UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione

Lesson 4.

Logistic regression

DATA SCIENCE AND
AUTOMATION COURSE

MASTER DEGREE SMART
TECHNOLOGY ENGINEERING

TEACHER
Mirko Mazzoleni

PLACE
University of Bergamo

Outline

1. Maximum Likelihood Estimation (MLE)
2. Logistic regression: problem formulation
3. Logistic regression: maximum likelihood estimate
4. Exercise

Outline

1. **Maximum Likelihood Estimation (MLE)**
2. Logistic regression: problem formulation
3. Logistic regression: maximum likelihood estimate
4. Exercise

Maximum likelihood estimation

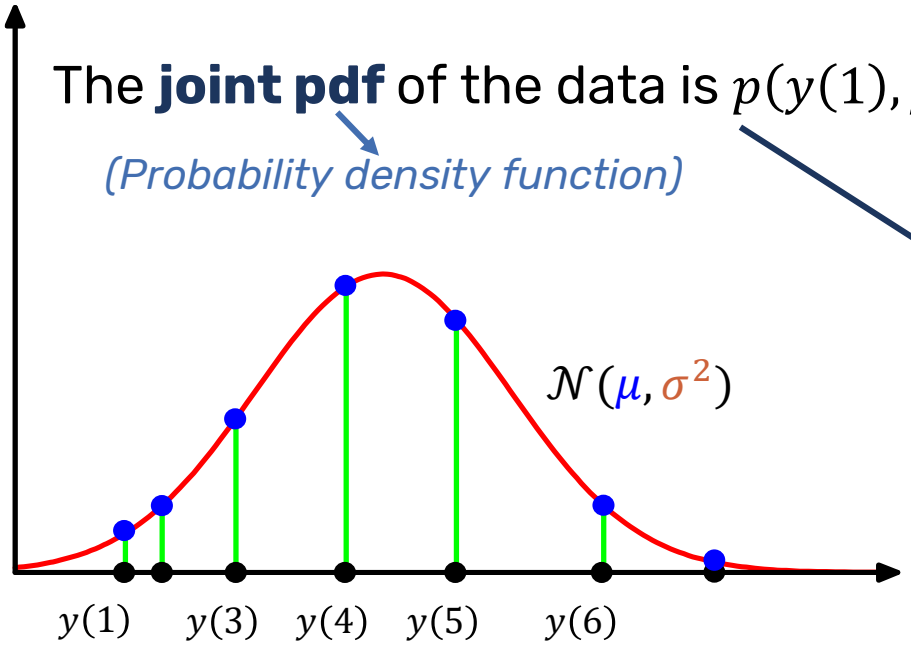
The Maximum Likelihood Estimation (MLE) method is an estimation procedure that, **given a probabilistic model**, estimates its **parameters** in such a way that they are **most consistent** with the observed data

Suppose to have at disposal N observations $Y = [y(1), y(2), \dots, y(N)]^T$, where $y(i) \sim \mathcal{N}(\mu, \sigma^2)$ $i.i.d.$

The **joint pdf** of the data is $p(y(1), y(2), \dots, y(N)|\mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y(i)|\mu, \sigma^2)$

(Probability density function)

$$p(y(i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2\right]$$



When $p(Y|\mu, \sigma^2)$ (the joint pdf) is seen as dependent on parameters μ and σ, it is called **likelihood** $\mathcal{L}(\mu, \sigma^2|Y)$

Maximizing the likelihood means changing the values of the parameters μ and σ^2 s.t. the **product of blue dots is maximized** (see Figure – taken from [8])

Maximum likelihood estimation

In the previous example, $\boldsymbol{\theta} = [\mu, \sigma^2] \in \mathbb{R}^{2\times 1}$

The Maximum Likelihood Estimation (MLE) estimate is therefore $\widehat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|Y)$

It is equivalent to write $\widehat{\boldsymbol{\theta}}_{ML} = \arg\min_{\boldsymbol{\theta}} -\ln[\,\mathcal{L}(\boldsymbol{\theta}|Y)\,]$ → We have a minimization problem as in linear regression, $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

Example

Let $y(i) \sim \mathcal{N}(\mu, \sigma^2 = 1), i = 1,2, i.i.d.$ Estimate the value of μ using MLE, when the observed values are $y(1) = 4, \quad y(2) = 6.$

The pdf of the data is $p(y(i)|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y(i)-\mu}{\sigma}\right)^2\right] = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}(y(i)-\mu)^2\right]$

The density in correspondence to the two observation is:

$$p(y(1) = 4|\mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(4-\mu)^2\right]$$

$$p(y(2) = 6|\mu, \sigma^2 = 1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(6-\mu)^2\right]$$

Maximum likelihood estimation

The joint pdf is the product of the two pdfs:

$$p(y(1), y(2)|\mu, \sigma^2 = 1) = \left(\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(4-\mu)^2\right]\right)\left(\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(6-\mu)^2\right]\right)$$

This expression is now function of μ. With this interpretation, the joint pdf is the **likelihood function** $\mathcal{L}(\mu|Y) = \left(\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(4-\mu)^2\right]\right)\left(\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(6-\mu)^2\right]\right)$

The value of μ that **maximizes** the likelihood, $\hat{\mu}_{ML}$, is taken as estimated value.

It is more convenient to maximize the logarithm of the likelihood. This new function (the **log-likelihood**) has the same maximum of the previous one since the logarithm is a monotonic function

Maximum likelihood estimation

Summarising, the ML estimation has the form:

$$\hat{\mu}_{ML} = \arg\max_{\mu} \ln[\mathcal{L}(\mu | y(1) = 4, y(2) = 6]$$

Let's compute the log-likelihood:

$$\ln(\mathcal{L}) = \ln\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(4-\mu)^2 \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(6-\mu)^2 \right) \right]$$

$$= \ln\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(4-\mu)^2 \right) \right] + \ln\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(6-\mu)^2 \right) \right]$$

$$= \ln\frac{1}{\sqrt{2\pi}} + \ln\left[\exp\left(-\frac{1}{2}(4-\mu)^2 \right) \right] + \ln\frac{1}{\sqrt{2\pi}} + \ln\left[\exp\left(-\frac{1}{2}(6-\mu)^2 \right) \right]$$

$$= 2 \cdot \ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2$$

Maximum likelihood estimation

Maximizing the obtained expression with respect to μ

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \Rightarrow (4-\mu) + (6-\mu) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{4+6}{2} = 5$$

The maximum likelihood estimate of the parameter μ for the defined Gaussian model is the arithmetic mean of the observed data

It is important to notice that maximizing the log-likelihood is equivalent to **minimizing** the negative log-likelihood

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ln[\, \mathcal{L}(\mu|Y) \,]$$

$$= \arg\min_{\theta} -\ln[\, \mathcal{L}(\mu|Y) \,]$$

Outline

1. Maximum Likelihood Estimation (MLE)

2. **Logistic regression: problem formulation**

3. Logistic regression: maximum likelihood estimate

4. Exercise

Logistic regression

Metric data
- Describe a quantity
- An ordering is defined
- A distance is defined

Categorical data
- Describe membership categories
- It is not meaningful to apply an ordering
- It is not meaningful to compute distances

Linear regression modeled **metric data** using a linear model, using regressors (features)

A frequent problem is the modeling of **dichotomic categorical data**, where the output y can assume only two values, e.g. in $\{0, 1\}$

- Predict who of two candidates will get elected
- Predict if a baseball player will hit or not the ball, given its role

In these cases, it is **not meaningful** to use a linear model $y(i) = \boldsymbol{x}(i)^T \boldsymbol{\theta} + \epsilon(i)$

- It is not meaningful to add a continuous error $\epsilon(i) \in \mathbb{R}$ to a variable $y(i)$ that can assume only discrete values
- The model could predict values < 0 or > 1, that are not admissible. There is nothing that "saturates" the output between 0 and 1. ➡ **Logistic function (Sigmoid)**

Logistic regression

Purpose: Estimate the probability that a set of input variables $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$ belong to one of two classes $y \in \{0, 1\}$

- **Linear regression:** $h(\boldsymbol{x}) = a$
- **Logistic regression:** $h(\boldsymbol{x}) = s(a)$

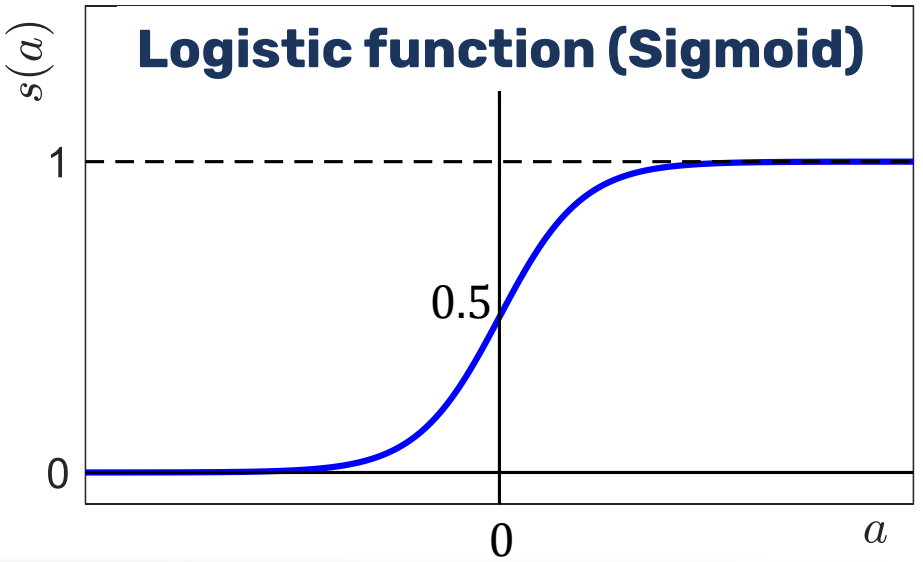
Define the linear combination quantity

$$a = \sum_{i=0}^{d-1} x_i \theta_i = \boldsymbol{x}^T \boldsymbol{\theta}$$

The formula $s(a)$ is the logistic function

$$s(a) = \frac{1}{1 + e^{-a}} = \frac{e^a}{1 + e^a}$$

- $a \gg 0 \Rightarrow s(a) = 1$
- $a \ll 0 \Rightarrow s(a) = 0$



Logistic regression

Purpose: Estimate the probability that a set of input variables $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$ belong to one of two classes $y \in \{0, 1\}$

$$P(y = 1|\boldsymbol{x}) = s(a) = s(\boldsymbol{x}^T\boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{x}^T\boldsymbol{\theta}}}$$

The output of $s(\boldsymbol{x}^T\boldsymbol{\theta})$ is interpreted as a probability

- $\boldsymbol{x}^T\boldsymbol{\theta} \gg 0 \Rightarrow s(\boldsymbol{x}^T\boldsymbol{\theta}) \gg 0.5 \Rightarrow P(y = 1|\boldsymbol{x}) \approx 1$
- $\boldsymbol{x}^T\boldsymbol{\theta} \ll 0 \Rightarrow s(\boldsymbol{x}^T\boldsymbol{\theta}) \ll 0.5 \Rightarrow P(y = 1|\boldsymbol{x}) \approx 0$

Linear regression

$$\mu = \boldsymbol{x}^T\boldsymbol{\theta} = \theta_0 + \theta_1 x_1 + \cdots + \theta_{d-1} x_{d-1}$$

$$y = \mathcal{N}(\mu, \sigma^2)$$

Logistic regression

$$\pi = s(\boldsymbol{x}^T\boldsymbol{\theta}) = s(\theta_0 + \theta_1 x_1 + \cdots + \theta_{d-1} x_{d-1})$$

$$y = \text{Bernoulli}(\pi)$$

Linear and logistic regression are part of a general family of models called **Generalized Linear Models (GLM)**

Outline

1. Maximum Likelihood Estimation (MLE)

2. Logistic regression: problem formulation

3. **Logistic regression: maximum likelihood estimate**

4. Exercise

Logistic regression – MLE

Suppose to have at disposal a dataset $\mathcal{D} = \{(\boldsymbol{x}(1), y(1)), \dots, (\boldsymbol{x}(N), y(N))\}$ where $\boldsymbol{x} \in \mathbb{R}^{d \times 1}$ and $y(i) \in \{0, 1\}, i = 1, \dots, N,\ i.i.d.$

Estimate a logistic regression model $P(y = 1|\boldsymbol{x}) = \frac{1}{1+e^{-x^T\boldsymbol{\theta}}} \equiv \pi$

Interpret the data as $y \sim \text{Bernoulli}(\pi)$

Computation of the likelihood

$$\mathcal{L}(\pi|Y) = \prod_{i=1}^{N} \pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)} \Rightarrow$$ Compute minus log-likelihood $\Rightarrow$

$$\Rightarrow \quad -\ln[\mathcal{L}(\pi|Y)] = -\ln\left[\prod_{i=1}^{N} \pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)}\right]$$

Logistic regression – MLE

$$-\ln[\mathcal{L}(\pi|Y)] = -\ln\left[\prod_{i=1}^{N} \pi(i)^{y(i)} \cdot \big(1-\pi(i)\big)^{1-y(i)}\right] = -\sum_{i=1}^{N} \ln\left[\pi(i)^{y(i)} \cdot \big(1-\pi(i)\big)^{1-y(i)}\right]$$

$$= -\sum_{i=1}^{N} \left(\ln\left[\pi(i)^{y(i)}\right] + \ln\left[\big(1-\pi(i)\big)^{1-y(i)}\right]\right)$$

$$= -\sum_{i=1}^{N} \big(y(i) \cdot \ln \pi(i) + \big(1-y(i)\big) \cdot \ln[1-\pi(i)] \big) \equiv J(\boldsymbol{\theta})$$

Logistic regression – MLE

Cost function interpretation

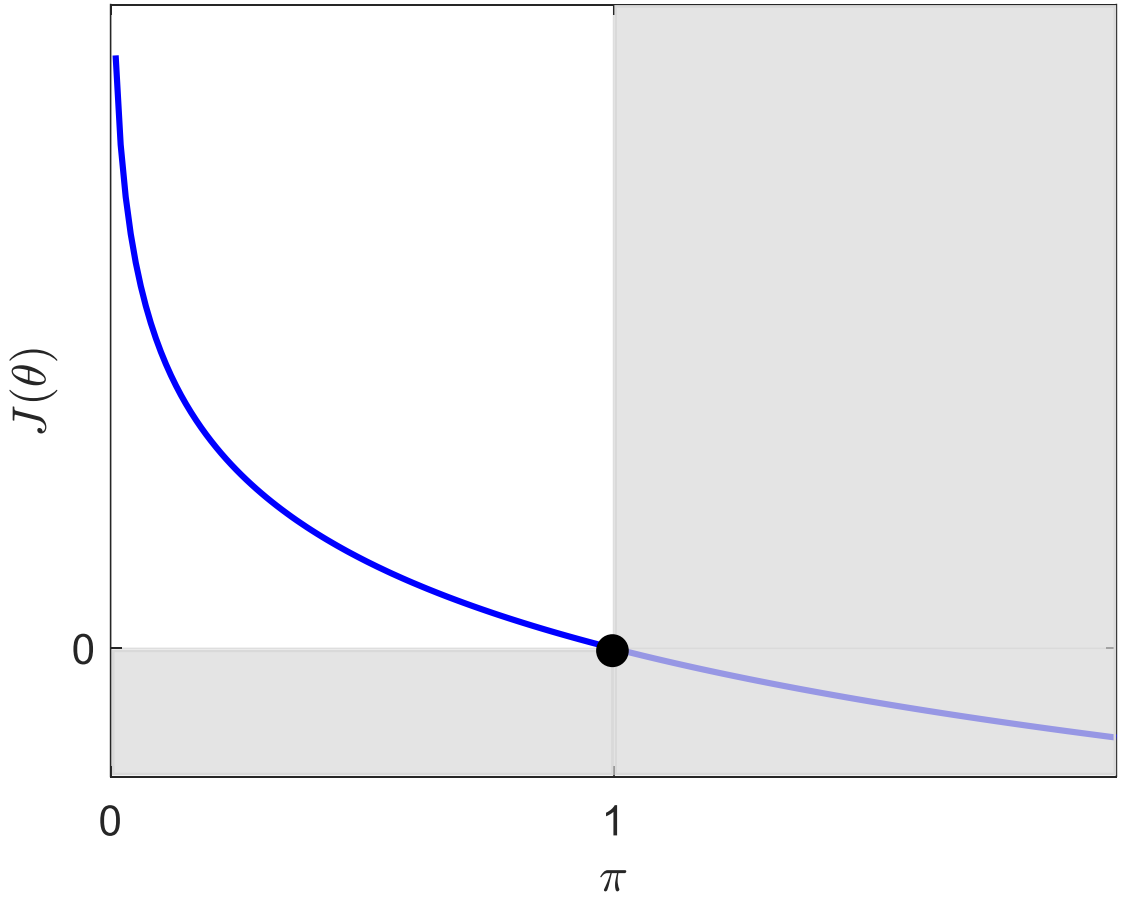
Suppose there is only one datum $\mathcal{D} = \{(\boldsymbol{x}, y)\}$

$$\Rightarrow J(\boldsymbol{\theta}) = \begin{cases} -\ln \pi & \textit{if} \quad y = 1 \\ -\ln[1-\pi] & \textit{if} \quad y = 0 \end{cases}$$

Case $y = 1$

$$J(\boldsymbol{\theta}) = -\ln \pi$$

- $J(\boldsymbol{\theta}) \approx 0$ if $y = 1$ and $\pi \approx 1$
- $J(\boldsymbol{\theta}) \approx +\infty$ if $y = 1$ and $\pi \approx 0$



Logistic regression – MLE

Cost function interpretation

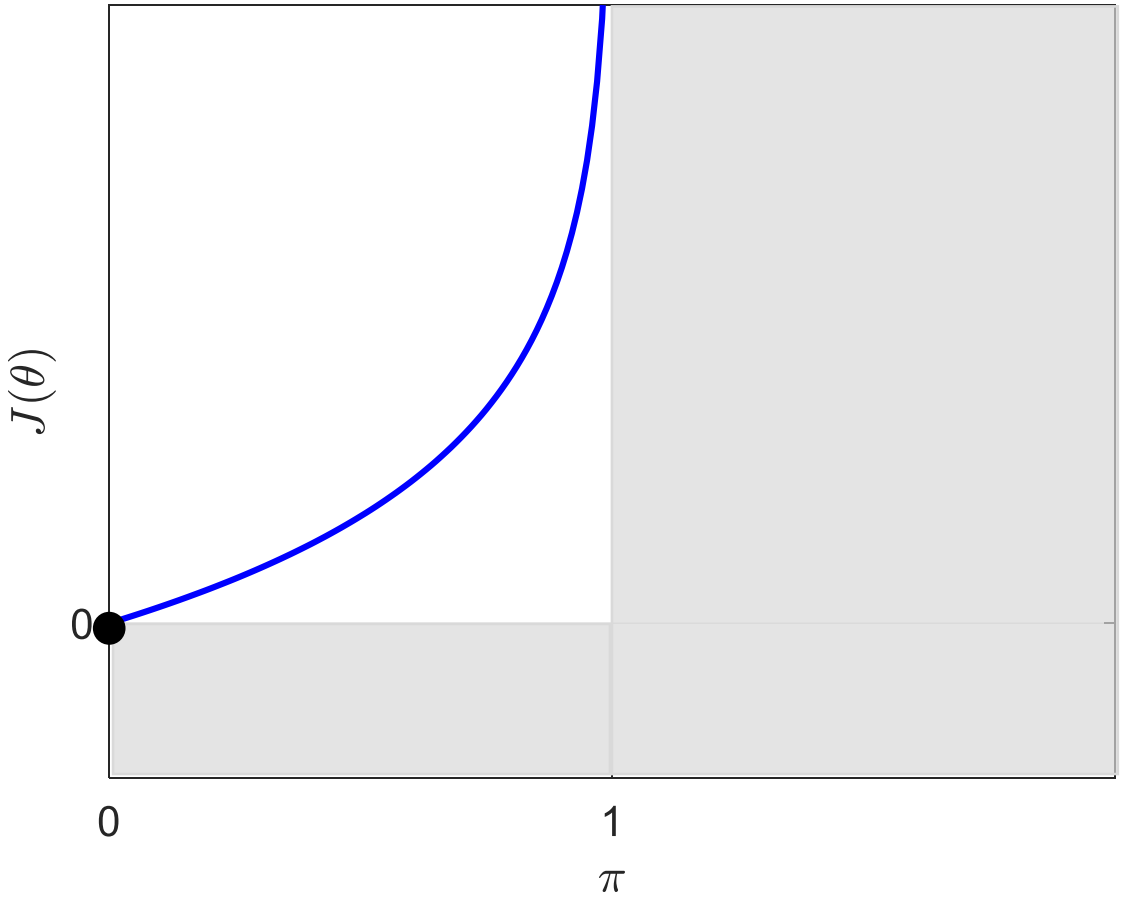
Suppose there is only one datum $\mathcal{D} = \{(\boldsymbol{x}, y)\}$

$$\Rightarrow J(\boldsymbol{\theta}) = \begin{cases} -\ln \pi & \text{if } \; y = 1 \\ -\ln[1-\pi] & \text{if } \; y = 0 \end{cases}$$

Case $y = 0$

$$J(\boldsymbol{\theta}) = -\ln[1-\pi]$$

- $J(\boldsymbol{\theta}) \approx 0$ if $y = 0$ and $\pi \approx 0$
- $J(\boldsymbol{\theta}) \approx +\infty$ if $y = 0$ and $\pi \approx 1$



IN-DEPTH ANALYSIS

Computation of the minimum of $J(\boldsymbol{\theta})$

We have to compute the gradient of $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$. First, compute the derivative of $s(a) = \frac{1}{1+e^{-a}}$

$$\frac{\partial s(a)}{\partial a} = \frac{\partial}{\partial a}\left[\frac{1}{1+e^{-a}}\right] = \frac{\partial}{\partial a}[(1+e^{-a})^{-1}] = -(1+e^{-a})^{-2}(e^{-a})(-1) = -(1+e^{-a})^{-2}(-e^{-a}) = \frac{e^{-a}}{(1+e^{-a})^2}$$

$$= \frac{1}{(1+e^{-a})}\cdot\frac{e^{-a}}{(1+e^{-a})} = \frac{1}{(1+e^{-a})}\cdot\frac{(1+e^{-a})-1}{1+e^{-a}} = \frac{1}{1+e^{-a}}\cdot\left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right) = \boxed{s(a)\cdot[1-s(a)]}$$

In the case where $a = \boldsymbol{x}^T\boldsymbol{\theta}$, we have that

$$\underset{d\times 1}{\nabla_{\boldsymbol{\theta}} s(\boldsymbol{x}^T\boldsymbol{\theta})} = \underset{d\times 1}{\boldsymbol{x}}\cdot \underset{1\times 1}{s(\boldsymbol{x}^T\boldsymbol{\theta})}\cdot[1-\underset{1\times 1}{s(\boldsymbol{x}^T\boldsymbol{\theta})}] = \boxed{\boldsymbol{x}\cdot\pi\cdot[1-\pi]}$$

IN-DEPTH ANALYSIS

Computation of the minimum of $J(\theta)$

We can now compute the gradient of $J(\boldsymbol{\theta})$

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\big(y(i)\ln\pi(i) + \big(1-y(i)\big)\ln[1-\pi(i)]\big) \qquad \pi(i) = \frac{1}{1+e^{-\boldsymbol{x}(i)^T\boldsymbol{\theta}}}$$

$$\underset{d\times 1}{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})} = -\sum_{i=1}^{N}\left(y(i)\frac{\pi'(i)}{\pi(i)} + \big(1-y(i)\big)\frac{-\pi'(i)}{1-\pi(i)}\right) = -\sum_{i=1}^{N}\left(y(i)\frac{\boldsymbol{x}(i)\pi(i)[1-\pi(i)]}{\pi(i)} + \big(1-y(i)\big)\frac{-\boldsymbol{x}(i)\pi(i)[1-\pi(i)]}{1-\pi(i)}\right)$$

$$= \sum_{i=1}^{N}\big(-y(i)\boldsymbol{x}(i)[1-\pi(i)] - \big(1-y(i)\big)\big(-\boldsymbol{x}(i)\pi(i)\big)\big) = \sum_{i=1}^{N}\big(\boldsymbol{x}(i)\cdot[-y(i)+y(i)\pi(i)] + \boldsymbol{x}(i)\cdot[\pi(i)-y(i)\pi(i)]\big)$$

$$= \sum_{i=1}^{N}\big(\boldsymbol{x}(i)\cdot[-y(i)+y(i)\pi(i)-y(i)\pi(i)+\pi(i)]\big) \boxed{= \sum_{i=1}^{N}\underset{d\times 1}{\boldsymbol{x}(i)}\cdot\underset{1\times 1}{\big(\pi(i)-y(i)\big)}}$$

Gradient descent

It can be shown that:

- The cost function $J(\boldsymbol{\theta})$ is **convex** and admits a **unique minimum**
- The equations found by posing $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbf{0}$ are nonlinear in $\boldsymbol{\theta}$ and it **is not possible** to find a solution in **closed-form**
  - ✓ For this reason, we need to resort to **iterative optimization algorithms**

Use **gradient descent:**

$$\underset{d\times 1}{\widehat{\boldsymbol{\theta}}(t+1)} = \underset{d\times 1}{\widehat{\boldsymbol{\theta}}(t)} - \underset{1\times 1}{\alpha} \cdot \underset{d\times 1}{\nabla J(\boldsymbol{\theta})}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(t)}$$

$\alpha \in \mathbb{R}_{>0}$: learning rate

Logistic regression recap

- The logistic regression model, despite its name, is not used for regression, but for **classification**
- Once the model predicts the probability of a class, we can classify a point to a particular class if the probability for that class is **above a threshold** (usually 0.5)
- The function that now we are trying to predict is: $f(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$
- The logistic regression tries to model f by using the hypothesis:

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

- The point $\boldsymbol{x}$ can then be classified to class $y = +1$ if $h(\boldsymbol{x}) \geq 0.5$

Logistic regression recap

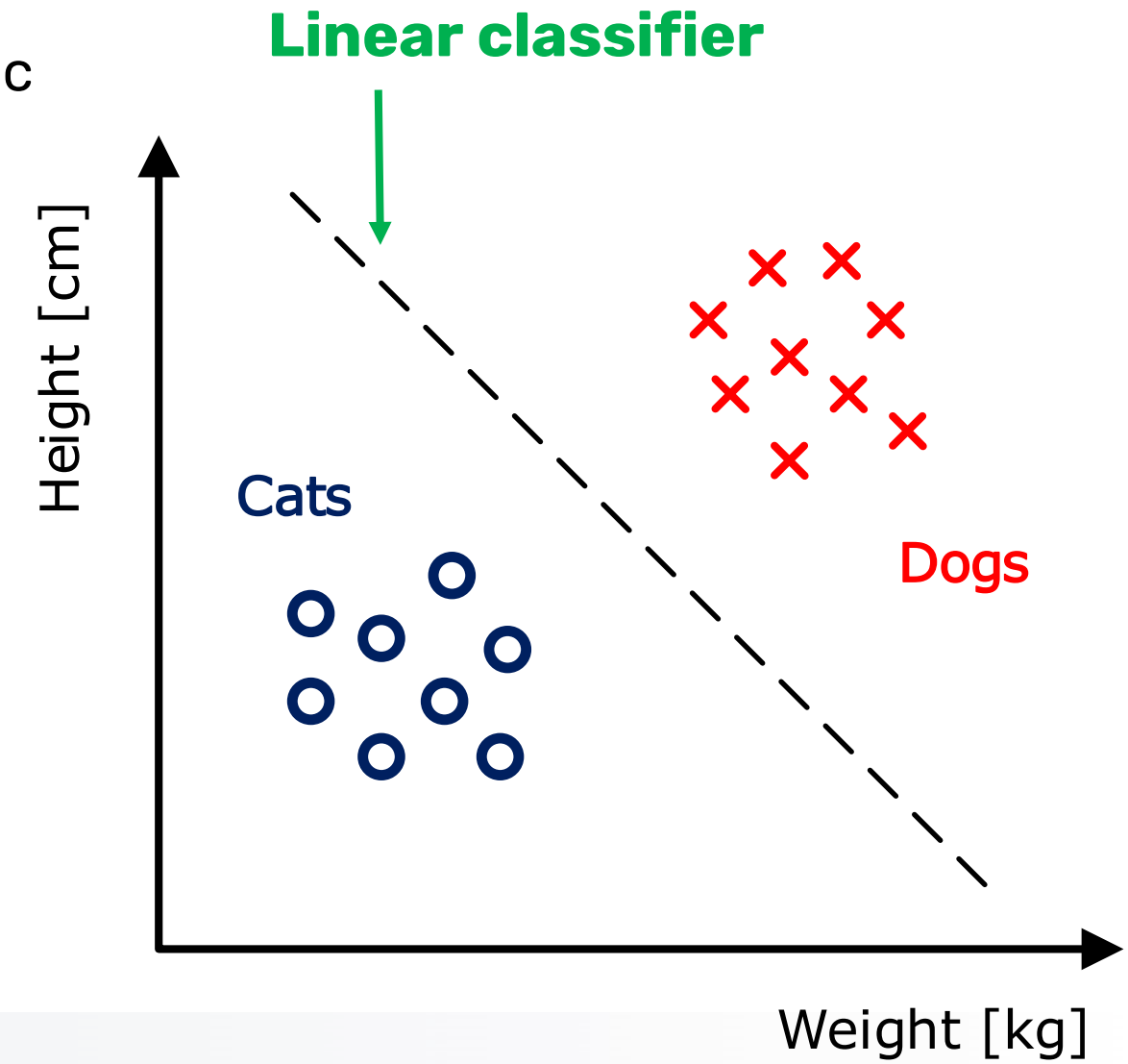
The classification boundary found by logistic regression is **linear**

Infact, classifying with the rule:

$$y = 1 \text{ if } h(\boldsymbol{x}) \geq 0.5$$

is the same as saying

$$y = 1 \text{ if } \boldsymbol{x}^T \boldsymbol{\theta} \geq 0$$



Outline

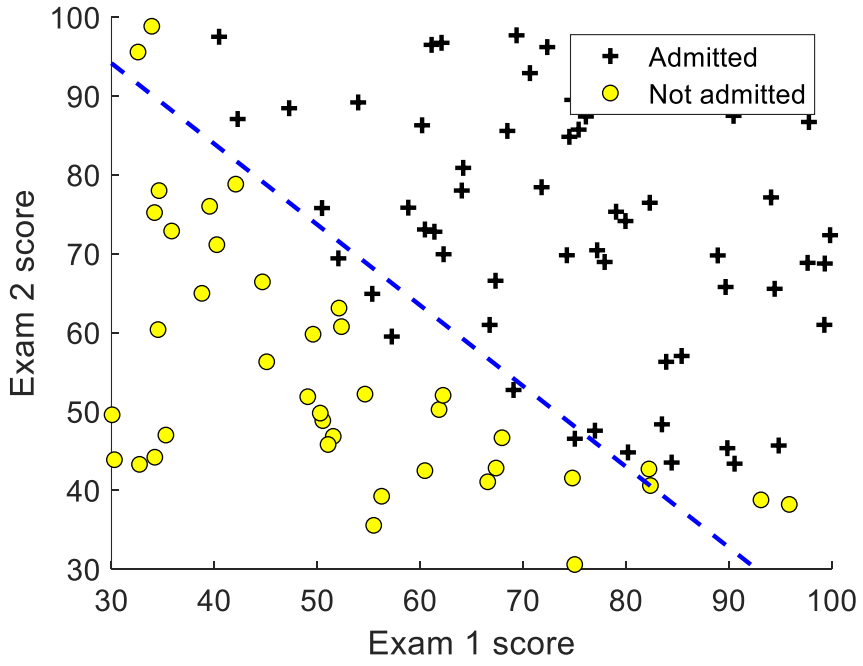
1. Maximum Likelihood Estimation (MLE)
2. Logistic regression: problem formulation
3. Logistic regression: maximum likelihood estimate
4. **Exercise**

Logistic regression laboratory: students admissions

We want to **predict if a student will get admitted** to a university given the results on two exams (x_1, x_2)

- The **training set** consists of $N = 100$ students with $x_1(i), x_2(i)$ and $y(i) \in \{0,1\}$, for $i = 1, \dots, N$

- $X \in \mathbb{R}^{100\times 3}$
- $Y \in \mathbb{R}^{100\times 1}$
- $\boldsymbol{\theta} \in \mathbb{R}^{3\times 1}$




```
% Read data from file
data = load('studentsdata.csv');
X = data(:, [1, 2]); y = data(:, 3);

% Setup the data matrix appropriately, and
add ones for the intercept term
[N, m] = size(X); d = m + 1;

% Add intercept term
X = [ones(N, 1) X];

% Initialize fitting parameters
initial_theta = zeros(d, 1);
pi_s = sigmoid(X*theta)

J = ( -y'*log(pi_s) - (1-y)'*log(1-pi_s));
grad = X'*( pi_s - y);
```

Embed in a function and pass the function to an optimization algoritm that iteratively computes the gradient

References

1. Provost, Foster, and Tom Fawcett. "*Data Science for Business: What you need to know about data mining and data-analytic thinking*". O'Reilly Media, Inc., 2013.
2. Brynjolfsson, E., Hitt, L. M., and Kim, H. H. *"Strength in numbers: How does data driven decision making affect firm performance?"* Tech. rep., available at SSRN: http://ssrn.com/abstract=1819486, 2011.
3. Pyle, D. "*Data Preparation for Data Mining*". Morgan Kaufmann, 1999.
4. Kohavi, R., and Longbotham, R. *"Online experiments: Lessons learned".* Computer, 40 (9), 103–105, 2007.
5. Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. *"Learning from data"*. AMLBook, 2012.
6. Andrew Ng. *"Machine learning"*. Coursera MOOC. (https://www.coursera.org/learn/machine-learning)
7. Domingos, Pedro. "*The Master Algorithm".* Penguin Books, 2016.
8. Christopher M. Bishop, "*Pattern recognition and machine learning", Springer-Verlag New York, 2006.*