MATHEMATICS

FOR MACHINE LEARNING

Nagoya University, Fall 2023

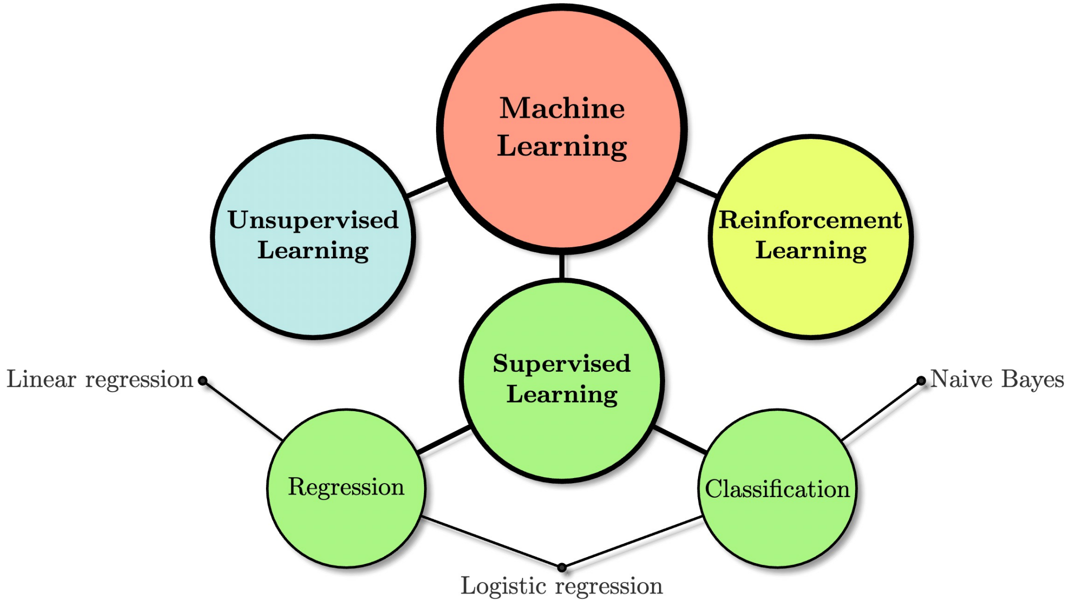
Lecture 3

Polynomial regression & Logistic regression

https://www.henrikbachmann.com/mml2023.html

Lecture notes & Tutorial

- You can now find the first version of the lecture notes on the homepage. (may contain typos)
- Anyone interested in helping out with writing them (latex in overleaf) feel free to contact me.

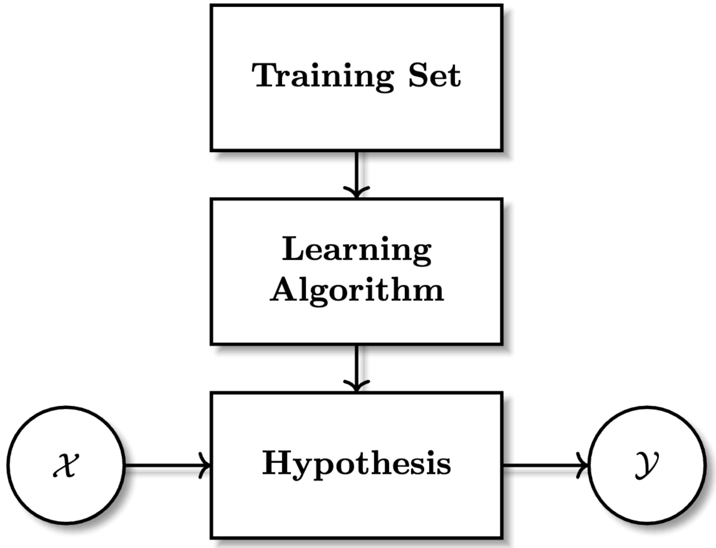


This week the Tutorial is Friday 6th period!

27th October

1 Supervised learning: Notations Recall

- Input values (Feature space): $\mathcal{X}$
- Output value (Label space): $\mathcal{Y}$
- Trainings example: $(x, y) \in \mathcal{X} \times \mathcal{Y}$.
- Trainings set (with n training examples): $\mathcal{T} = \big((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\big) \in (\mathcal{X} \times \mathcal{Y})^n$.
- hypothesis: A function $h : \mathcal{X} \to \mathcal{Y}$.
- Learning algorithm: An algorithm to create a hypothesis h out of a trainings set $\mathcal{T}$.



- Learning algorithms make an Ansatz (educated guess) for a hypothesis involving certain **parameters**.
- Learning: Find good parameters depending on the trainings set.

1 Supervised learning – Linear Regression **Recall**

Learning Algorithm: Linear Regression

Let $\mathcal{X} = \mathbb{R}^d$, i.e. we have d features, and $\mathcal{Y} = \mathbb{R}$. As an Ansatz for the hypothesis we set

$$h_\theta(x) := \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d = \sum_{i=0}^{d} \theta_i x_i \,,$$

with **parameters (weights)** $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T \in \mathbb{R}^{d+1}$. In the second equation we set $x_0 := 1$.

For a given training set $\mathcal{T} = \big((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\big)$ we define the **cost function** by

$$J(\theta) = \frac{1}{2} \sum_{j=1}^{n} (h_\theta(x^{(j)}) - y^{(j)})^2 \,.$$

The cost function is a function $J : \mathbb{R}^{d+1} \to \mathbb{R}$, which we want to minimize.

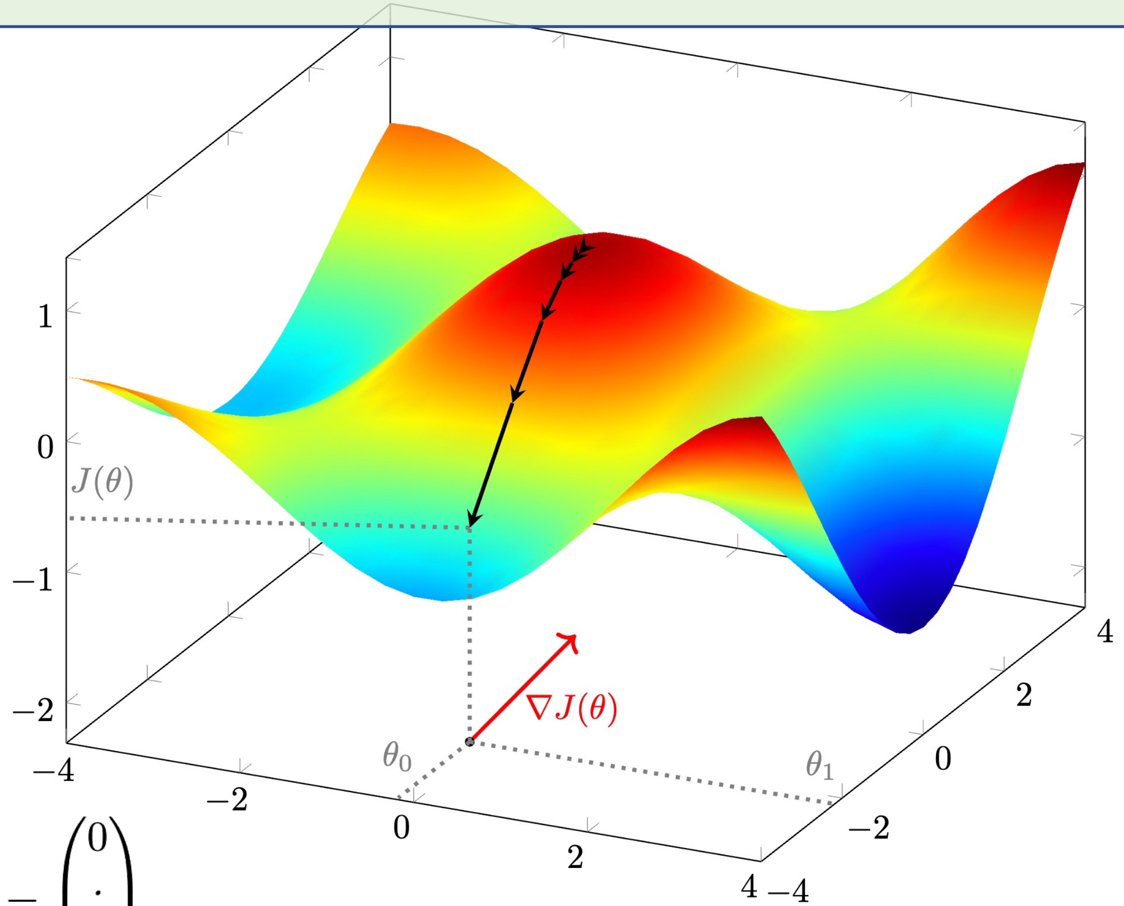
Goal: Minimize the cost function for a given trainings set.

Recall

1 Supervised learning – Solution 1: Gradient descent

The **gradient of** J is defined by

$$\nabla J = \begin{pmatrix} \frac{\partial}{\partial \theta_0} J \\ \frac{\partial}{\partial \theta_1} J \\ \vdots \\ \frac{\partial}{\partial \theta_d} J \end{pmatrix}$$



Start with a random starting value, e.g. $\theta = 0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$.

Do several times

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

Learning rate

Recall

1 Linear Regression: Solution 2 - Use linear algebra

For a training set $\mathcal{T} = ((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}))$ we define

$$A = \begin{pmatrix} | & & | \\ x^{(1)} & \ldots & x^{(n)} \\ | & & | \end{pmatrix}^T = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_d^{(1)} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_1^{(n)} & \ldots & x_d^{(n)} \end{pmatrix} \in \mathbb{R}^{n \times d+1} \qquad y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}$$

Proposition 3.5. *If $\theta \in \mathbb{R}^{d+1}$ is a solution to*

Normal equation

$$A^T A\theta = A^T y\,,$$

then $||A\theta - y||$ is minimal and consequently $J(\theta)$ is minimal.

$$\theta = (A^T A)^{-1} A^T y\,.$$

$$d=2 \quad T=\left(\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right)\right)$$

$$x^{(i)} \in \mathbb{R}^2, \quad y^{(i)} \in \mathbb{R}$$

$$T = \left(\left(\begin{pmatrix} 5 \\ 100 \end{pmatrix}, 6 \right), \left(\begin{pmatrix} 7 \\ 150 \end{pmatrix}, 10 \right) \right)$$

($5 = x_1^{(1)}$, $100 = x_2^{(1)}$, $\begin{pmatrix} 5 \\ 100 \end{pmatrix} = x^{(1)}$, $6 = y^{(1)}$; $7 = x_1^{(2)}$, $150 = x_2^{(2)}$, $\begin{pmatrix} 7 \\ 150 \end{pmatrix} = x^{(2)}$, $10 = y^{(2)}$)

$$A = \begin{pmatrix} 1 & 5 & 100 \\ 1 & 7 & 150 \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \end{pmatrix}$$

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

$$y = \begin{pmatrix} 6 \\ 10 \end{pmatrix}$$

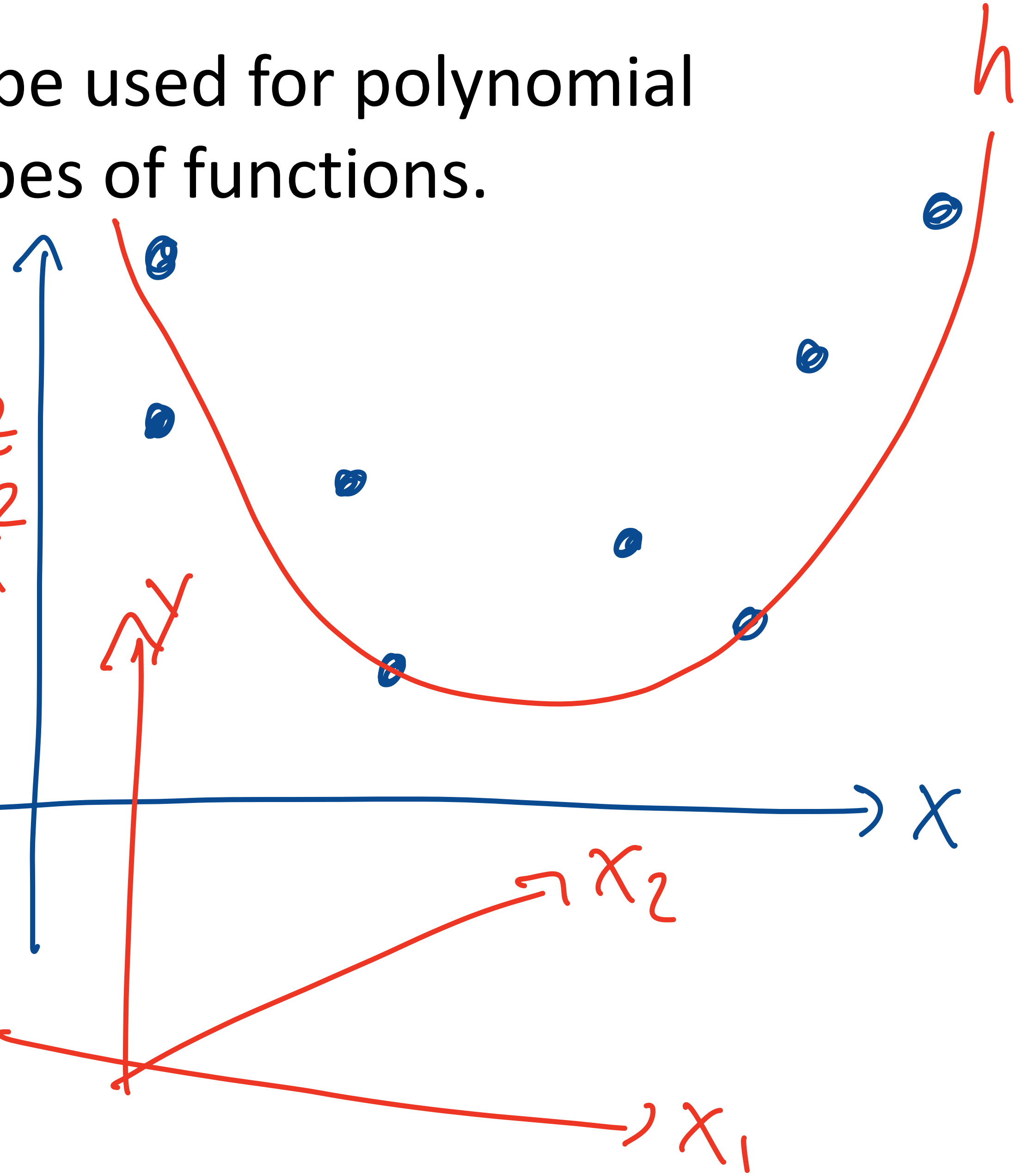
Solve $A^T A \theta = A^T y$

$$\underbrace{\begin{pmatrix} 1 & 1 \\ 5 & 7 \\ 100 & 150 \end{pmatrix} \begin{pmatrix} 1 & 5 & 100 \\ 1 & 7 & 150 \end{pmatrix}}_{\begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 5 & 7 \\ 100 & 150 \end{pmatrix} \begin{pmatrix} 6 \\ 10 \end{pmatrix}$$

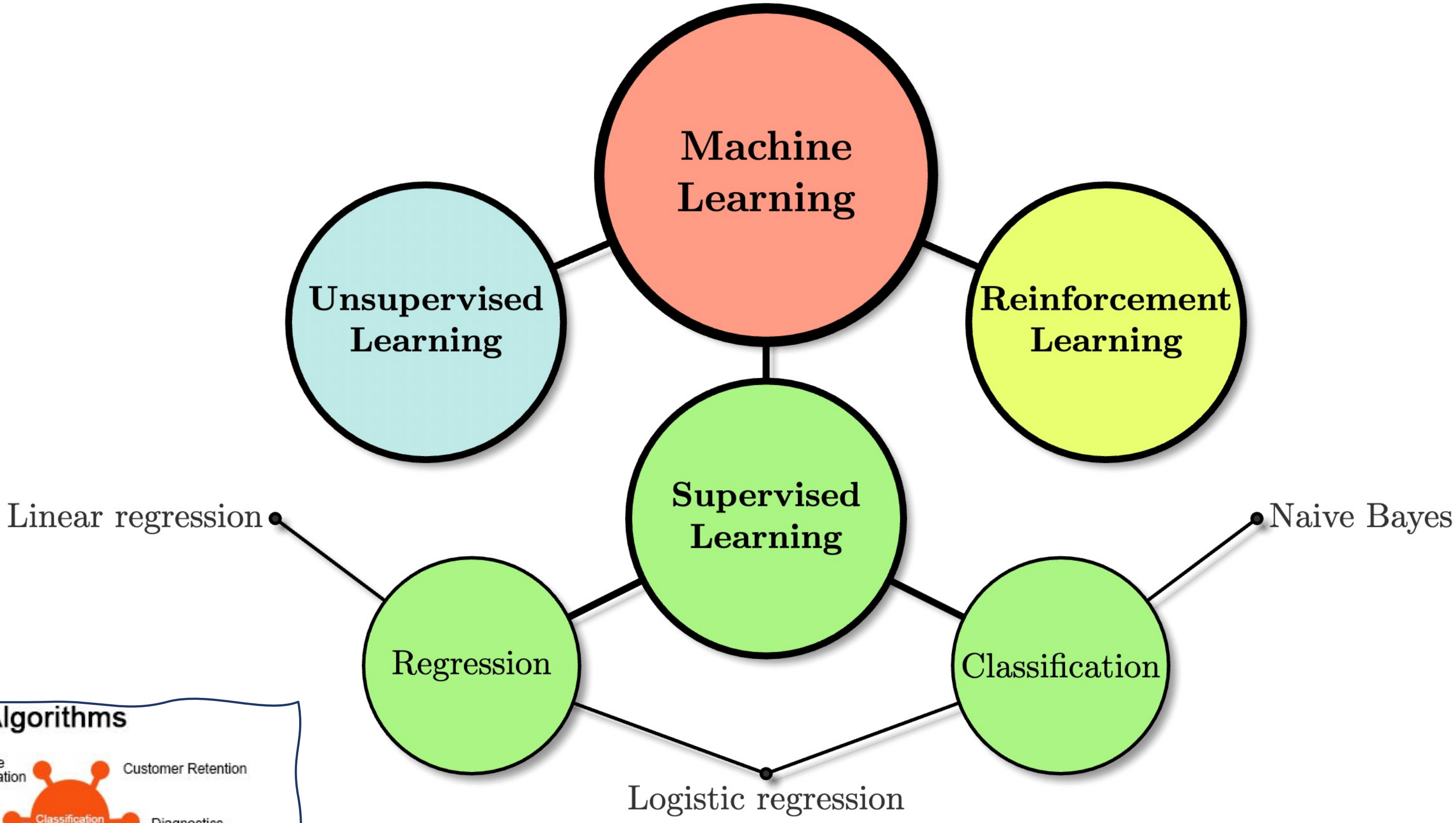
1 Linear Regression – Not only for "linear data"

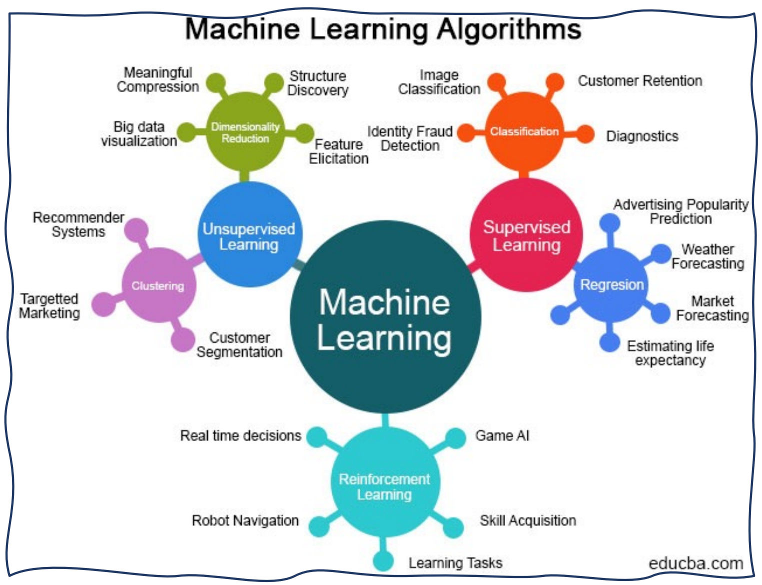
Linear regression can also be used for polynomial interpolation and other types of functions.

$$h_\theta(x) = \theta_0 + \theta_1 \underset{x_1}{\underbrace{x}} + \theta_2 \underset{x_2}{\underbrace{x^2}}$$



2 Binary Classification - Logistic Regression



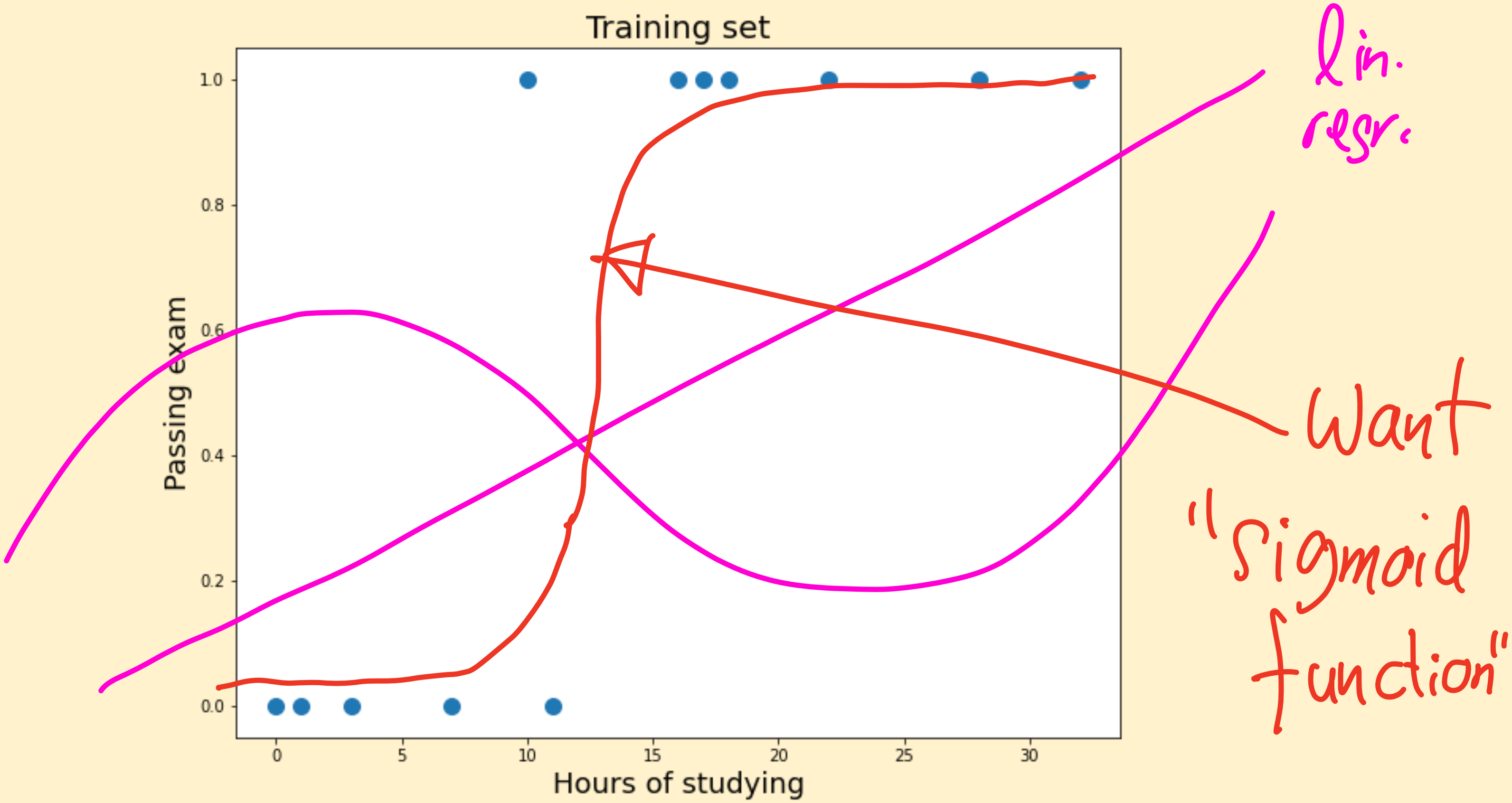


2 Binary Classification – Passing exam

Binary Classification $\mathcal{Y} = \{0, 1\}$

Passing exam example

Features: Hours studied for the exam
Labels: Failed exam (y=0), Passed exam (y=1)



Question: How to model the hypothesis?

2 Binary Classification – Logistic regression

The **logistic function** is defined by

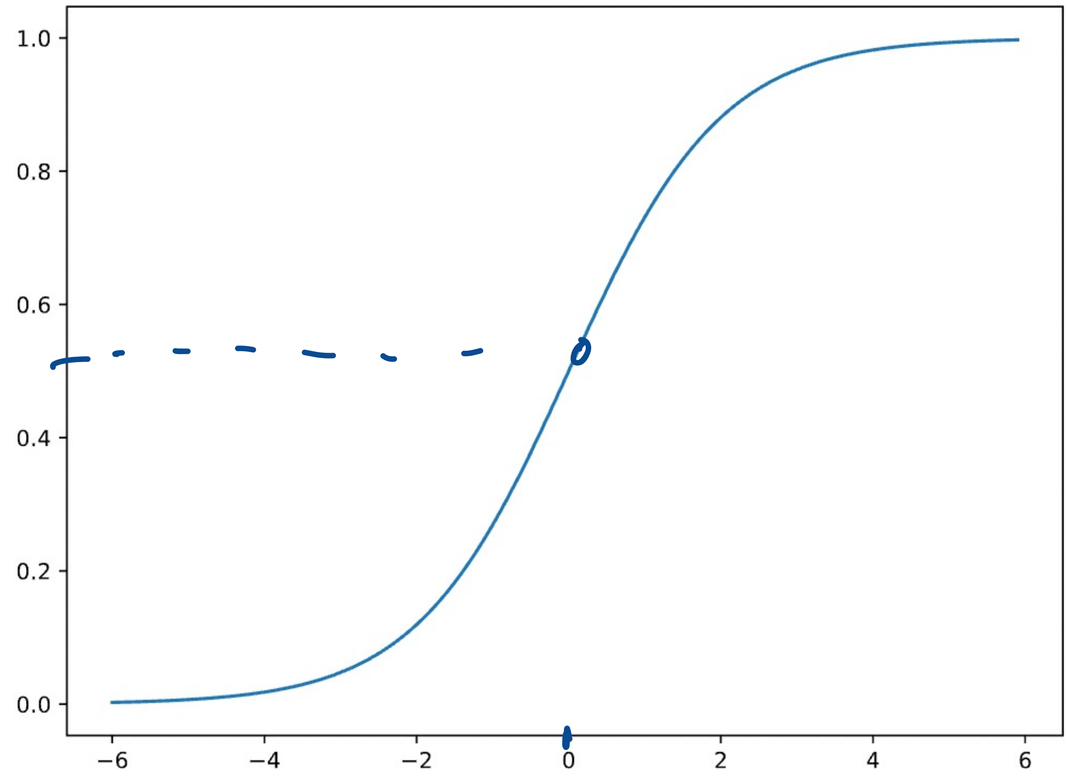
$$S(x) = \frac{1}{1 + e^{-x}} \,,$$

and its graph looks as follows

$S(0) = \frac{1}{2}$

$\lim_{x \to \infty} S(x) = 1$

$\lim_{x \to -\infty} S(x) =$



2 Binary Classification – Logistic regression

Recall: Linear regression

Hypothesis: $h_\theta(x) := \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d = \sum_{i=0}^{d} \theta_i x_i = \theta^T x$

Logistic regression

$$S(x) = \frac{1}{1 + e^{-x}}.$$

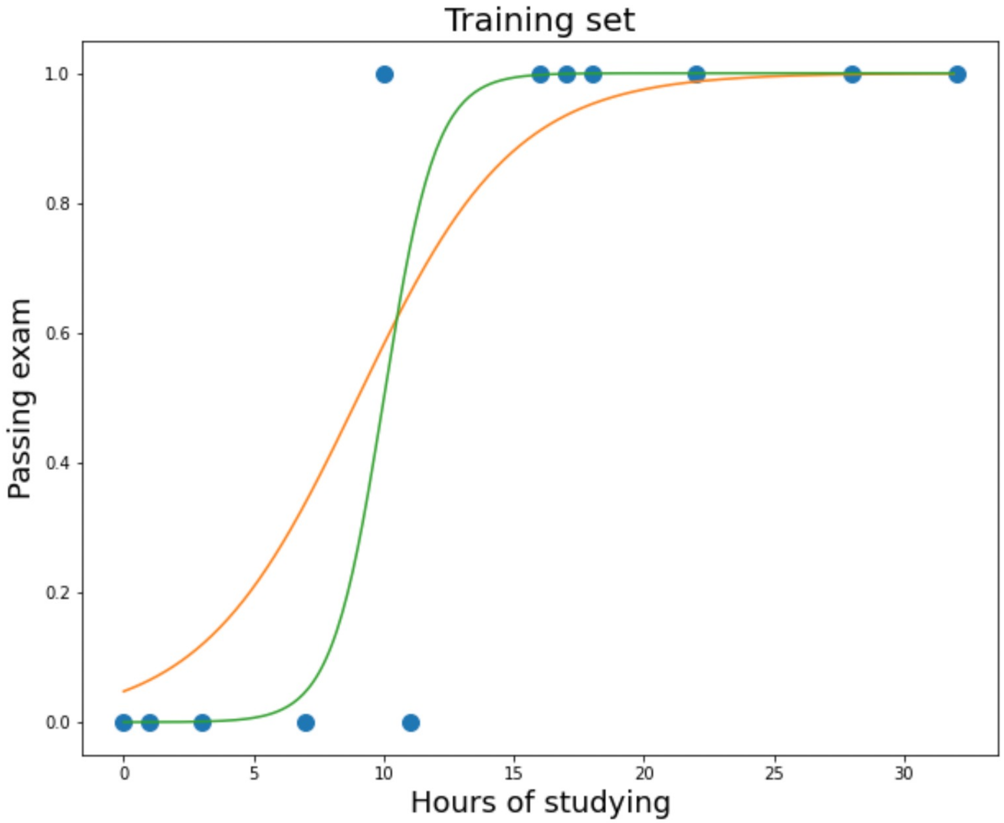
Hypothesis:

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

2 Binary Classification – Logistic regression

Passing exam example

$$h_\theta(x) = S(\theta_0 + \theta_1 x_1) = \frac{1}{1 + e^{-\theta_0 - \theta_1 x_1}}$$



Plot of $h_\theta(x)$ for $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} -10 \\ 1 \end{pmatrix}$

Plot of $h_\theta(x)$ for $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} -3 \\ \frac{1}{3} \end{pmatrix}$

Question: How to find the best weights?

2 Logistic regression – Probabilities

Notation: P(A|B) refers to the **conditional probability** that event A occurs, given that event B has occurred.

For fixed θ, the hypothesis $h_\theta(x)$ can be interpreted as the conditional probability of passing the exam $(y = 1)$ assuming that one studied x hours.

$$P(y = 1 \mid x; \theta) = h_\theta(x)\,.$$

The probability of failing the exam is therefore:

$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)\,.$$

2 Logistic regression – Probabilities

Notation: P(A|B) refers to the **conditional probability** that event A occurs, given that event B has occurred.

For fixed θ, the hypothesis $h_\theta(x)$ can be interpreted as the conditional probability of passing the exam ($y = 1$) assuming that one studied x hours.

$$P(y = 1 \mid x; \theta) = h_\theta(x)\,.$$

The probability of failing the exam is therefore:

$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)\,.$$

We can combine both into one single function, which gives back the above cases for $y \in \{0, 1\}$:

$$P(y \mid x; \theta) = h_\theta(x)^y \cdot (1 - h_\theta(x))^{1-y}\,.$$

2 Logistic regression – Maximum likelihood

Likelihood: “measures the goodness of fit of a statistical model to a sample of data"

Likelihood of parameters = Product over all probabilities in the training set

2 Logistic regression – Maximum likelihood

Likelihood: "measures the goodness of fit of a statistical model to a sample of data"

Likelihood of parameters = Product over all probabilities in the training set

For a training set $\mathcal{T} = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$, we define the **likelihood** of θ by

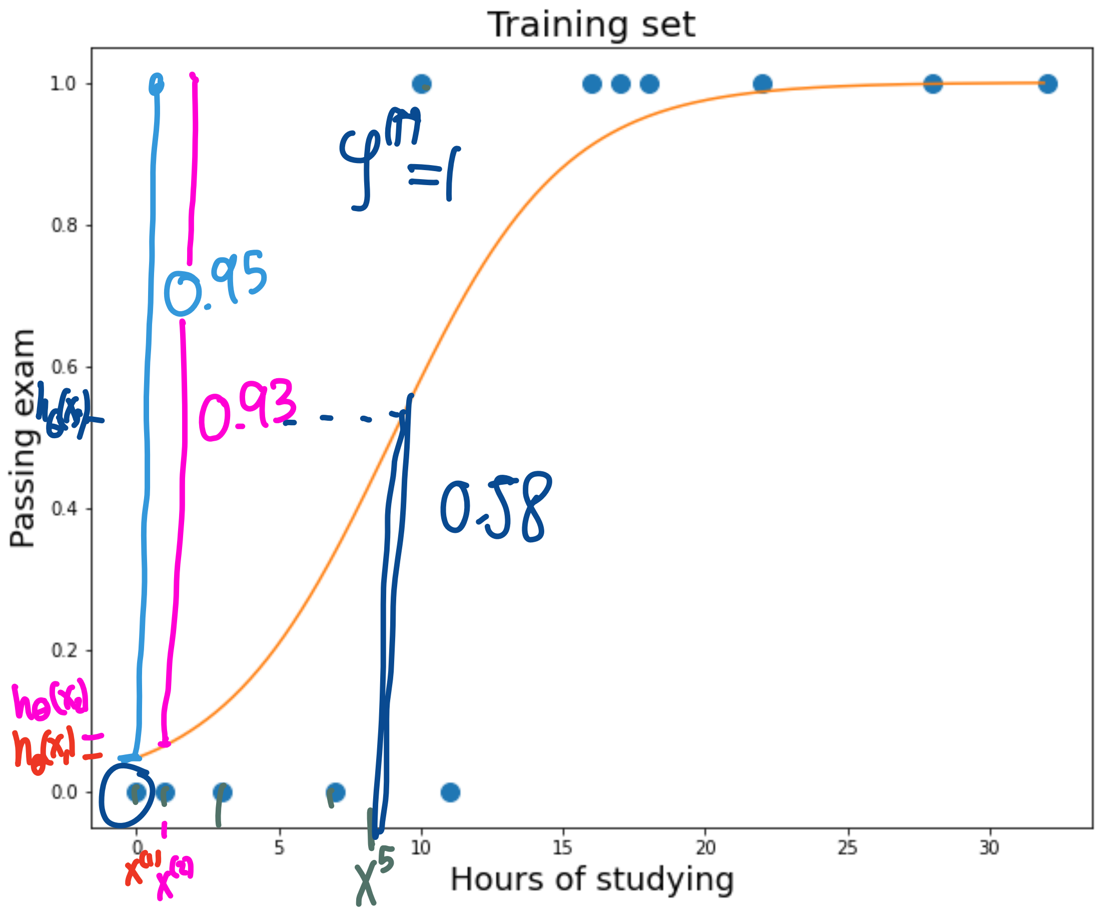
$$L(\theta) = \prod_{j=1}^{n} P(y^{(j)} \mid x^{(j)}; \theta)$$

$$= \prod_{j=1}^{n} h_\theta(x^{(j)})^{y^{(j)}} \cdot (1 - h_\theta(x^{(j)}))^{1-y^{(j)}}$$

Goal: Given a training set, find the parameters with the maximal **likelihood**.

2 Logistic regression – Maximum likelihood - Example

$$L(\theta) = \prod_{j=1}^{n} P(y^{(j)} \mid x^{(j)}; \theta) = \prod_{j=1}^{n} h_\theta(x^{(j)})^{y^{(j)}} \cdot (1 - h_\theta(x^{(j)}))^{1-y^{(j)}}$$

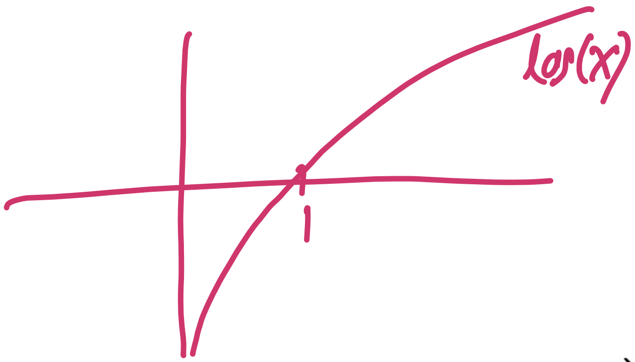


Plot of $h_\theta(x)$ for $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} -3 \\ \frac{1}{3} \end{pmatrix}$

2 Logistic regression – Maximum log likelihood

$$L(\theta) = \prod_{j=1}^{n} P(y^{(j)} \mid x^{(j)}; \theta) = \prod_{j=1}^{n} h_\theta(x^{(j)})^{y^{(j)}} \cdot (1 - h_\theta(x^{(j)}))^{1-y^{(j)}}$$

- Often it is easier to maximize the logarithm of the likelihood.
- The logarithm is monotonically increasing.
- The logarithm turns products into sums.



The **log likelihood** of θ is given by

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} \left(y^{(j)} \log h_\theta(x^{(j)}) + (1 - y^{(j)}) \log(1 - h_\theta(x^{(j)})) \right)$$

Goal: Given a training set, find the parameters with the maximal **log likelihood**.

1 Binary Classification – Gradient ascent

The **log likelihood** of θ is given by

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} y^{(j)} \log h_\theta(x^{(j)}) + (1 - y^{(j)}) \log(1 - h_\theta(x^{(j)}))$$

We want to maximize the log likelihood by using gradient ascent.

$$\theta := \theta + \alpha \nabla \ell(\theta).$$

Recall: Linear regression

Minimized the cost function J by gradient descent:

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

2 Logistic regression – Gradient ascent

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} y^{(j)} \log h_\theta(x^{(j)}) + (1 - y^{(j)}) \log(1 - h_\theta(x^{(j)}))$$

Gradient:

$$\nabla \ell(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \ell(\theta) \\ \frac{\partial}{\partial \theta_1} \ell(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_d} \ell(\theta) \end{pmatrix}$$

2 Logistic regression – Gradient ascent

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} y^{(j)} \log h_\theta(x^{(j)}) + (1 - y^{(j)}) \log(1 - h_\theta(x^{(j)}))$$

Gradient: $$\nabla \ell(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_0}\ell(\theta) \\ \frac{\partial}{\partial \theta_1}\ell(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_d}\ell(\theta) \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = x \in \mathbb{R}^{d+1}$$

$$S(\theta_0 + \theta_1 x_1 + \dots + \theta_d x_d) = h_\theta(x) = S(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Lemma 3.8. *The logistic function satisfies the following differential equation*

$$S(x) = (1 + e^{-x})^{-1} \qquad S'(x) = S(x)(1 - S(x)) . \qquad 1 - \frac{1}{1+e^{-x}}$$

$$\frac{d}{dx} S(x) = e^{-x} \frac{1}{(1+e^{-x})^2} = \underbrace{\frac{1}{1+e^{-x}}}_{S(x)} \frac{e^{-x}}{1+e^{-x}}$$

2 Logistic regression – Gradient ascent

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} y^{(j)} \log h_\theta(x^{(j)}) + (1 - y^{(j)}) \log(1 - h_\theta(x^{(j)}))$$

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

Lemma 3.8. *The logistic function satisfies the following differential equation*

$$S'(x) = S(x)(1 - S(x))\,.$$

$$\nabla \ell(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \ell(\theta) \\ \frac{\partial}{\partial \theta_1} \ell(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_d} \ell(\theta) \end{pmatrix}$$

Proposition 3.9. *The gradient of the log likelihood function* ℓ *is given by*

$$\nabla \ell(\theta) = \sum_{j=1}^{n} \left(y^{(j)} - h_\theta(x^{(j)}) \right) x^{(j)}\,.$$

2 Logistic regression – Gradient ascent

Lemma 3.1. *The sigmoid function satisfies the following differential equation*

$$S'(x) = S(x)(1 - S(x))\,.$$

Proposition 3.2. *The gradient of the log likelihood function ℓ is given by*

$$\nabla \ell(\theta) = \sum_{j=1}^{n} \left(y^{(j)} - h_\theta(x^{(j)}) \right) x^{(j)}\,.$$

The update rule for the gradient ascent is therefore

$$\theta := \theta + \alpha \sum_{j=1}^{n} \left(y^{(j)} - h_\theta(x^{(j)}) \right) x^{(j)}$$

2 Logistic regression

In the colab notebook of Lecture 3 you can see the gradient ascent for the passing exam example.