Machine Learning and Intelligent Systems

Linear Models for Regression

Maria A. Zuluaga

October 13, 2023

EURECOM - Data Science Department

Table of contents

Intro & Recap

Recap: Definition

- We have $y \in \mathcal{C}$ and $\mathbf{x} \in \mathbb{R}^D$
- They are related by an unknown function $f : \mathbb{R}^D \longrightarrow \mathcal{C}$

Recap: Definition

- We have $y \in \mathcal{C}$ and $\mathbf{x} \in \mathbb{R}^D$
- They are related by an unknown function $f : \mathbb{R}^D \longrightarrow \mathcal{C}$

y	$\mathbf{x}$
Output	Input
Target	Feature vector
Label	Attributes
Dependent variable	Independent variable

Recap: Definition

- We have $y \in \mathcal{C}$ and $\mathbf{x} \in \mathbb{R}^D$
- They are related by an unknown function $f : \mathbb{R}^D \longrightarrow \mathcal{C}$

y	$\mathbf{x}$
Output	Input
Target	Feature vector
Label	Attributes
Dependent variable	Independent variable

Goal

To predict y using $\mathbf{x}$ **but** we don't know the true relationship, f, between y and $\mathbf{x}$

Recap: Definition

- We have $y \in \mathcal{C}$ and $\mathbf{x} \in \mathbb{R}^D$
- They are related by an unknown function $f : \mathbb{R}^D \longrightarrow \mathcal{C}$

y	$\mathbf{x}$
Output	Input
Target	Feature vector
Label	Attributes
Dependent variable	Independent variable

Goal

To predict y using $\mathbf{x}$ **but** we don't know the true relationship, f, between y and $\mathbf{x}$

Question: I am using f instead of h, why?

Recap: Data

- To discover the relationship between $\mathbf{x}$ and y, we have access to data.
- It consists of a set of N inputs

$$\{\mathbf{x}_i\} \quad i = 1, \ldots, N,$$

and the corresponding set of outputs

$$\{y_i\}$$

Recap: Data

- To discover the relationship between $\mathbf{x}$ and y, we have access to data.
- It consists of a set of N inputs

$$\{\mathbf{x}_i\} \quad i = 1, \dots, N,$$

  and the corresponding set of outputs

$$\{y_i\}$$

- The paired inputs-outputs set

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$$

  is denoted the training set.

Recap: Notation

Symbol	Reads as
X	Input variable ($\mathbb{R}^D$)
$\mathbf{x}_i$	i^{th} feature vector. Observed value of X.
$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T$	Matrix of N input $D-$dimensional vectors $\mathbf{x}_i$
x_j	j^{th} element of the i^{th} input vector $\mathbf{x}_i$, i.e. x_i^j
Y	Output variable ($\mathcal{C}$)
y_i	i^{th} output label
$\mathbf{y} = (y_1, \ldots, y_N)^T$	Observed vector of outputs y_i
$\mathbf{x}^*$	Test point (unseen data)
$\hat{y}$	Prediction for $\mathbf{x}^*$

Table 1: Different notation for the input and output variables

Note

For regression, we deal with $y \in \mathcal{C} = \mathbb{R}^{O=1}$

Recap: Hypothesis class

- The goal of supervised learning is to use $\mathcal{D}$ to learn a function $h : \mathbb{R}^D \longleftrightarrow \mathcal{C}, h \in \mathcal{H}$ that can predict y from $\mathbf{x}$.
- The first hypothesis class we studied was nearest neighbors

Recap: Hypothesis class

- The goal of supervised learning is to use $\mathcal{D}$ to learn a function $h : \mathbb{R}^D \longleftrightarrow \mathcal{C}, h \in \mathcal{H}$ that can predict y from $\mathbf{x}$.
- The first hypothesis class we studied was nearest neighbors

In this lecture: Second family - Linear Models for Regression

Linear Models for Regression

No Free Lunch: Assumptions

Data Assumptions:

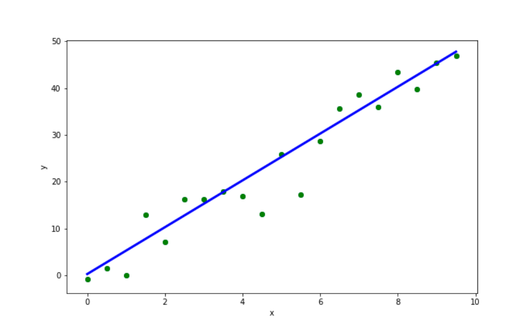
No Free Lunch: Assumptions

Data Assumptions:

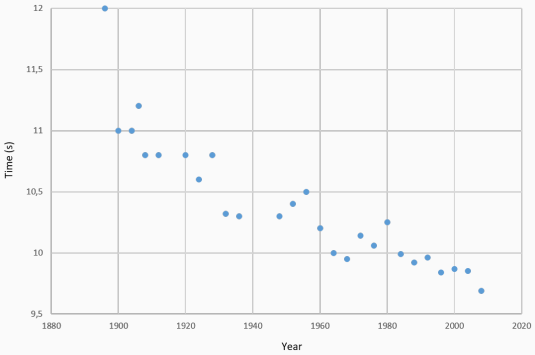
$$y \in \mathbb{R}$$

Model Assumptions:

$$y = f(\mathbf{x})$$
$$y_i = \mathbf{w}^T \mathbf{x}_i$$



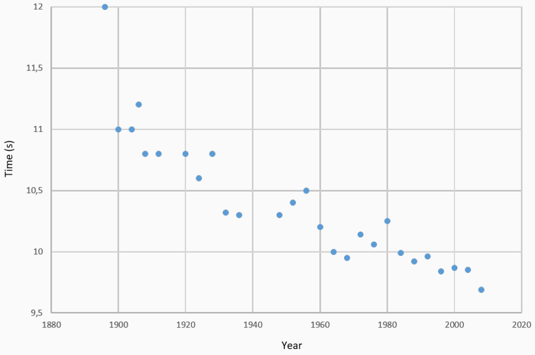
An Example: 100m at the Olympics



Winning times of men's 100m at the Olympics 1896 - 2012

Can we use this information to predict the times of Rio 2016 and Japan 2020?

100m at the Olympics: The Data

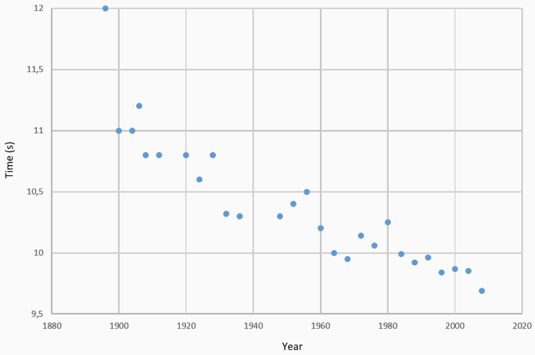


Winning times of men's 100m at the Olympics 1896 - 2012

	Year	Time
1	1896	12.00
2	1900	11.00
3	1904	11.00
4	1908	10.80
5	1912	10.80
6	1920	10.80
7	1924	10.60
8	1928	10.80
9	1932	10.32
10	1936	10.30
11	1948	10.30
12	1952	10.40
13	1956	10.50
14	1960	10.20
15	1964	10.00
16	1968	9.95
17	1972	10.14
18	1976	10.06
19	1980	10.25
20	1984	9.99
21	1988	9.92
22	1992	9.96
23	1996	9.84
24	2000	9.87
25	2004	9.85
26	2008	9.69
27	2012	9.63

- y -
- $\mathbf{x}$ -
- D -
- N -
- $(\mathbf{x}_3, y_3)$ -
- $\hat{y}$ -

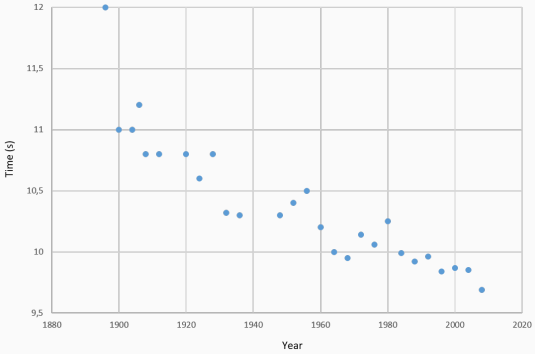
100m at the Olympics: The Hypothesis Class



Winning times of men's 100m at the Olympics 1896 - 2008

Do we have any prior knowledge about $f(\mathbf{x})$?

100m at the Olympics: The Hypothesis Class



Winning times of men's 100m at the Olympics
1896 - 2008

Do we have any prior knowledge about $f(\mathbf{x})$?
Which are our assumptions?

Wrap-up

- We have identified the data

Wrap-up

- We have identified the data
- We assumed that there is an unknown function f that maps the Olympic year ($\mathbf{x}$) to the men's Olympic 100m winning time (y)

Wrap-up

- We have identified the data
- We assumed that there is an unknown function f that maps the Olympic year ($\mathbf{x}$) to the men's Olympic 100m winning time (y)
- Some other assumptions:
  - $y \in \mathbb{R}$ and $y > 0$
  - f is a decreasing function: $f(\mathbf{x}_i) \geq f(\mathbf{x}_{i+1})$
  - $\mathbf{x}$, y have a linear relationship

Wrap-up

- We have identified the data
- We assumed that there is an unknown function f that maps the Olympic year ($\mathbf{x}$) to the men's Olympic 100m winning time (y)
- Some other assumptions:
  - $y \in \mathbb{R}$ and $y > 0$
  - f is a decreasing function: $f(\mathbf{x}_i) \geq f(\mathbf{x}_{i+1})$
  - $\mathbf{x}, y$ have a linear relationship

It seems that linear regression is a good choice to find an appropriate $h(\mathbf{x})$

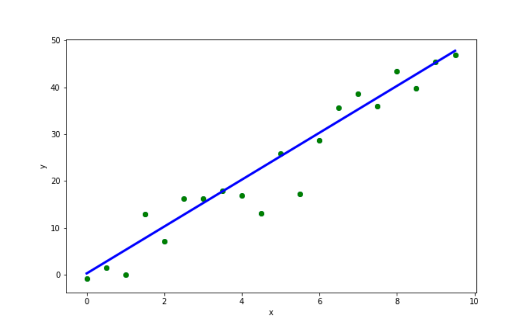
Back to assumptions

Data Assumptions:

$$y \in \mathbb{R}$$

Model Assumptions:

$$y = f(\mathbf{x})$$
$$y_i = \mathbf{w}^T \mathbf{x}_i$$



How realistic are these assumptions?

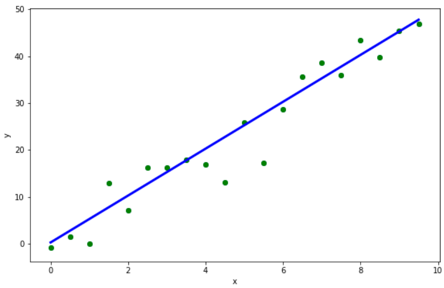
Back to assumptions

Data Assumptions:

$$y \in \mathbb{R}$$

Model Assumptions:

$$y = f(\mathbf{x})$$
$$y_i = \mathbf{w}^T \mathbf{x}_i$$



How realistic are these assumptions?

They do not consider that observations are noisy

Observation Errors: Noise

- Inherent errors in the measurement tools
- Missing variables

Errors

A more accurate model assumption should account for observation errors (or noise).
This is the additive error model

$$y = h(\mathbf{x}) = f(\mathbf{x}) + \epsilon$$

Errors

A more accurate model assumption should account for observation errors (or noise).
This is the additive error model

$$y = h(\mathbf{x}) = f(\mathbf{x}) + \epsilon$$

Error Assumptions:

- The error can be positive or negative
- It is **independent** for each $\mathbf{x}_i$:
  - It may be different for every input sample $\mathbf{x}_i$
  - Holds no relationship between errors along $\mathbf{X}$
- It cannot be modeled exactly
- It can be modeled as a **continuous random variable**

Parenthesis: Probability Refresher

Parenthesis: Random Variables

Deterministic: Fixed outcome that can be estimated exactly

- Kelvin = Celsius + 273.15
- Amount of money in your bank account next month
- Odds of obtaining a five when rolling a dice once

Parenthesis: Random Variables

Deterministic: Fixed outcome that can be estimated exactly

- Kelvin = Celsius + 273.15
- Amount of money in your bank account next month
- Odds of obtaining a five when rolling a dice once

Random: Variable whose value is determined by chance

- Tomorrows temperature
- Price of Amazons stock next month
- The roll of a dice is a prime number

Parenthesis: Random Variables

Deterministic: Fixed outcome that can be estimated exactly

- Kelvin = Celsius + 273.15
- Amount of money in your bank account next month
- Odds of obtaining a five when rolling a dice once

Random: Variable whose value is determined by chance

- Tomorrows temperature
- Price of Amazons stock next month
- The roll of a dice is a prime number

More formally:
A random variable is a measurable function $X : \Omega \rightarrow E$ from a set of possible outcomes Ω to a measurable space E.

Probability Density Functions

- The probability density function (PDF), or density of a continuous random variable, is a function that describes the relative **likelihood** for this random variable to take on a given value.

Probability Density Functions

- The probability density function (PDF), or density of a continuous random variable, is a function that describes the relative **likelihood** for this random variable to take on a given value.
- The probability of the random variable to fall within a particular region is given by the integral of this variable's PDF over the region:

$$P(a \leq X \leq b) = \int_a^b p(X)dX$$

The Gaussian Distribution

- The Gaussian (or Normal) distribution is a PDF for continuous random variables.

The Gaussian Distribution

- The Gaussian (or Normal) distribution is a PDF for continuous random variables.
- Its general form is:

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

where:

- z is a continuous random variable
- μ is its mean (of z)
- σ^2 is its variance

The Gaussian Distribution

- The Gaussian (or Normal) distribution is a PDF for continuous random variables.
- Its general form is:

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

where:

  - z is a continuous random variable
  - μ is its mean (of z)
  - σ^2 is its variance

- It is often referred to as $\mathcal{N}(\mu, \sigma^2)$

The Multivariate Gaussian Distribution

- Let us now denote $\mathbf{z} = (z_1, \ldots, z_D)^T$, the multivariate Gaussian or joint normal distribution of $\mathbf{z}$ is denoted by:

The Multivariate Gaussian Distribution

- Let us now denote $\mathbf{z} = (z_1, \ldots, z_D)^T$, the multivariate Gaussian or joint normal distribution of $\mathbf{z}$ is denoted by:

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\}$$

  where:

  - $\boldsymbol{\mu} \in \mathbb{R}^D$ is its mean (of z)
  - $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ is the covariance matrix
  - $|\cdot|$ is the determinant

The Multivariate Gaussian Distribution

- Let us now denote $\mathbf{z} = (z_1, \ldots, z_D)^T$, the multivariate Gaussian or joint normal distribution of $\mathbf{z}$ is denoted by:

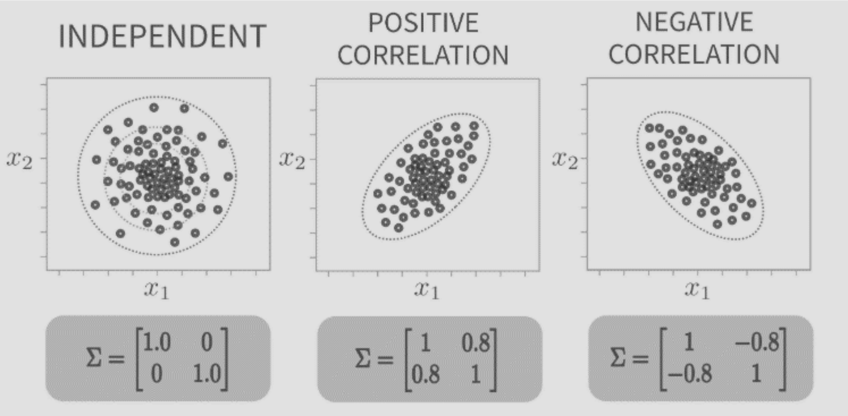
$$p(\mathbf{z}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})\right\}$$

  where:

  - $\boldsymbol{\mu} \in \mathbb{R}^D$ is its mean (of z)
  - $\mathbf{\Sigma} \in \mathbb{R}^{D \times D}$ is the covariance matrix
  - $|\cdot|$ is the determinant

- It is denoted $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$

The Covariance Matrix



Source: `http://complx.me/2017-01-22-mle-linear-regression/`

Linear Models for Regression (back)

Back to Assumptions

Data Assumptions:

$$y_i \in \mathbb{R}$$

Back to Assumptions

Data Assumptions:

$$y_i \in \mathbb{R}$$

Model Assumptions:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \boldsymbol{\sigma}^2)$.

Back to Assumptions

Data Assumptions:

$$y_i \in \mathbb{R}$$

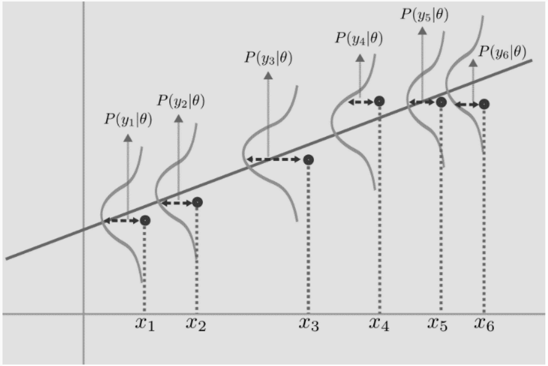
Model Assumptions:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

This is equivalent to:

$$y_i | \mathbf{x}_i \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_i, \sigma^2)$$



Source: `http://complx.me/2017-01-22-mle-linear-regression/`

Recap: Learning

- The task of learning consists in finding a hypothesis $h : \mathbb{R}^D \to \mathcal{C}$ that can make a good prediction of y, given $\mathbf{x}$

Recap: Learning

- The task of learning consists in finding a hypothesis $h : \mathbb{R}^D \rightarrow \mathcal{C}$ that can make a good prediction of y, given $\mathbf{x}$

$$\hat{y} = h(\mathbf{x})$$

- $\hat{y}$ denotes the predicted value of y

Recap: Learning

- The task of learning consists in finding a hypothesis $h : \mathbb{R}^D \to \mathcal{C}$ that can make a good prediction of y, given $\mathbf{x}$
$$\hat{y} = h(\mathbf{x})$$
- $\hat{y}$ denotes the predicted value of y
- **Final goal:** $\hat{y} \approx y$ for unseen $(\mathbf{x}, y)$ pairs

Recap: Learning

- The task of learning consists in finding a hypothesis $h : \mathbb{R}^D \to \mathcal{C}$ that can make a good prediction of y, given $\mathbf{x}$
$$\hat{y} = h(\mathbf{x})$$
- $\hat{y}$ denotes the predicted value of y
- **Final goal:** $\hat{y} \approx y$ for unseen $(\mathbf{x}, y)$ pairs
- To achieve this make us of the data and of any prior knowledge we might have about f.
- Example:
  - $y \geq 0$
  - Continuity and smoothness of the function
  - Linear relationship

Learning

$$p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}$$

Learning

$$p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}$$

- $\mathbf{w} = \{w_j\}$, σ^2 are the **parameters** of the model

Learning

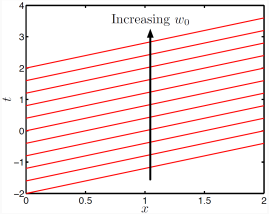
$$p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}$$

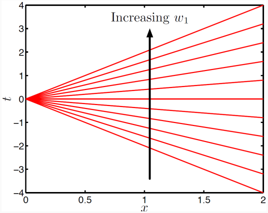
- $\mathbf{w} = \{w_j\}$, σ^2 are the **parameters** of the model
- Question: What were the parameters in k-NN?

A deeper look into w

Example: 100m Olympics (let's ignore σ^2 for a bit)

$$y = w_0 + w_1 x_1$$

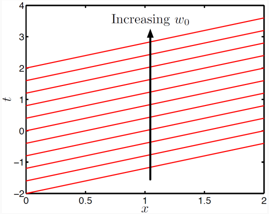


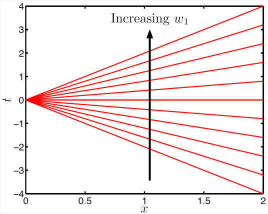


A deeper look into w

Example: 100m Olympics (let's ignore σ^2 for a bit)

$$y = w_0 + w_1 x_1$$





Important

The term "linear" comes from the fact that y is linear w.r.t the parameters $\mathbf{w}$

Learning or Model Fitting

- Model fitting is the process of finding a good estimate of $h(\mathbf{x})$

Learning or Model Fitting

- Model fitting is the process of finding a good estimate of $h(\mathbf{x})$
- It accounts to fitting the training data $\mathcal{D}$ into the linear model to estimate the **model's parameters**:

$$p(y_i|\mathbf{x}_i;\mathbf{w},\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}$$

Learning or Model Fitting

- Model fitting is the process of finding a good estimate of $h(\mathbf{x})$
- It accounts to fitting the training data $\mathcal{D}$ into the linear model to estimate the **model's parameters**:

$$p(y_i|\mathbf{x}_i;\mathbf{w},\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}$$

- **In words:** For each data point with features $\mathbf{x}_i$, the label y is drawn from a Gaussian with mean $\mathbf{w}^T\mathbf{x}_i$ and variance σ^2. Our task is to estimate the slope w_1 and intercept w_0 from the data by using the fact that $\hat{y} \approx y$

Wrap-up

- We introduced the basic terminology used in supervised learning
- We introduce linear regression models
- We introduced the concept of model parameters
- Next: How to obtain these parameters

References

Further Reading and Useful Material

Source	Notes
The Elements of Statistical Learning	Ch. 2 and 3
Pattern Recognition and Machine Learning	Sec 1.5.5, Ch. 3