x) \\ \frac{\mathrm{d}}{\mathrm{d}x}(\cos(x)) &= -\sin(x) \\ \frac{\mathrm{d}}{\mathrm{d}x}(\exp(x)) &= \exp(x) \end{aligned}$$

Example:
Compute the derivative of the function E

$$E(\theta_0, \theta_1) = \frac{1}{2n}\sum_{i=1}^{n}\left[h_\theta(x^{(i)}) - y^{(i)}\right]^2$$

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

Let: $g(\theta) = \left[h_\theta(x^{(i)}) - y^{(i)}\right] = \left[\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right]$

- **Derivative of $E(\theta_0, \theta_1)$ with respect to θ_0**

$$\frac{\partial}{\partial\theta_0}E(\theta) = \frac{\partial}{\partial\theta_0}\frac{1}{2n}\sum_{i=0}^{n}g(\theta)^2 = \frac{1}{2n}\sum_{i=0}^{n}2g(\theta)\frac{\partial g}{\partial\theta_0} = \boxed{\frac{1}{n}\sum_{i=0}^{n}\left[h_\theta(x^{(i)}) - y^{(i)}\right]}$$

- **Derivative of $E(\theta_0, \theta_1)$ with respect to θ_1**

$$\frac{\partial}{\partial\theta_1}E(\theta) = \frac{\partial}{\partial\theta_1}\frac{1}{2n}\sum_{i=0}^{n}g(\theta)^2 = \frac{1}{2n}\sum_{i=0}^{n}2g(\theta)\frac{\partial g}{\partial\theta_1} = \boxed{\frac{1}{n}\sum_{i=0}^{n}\left[h_\theta(x^{(i)}) - y^{(i)}\right]x^{(i)}}$$

Reading

- Please read the complementary document: *Revision of Linear Algebra and Probability.pdf* on blackboard, for a more exhaustive revision of math prerequisites.