

# COMS30035, Machine learning:

From regression to classification  
and neural networks:

Revising regression

(based on slides by Dima Damen)

Rui Ponte Costa

Department of Computer Science, SCEEM  
University of Bristol

October 6, 2021

# Textbooks

Chapter 3 of the Bishop book is directly relevant:

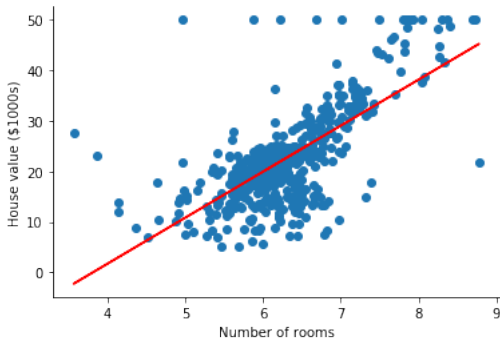
- ▶ Bishop, C. M., Pattern recognition and machine learning (2006). Available for free [here](#).
- ▶ **Note:** this first part is a revision of should be covered in Data Science for CS (on your 2nd year); old full lecture notes [here](#).

# Agenda

- ▶ **Revising linear and nonlinear regression** [see old SPS slides; Chapter 3, Bishop]
  - ▶ Linear regression
  - ▶ Nonlinear regression
  - ▶ Probabilistic models
  - ▶ Maximum likelihood estimation
- ▶ **Sequential Bayesian regression** [Chapter 3, Bishop]
  - ▶ Bayesian formulation
  - ▶ Maximum a posteriori
  - ▶ Example
- ▶ **Classification and neural networks** [Chapter 5, Bishop]
  - ▶ Architectures (Parametric model)
  - ▶ The supervised case
  - ▶ Optimising nnets using backprop
  - ▶ Highly flexible model → overfitting: early stopping/drop-out.

# Revisiting regression

- ▶ Goal: Finding a relationship between two variables (e.g. regress *house value* against *number of rooms*)
- ▶ Model: Linear relationship between *house value* and *number of rooms*?



## Revisiting regression – deterministic model

**Data:** a set of data points  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$  where  $x_i$  is the house value  $i$  and  $y_i$  is the number of rooms  $i$ .

**Task:** build a model that can predict the house value from the number of rooms

**Model Type:** parametric; assumes a polynomial relationship between house value and number of rooms

**Model Complexity:** assume the relationship is linear  
house value =  $a_0 + a_1 * \text{rooms}$

$$y_i = a_0 + a_1 x_i \quad (1)$$

**Model Parameters:** model has two parameters  $a_0$  and  $a_1$  which should be estimated.

- ▶  $a_0$  is the y-intercept
- ▶  $a_1$  is the slope of the line

# Least Squares Solution - matrix form

- ▶ To find a solution to the parameters  $\theta = \{a_0, a_1\}$  solve least squares problem which **in matrix form**, means to find  $\mathbf{a}_{LS}$ ; <sup>1</sup>

$$\|\mathbf{y} - \mathbf{X} \mathbf{a}_{LS}\|^2 = 0 \quad (2)$$

$$\mathbf{a}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (3)$$

- ▶ Matrix formulation also allows least squares method to be extended to **polynomial fitting**
- ▶ For a polynomial of degree  $p + 1$  we use (note:  $p > 1$  gives nonlinear regression)

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_p x_i^p$$

---

<sup>1</sup>  $\|\mathbf{A}\|^2 = \sqrt{\sum \sum |a_{ij}|^2}$  denotes the Frobenius norm, defined as the square root of the sum of the absolute squares of its elements. For a detailed derivation see [this derivation - p8](#)

# Least Squares Solution

## Example

Find the best least squares fit by a linear function to the data using  $p = 1$

x	-1	0	1	2
y	0	1	3	9

$$\mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 3 \\ 9 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad \mathbf{X}^T \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 6 \end{bmatrix}$$

$$\mathbf{a}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{1}{20} \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 3 \\ 9 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 2.9 \end{bmatrix}$$

$$y = 1.8 + 2.9x$$

# Regression with probabilistic models

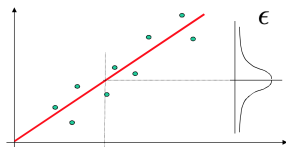
**Probabilistic models are a core part of ML**, as they allow us to also capture the uncertainty the model has about the data, which is critical for real world applications. For simplicity, let's drop  $a_0$  from the previous model and add a random variable  $\epsilon$  that captures the uncertainty

$$\text{house price} = a_1 \times \text{number of rooms} + \epsilon$$

We can assume, for example, that  $\epsilon$  is given by  $\mathcal{N}(\mu = 0, \sigma^2)$  which gives the likelihood

$$p(D|\theta) = \prod_{i=1}^N p(\text{price}_i | \text{rooms}_i, \theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \frac{(\text{price}_i - a_1 \text{rooms}_i)^2}{\sigma^2}}$$

This model has **two** parameters: the slope  $a_1$  and variance  $\sigma^2$



<sup>2</sup>Note that here  $\mu = a_0$  which, for simplicity, we assume to be zero.



# Maximum Likelihood Estimation

- ▶ Similar to building deterministic models, probabilistic model parameters need to be tuned/trained
- ▶ **Maximum-likelihood estimation (MLE)** is a method of estimating the parameters of a probabilistic model.
- ▶ Assume  $\theta$  is a vector of all parameters of the probabilistic model. (e.g.  $\theta = \{a_1, \sigma\}$ ).
- ▶ **MLE** is an extremum estimator<sup>3</sup> obtained by maximising an objective function of  $\theta$

---

<sup>3</sup>"Extremum estimators are a wide class of estimators for parametric models that are calculated through maximization (or minimization) of a certain objective function, which depends on the data." wikipedia.org

# Maximum Likelihood Estimation

## Definition

Assume  $f(\theta)$  is an objective function to be optimised (e.g. maximised), the *arg max* corresponds to the value of  $\theta$  that attains the maximum value of the objective function  $f$

$$\hat{\theta} = \mathit{arg\ max}_{\theta} f(\theta)$$

- ▶ Tuning the parameter is then equal to finding the maximum argument *arg max*

# Maximum Likelihood Estimation - General

- ▶ Maximum Likelihood Estimation (MLE) is a common method for solving such problems

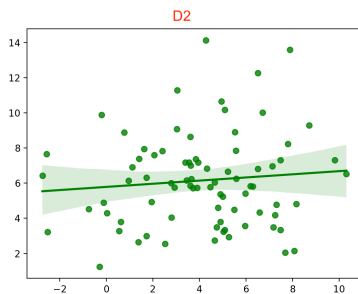
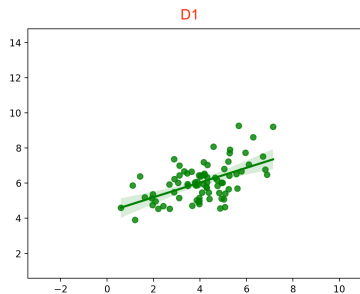
$$\begin{aligned}\theta_{MLE} &= \arg \max_{\theta} p(D|\theta) \\ &= \arg \max_{\theta} \ln p(D|\theta) \\ &= \arg \min_{\theta} -\ln p(D|\theta)\end{aligned}$$

## MLE Recipe

1. Determine  $\theta$ ,  $D$  and expression for likelihood  $p(D|\theta)$
2. Take the natural logarithm of the likelihood
3. Take the derivative of  $\ln p(D|\theta)$  w.r.t.  $\theta$ . If  $\theta$  is a multi-dimensional vector, take partial derivatives
4. Set derivative(s) to 0 and solve for  $\theta$

# Data Modelling - Deterministic vs Probabilistic

- ▶ **Probabilistic Models** can tell us **more**
- ▶ We could use the same MLE recipe to find  $\sigma_{ML}$ . This would tell us how uncertain our model is about the data  $D$ .
- ▶ For example: if we apply this method to two datasets ( $D_1$  and  $D_2$ ) what would the parameters  $\theta = \{a_1, \sigma\}$  be?



$$a_{1ML}^{D_1} > a_{1ML}^{D_2} \text{ [slope]} \text{ and } \sigma_{ML}^{D_1} < \sigma_{ML}^{D_2} \text{ [uncertainty]}^4$$

<sup>4</sup>The uncertainty ( $\sigma$ ) is represented by the light green bar in the plots. Test it yourself.

Quiz time!



Go to Blackboard unit page » Quizzes » Lecture 3.1

[Should take you less than 5 minutes]