COMS21202: Symbols, Patterns and Signals
Deterministic Data Models

[modified from Dima Damen lecture notes]

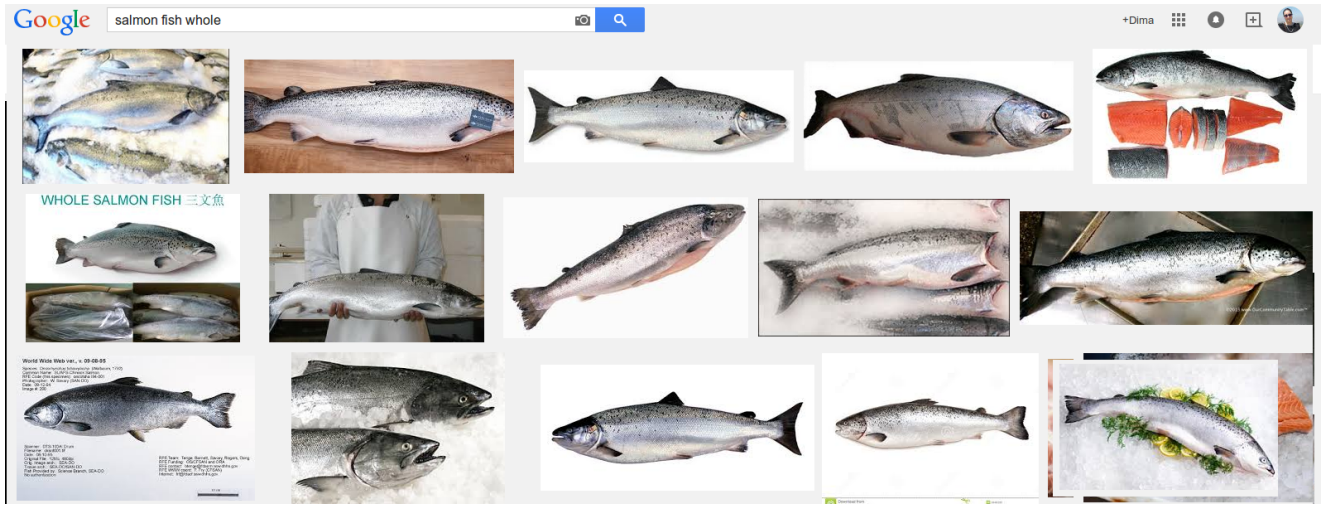
Rui Ponte Costa & Dima Damen
rui.costa@bristol.ac.uk

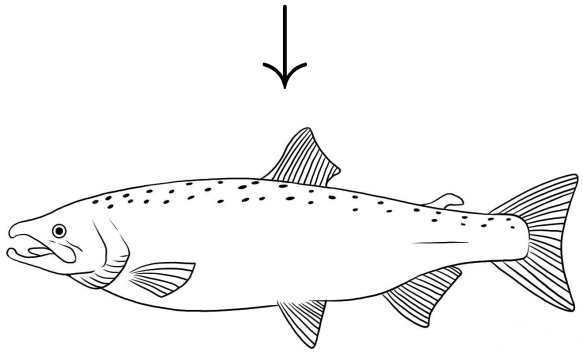
Department of Computer Science, University of Bristol
Bristol BS8 1UB, UK

March 9, 2020

Data Modelling

- From Data to a Model





Data Modelling

- Models are descriptions of the data
- They encode our assumptions about the data
- Enabling us to:
  - compare and contrast methods
  - quantify performance
- A model is ‘more than’ the data — it should be a ‘generalisation’ of the data

Data Modelling

e.g. build a model of Messi as he rolls the ball across the pitch

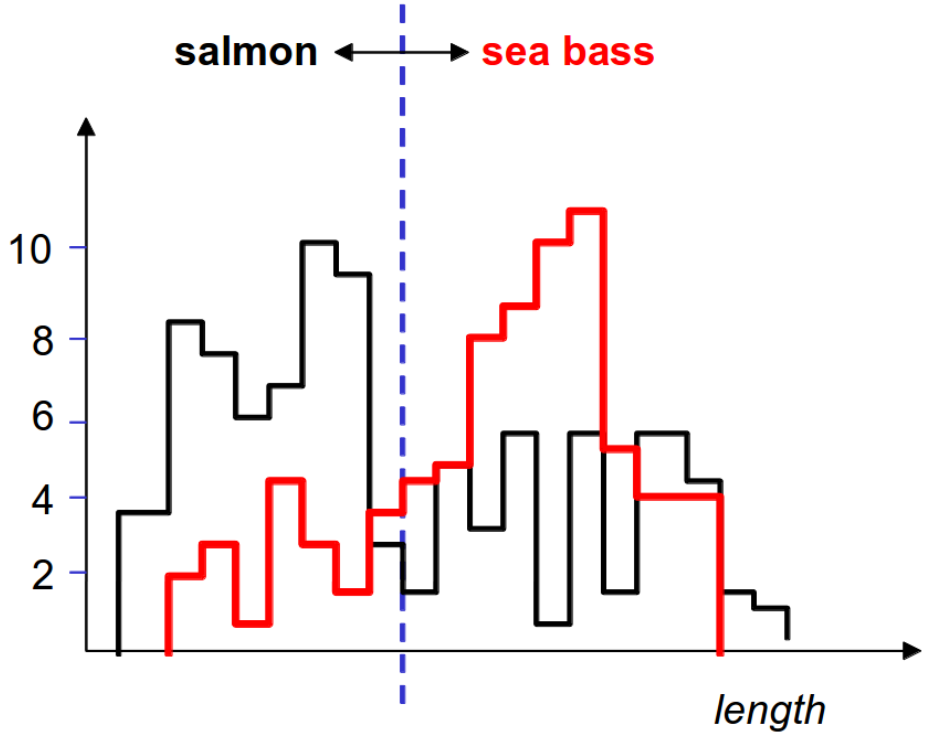
Data: collect data of body joints during action from multiple examples
Model: ?

Data Modelling

- **No need to play God**
- Models do not have to exactly describe the ‘real world’, nor exactly model how data was generated (e.g. the full human body)
- Instead, a model is *an abstraction of reality* – only approximates the underlying physical process (e.g. only joints)
- Models only need to enable us to define a method for a given task
- Performance of the method then depends on how well the model ‘maps’ the data onto the required solution
- Which model to use? Depends on its *practicality* as well as our *assumptions* about the data

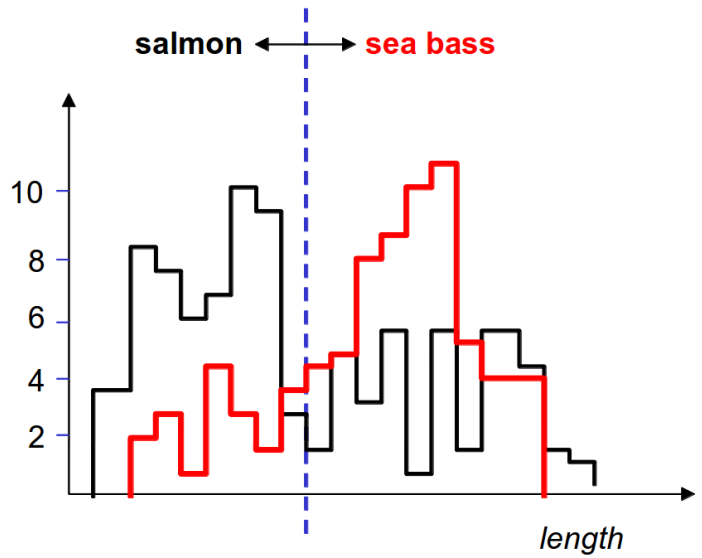
Fish Again,

- When classifying, we wish to find a model that can "understand" the difference between two classes by maximising their discrimination
- Model selected here is a linear classifier

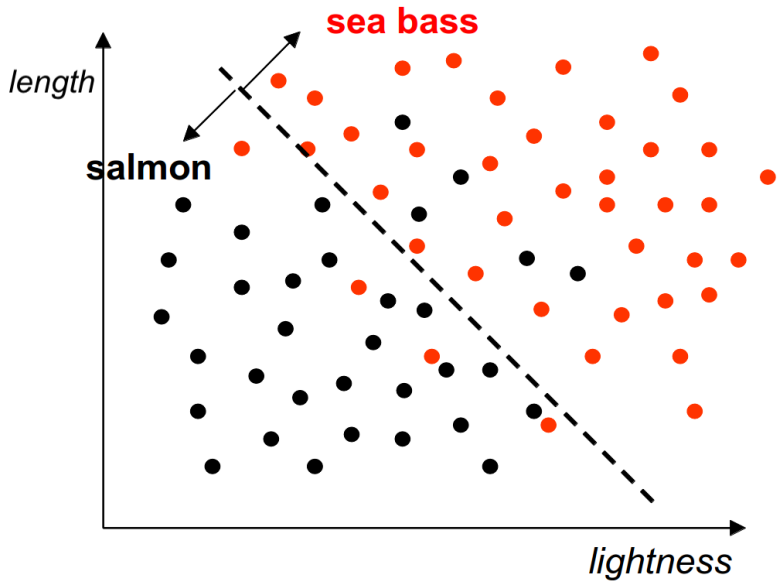


Model Parameters

- Models are defined in terms of parameters (one or more)
- These may be empirically obtained e.g. by trial and error
- or from training data by tuning or training the model



one parameter needed $x = t$



two parameters needed

$y = mx + c$

Generalisation vs. Overfitting

- **Generalisation** is the *probably* the most fundamental concept in machine learning.
- We do not *really* care about performance on training data - we already know the answer
- We care about whether we can take a decision on *new/unseen* data (i.e. outside the training data)
- A good performance on training data is only a means to an end, not a goal in itself
- In fact trying *too hard* on training data leads to a damaging phenomenon called **overfitting**

Generalisation vs. Overfitting

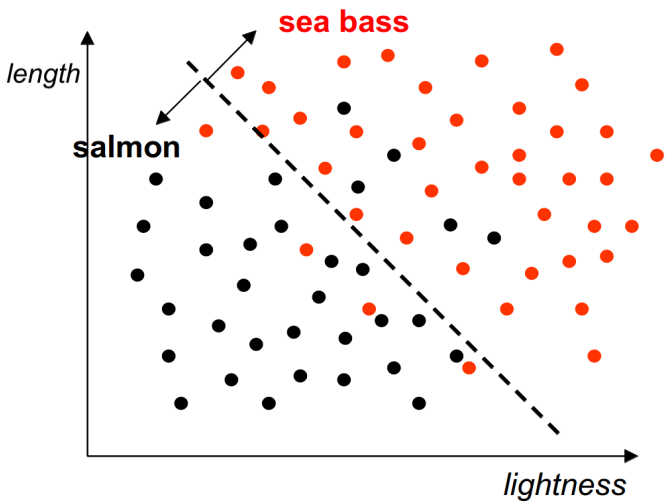
Example

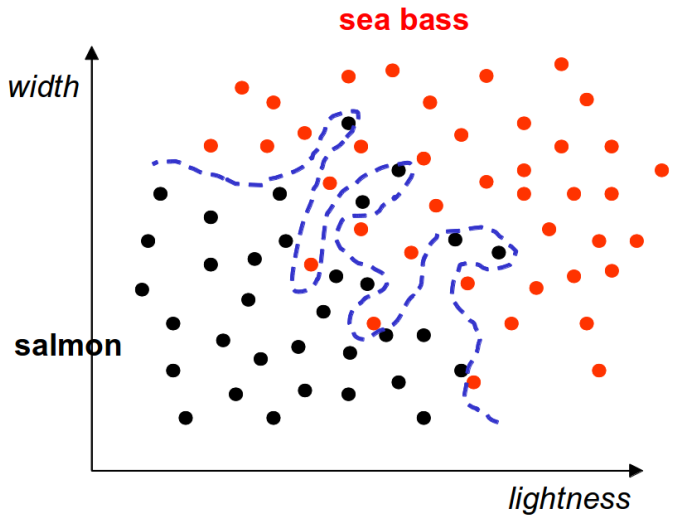
Imagine you are trying to prepare for *Symbols, Patterns and Signals* exam this June. You have access to previous exam papers and their worked answers available online. You begin by trying to answer the previous papers and comparing your answers with the model answers provided. *Next, you get carried away and spend all your time on memorising the model answers to all past papers.* Now if the upcoming exam completely consists of past questions, you are certainly to do well. But if the new exam asks different questions, you would be ill-prepared. In this case, you are *overfitting* the past exam papers and the knowledge you gained did not *generalise* to future exam questions.

source: Flach (2012), Machine Learning

Generalisation vs. Overfitting

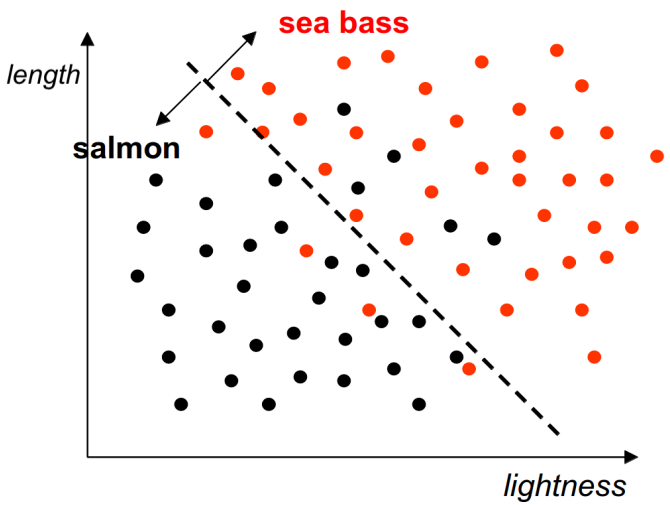
- Which model is more likely to **overfit** the data?





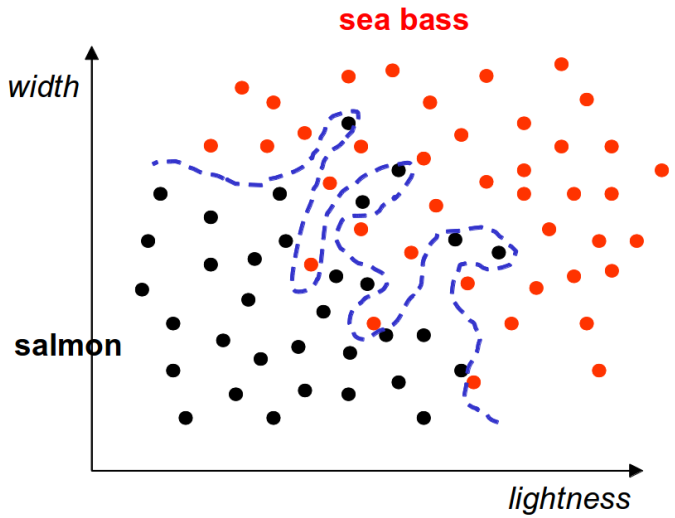
Generalisation vs. Overfitting

- **Simpler models** often give good performance and can be more **general**
- **highly complex models** over-fit the training data



two parameters needed

$y = mx + c$



A large number of parameters needs to be tuned

Deterministic Models

- **Deterministic models** produce an output without a confidence measure
- e.g. For the *fishy* model, prediction of whether the fish is salmon or sea bass is given, without an estimate of how good the prediction is
- Deterministic models do not encode the uncertainty in the data
- This is in contrast to probabilistic models (next lecture)

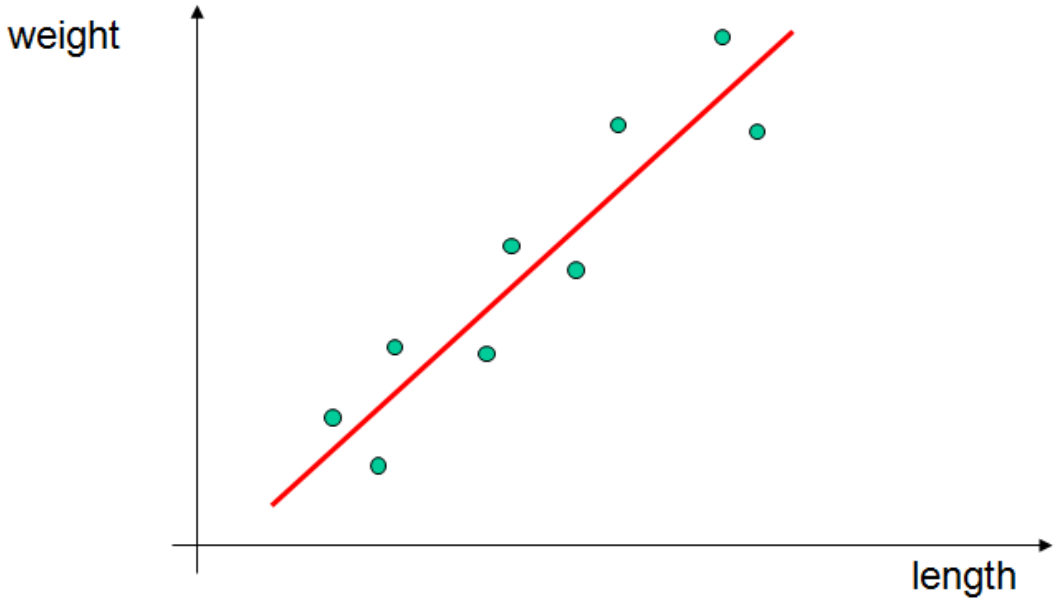
Deterministic Models

To build a deterministic model,

1. Understand the task
2. Hypothesise the model's type
3. Hypothesise the model's complexity
4. Tune/Train the model's parameters

Another Fish Problem: regression

- Goal: Finding a relationship between two variables (e.g. regress *weight* against *length*).
- Model: Linear relationship between weight and length?



Another Fish Problem: regression

Data: a set of data points $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ where x_i is the length of fish i and y_i is the weight of fish i.

Task: build a model that can predict the weight of a fish from its length

Model Type: assume there exists a polynomial relationship between length and weight

Model Complexity: assume the relationship is linear
$weight = a + b * length$

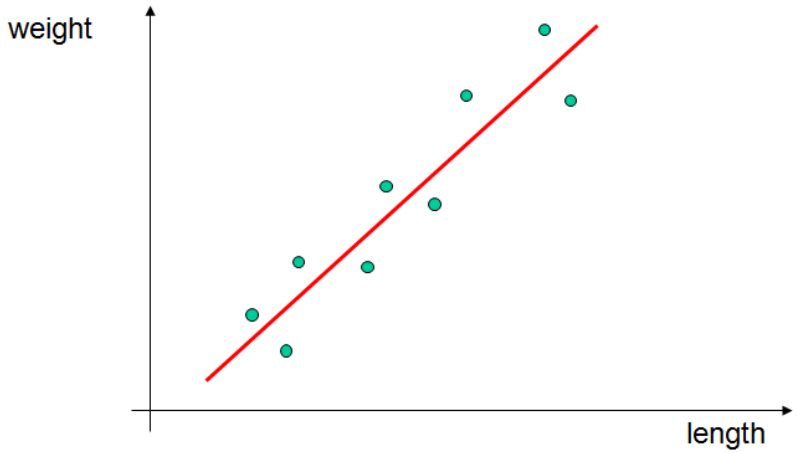
$$y_i = a + bx_i \tag{1}$$

Model Parameters: model has two parameters a and b which should be estimated.

- a is the y-intercept
- b is the slope of the line

Determinist Model - Line Fitting

- Finding the linear model parameters amounts to finding the *best fitting line* given the data
- **criterion:** The best fitting line is that which minimises a distance measure from the points to the line

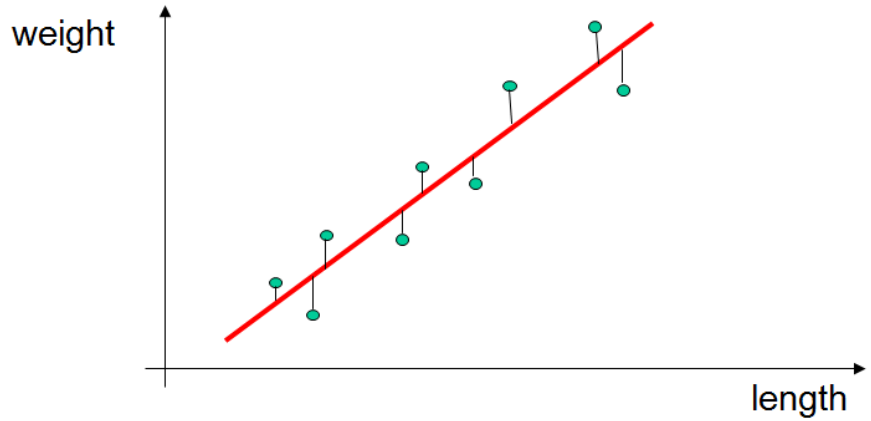


Determinist Model - Line Fitting

- Find a and b which minimises

$$R(a,b) = \sum_{i=1}^{N} (y_i - (a + bx_i))^2$$

- This is known as the residual/error[1]
- A method which gives a closed form solution is to minimise the sum of squared vertical offsets of the points from the line, **Method of Least-Squares**



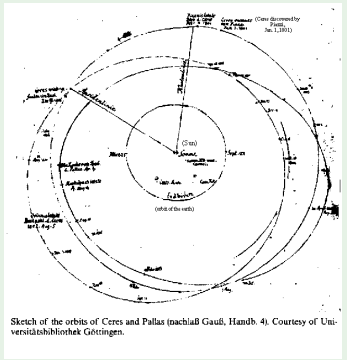
[1] Identical to the Euclidean (or L2 norm) distance we discussed in the last lecture.

Least Squares Solution

Example

The Ceres Orbit of Gauss:

On Jan 1, 1801, the Italian astronomer G. Piazzi discovered the asteroid Ceres. He was able to track the asteroid for six weeks but it was lost due to interference caused by the sun. A number of leading astronomers published papers predicting the orbit of the asteroid. Gauss also published a forecast, but his predicted orbit differed considerably from the others. Ceres was relocated by one observer on Dec 7 1801 and by another on Jan 1, 1802. In both cases the position was very close to that predicted by Gauss. Needless to say Gauss won instant fame in astronomical circles and for a time was more well known as an astronomer than as a mathematician. One of the keys to Gauss's success was his use of the method of least squares.



Sketch of the orbits of Ceres and Pallas (nachlaß Gauß, Handb. 4). Courtesy of Universitätsbibliothek Göttingen.

source: Leon (1994). Linear Algebra and its Applications

Least Squares Solution

Minimise residual by taking the partial derivatives w.r.t. the parameters (a,b), and setting them to zero (using chain rule) [2]

$$R(a,b) = \sum_i (y_i - (a + bx_i))^2$$

$$\frac{\partial R}{\partial a} = -2 \sum_i (y_i - (a + bx_i)) = 0$$

$$\frac{\partial R}{\partial b} = -2 \sum_i x_i (y_i - (a + bx_i)) = 0$$

Least Squares solution :

$$a_{LS} = \bar{y} - b_{LS}\bar{x}$$

$$b_{LS} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2}$$

$\bar{x} \equiv$ mean of $\{x_i\}$

[2] Full derivation here

Least Squares Solution Example

Example

Find the best least squares fit by a linear function to the data

x	-1	0	1	2
y	0	1	3	9

$\bar{x} = 0.5, \bar{y} = 3.25$

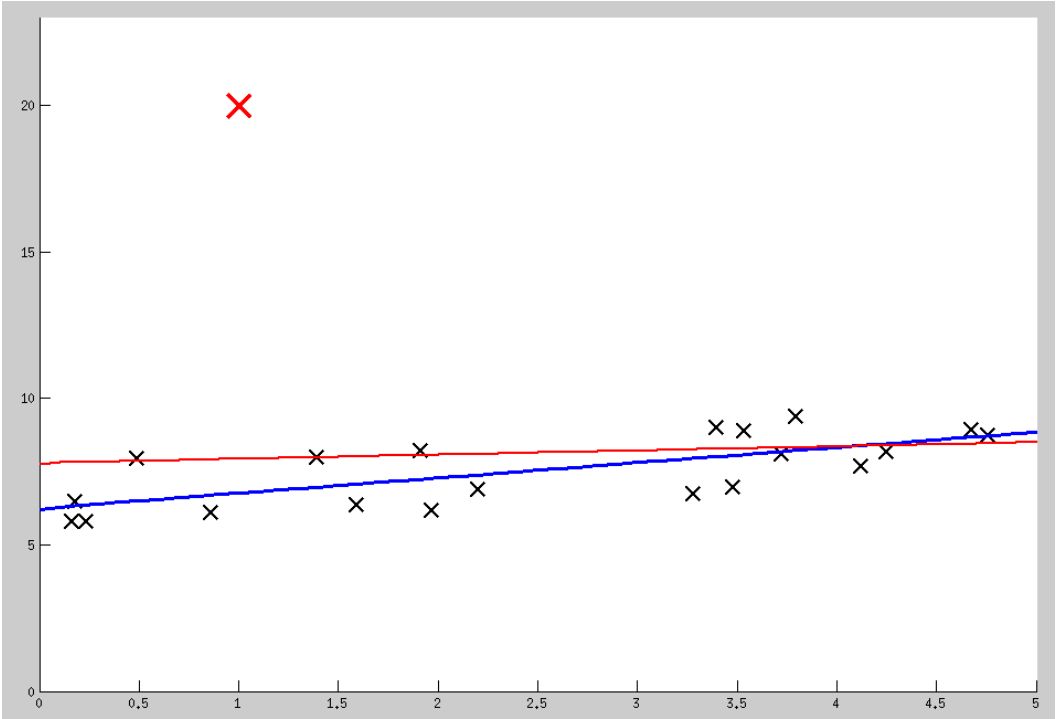
$b_{LS} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} = \frac{21 - 4 \times 0.5 \times 3.25}{6 - 4 \times 0.5^2} = 2.9$

$a_{LS} = \bar{y} - b_{LS}\bar{x} = 3.25 - 0.5 b_{LS} = 1.8$

$y = 1.8 + 2.9x$

Least Squares Solution - Outliers

- Outliers can have disproportionate effects on parameter estimates when using least squares
- Because residual is defined in terms of squared differences
- 'Best line' moves closer to outliers (Lab - week 15)



Least Squares Solution - matrix form

- Least squared solution can be defined using matrices and vectors
- Easier when dealing with variables

$$R(a,b) = \sum_i (y_i - (a + bx_i))^2 = \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2$$

where $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$, $\mathbf{a} = \begin{bmatrix} a \\ b \end{bmatrix}$

$$\mathbf{y} - \mathbf{X}\mathbf{a} = \begin{bmatrix} y_1 - a - bx_1 \\ \vdots \\ y_N - a - bx_N \end{bmatrix}$$

Least Squares Solution - matrix form

- To solve least squares in matrix form, find $\mathbf{a}_{LS}$; [3]

$$\|\mathbf{y} - \mathbf{X}\,\mathbf{a}_{LS}\|^2 = 0 \qquad \text{(minimise vector's length)}$$
$$\mathbf{y} - \mathbf{X}\,\mathbf{a}_{LS} = 0 \qquad \text{(optimal vector is of length 0)}$$
$$\mathbf{X}\,\mathbf{a}_{LS} = \mathbf{y} \qquad \text{(re-arrange)}$$
$$\mathbf{X}^T\mathbf{X}\,\mathbf{a}_{LS} = \mathbf{X}^T\,\mathbf{y} \qquad \text{(to get a square matrix)}$$
$$\mathbf{a}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y} \qquad \text{(matrix inverse)}$$

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

WARNING: This is not a derivation! It merely intends to give you intuition of the solution. For accurate understanding please refer to: this derivation - p8

[3] $\|\mathbf{A}\|^2 = \sqrt{\sum\sum |a_{ij}|^2}$ denotes the Frobenius norm, defined as the square root of the sum of the absolute squares of its elements.

Least Squares Solution Example - again

Example

Find the best least squares fit by a linear function to the data

x	-1	0	1	2
y	0	1	3	9

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

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

$$y = 1.8 + 2.9x$$

K-D Least Squares - matrix form

- Matrix formulation allows least squares method to be easily extended to data points in higher (K) dimensions
- Consider set of points $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$ where $\mathbf{x}_i$ has K dimensions
- For a model where y_i is linearly related to $\mathbf{x}_i$

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_K \; x_{iK} \tag{2}$$

K-D Least Squares - matrix form

- Solved in the same manner

$$\mathbf{y}_{(N\times 1)} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{X}_{(N\times(K+1))} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{NK} \end{bmatrix}, \mathbf{a}_{((K+1)\times 1)} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_K \end{bmatrix}$$

$$R(\mathbf{a}) = \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2$$

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

where $(\mathbf{X}^T\mathbf{X})$ is a $(K+1)\times(K+1)$ square matrix

General Least Squares - matrix form

- Matrix formulation also allows least squares method to be extended to **polynomial fitting**
- For a polynomial of degree $p + 1$

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

General Least Squares - matrix form

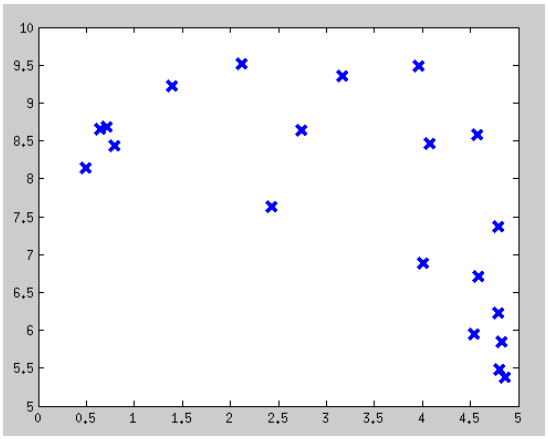
- Solved in the same manner

$$\mathbf{y}_{(N\times 1)} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{X}_{(N\times(p+1))} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^p \end{bmatrix}, \mathbf{a}_{((p+1)\times 1)} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix}$$

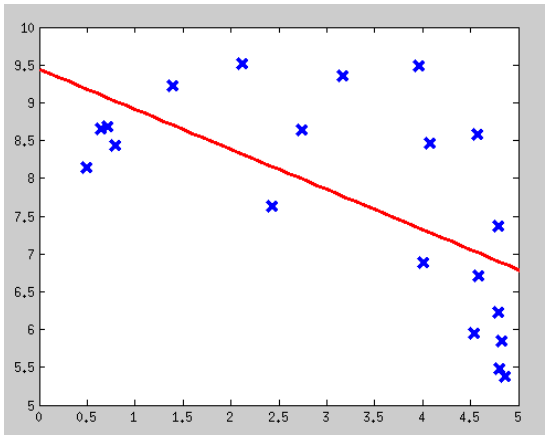
$$\mathbf{a}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$$

where $(\mathbf{X}^T\mathbf{X})$ is a $(p+1)\times(p+1)$ square matrix

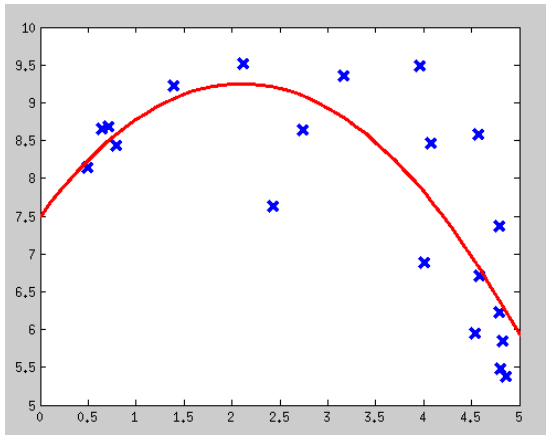
Generalisation and Overfitting - again



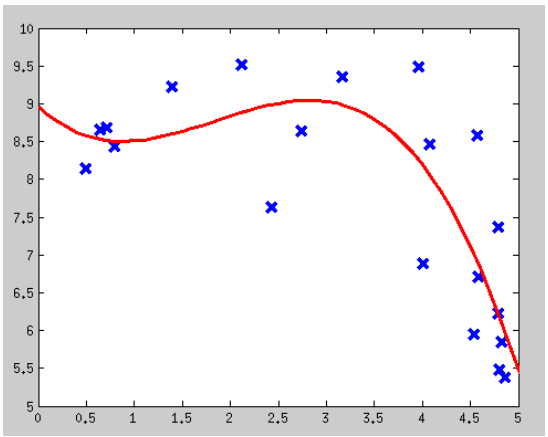
Data



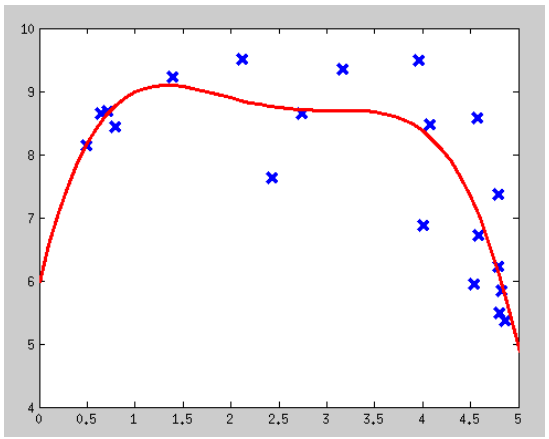
$p = 1$
Residual = 4.7557



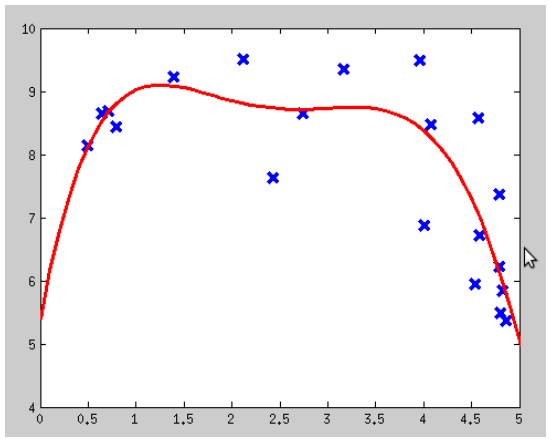
$p = 2$
Residual = 3.7405



$p = 3$
Residual = 3.5744



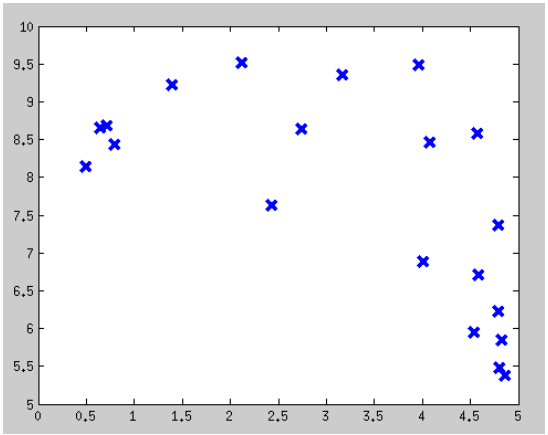
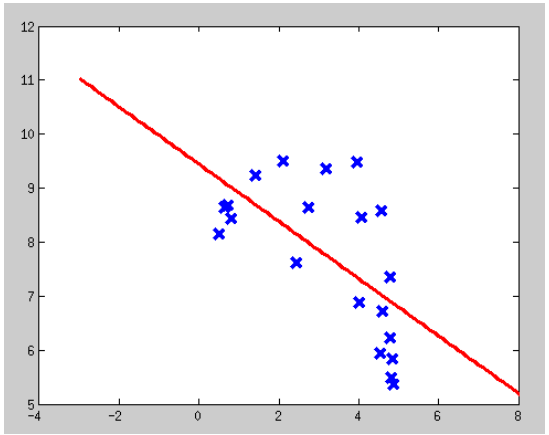
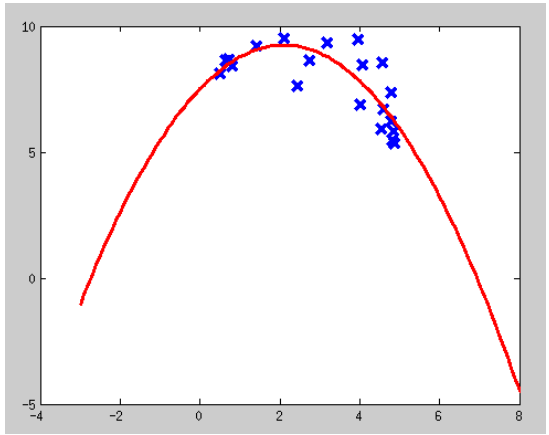
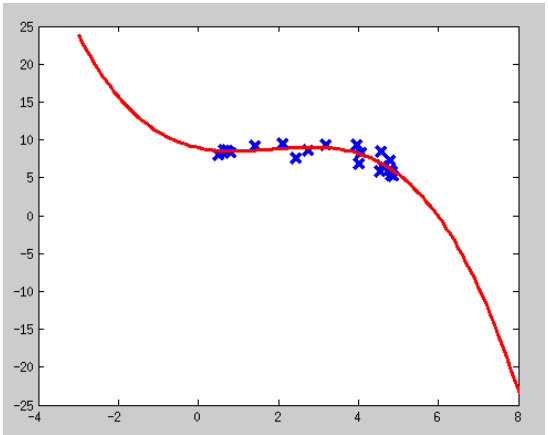
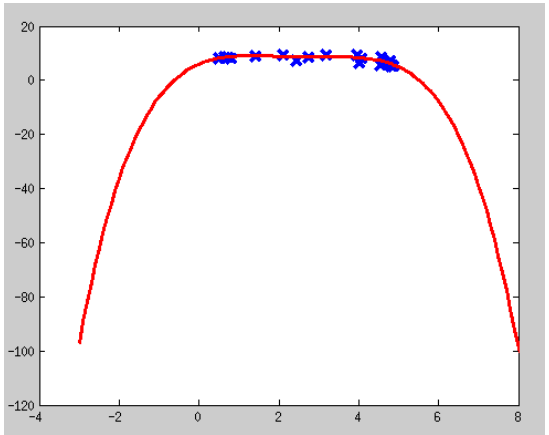
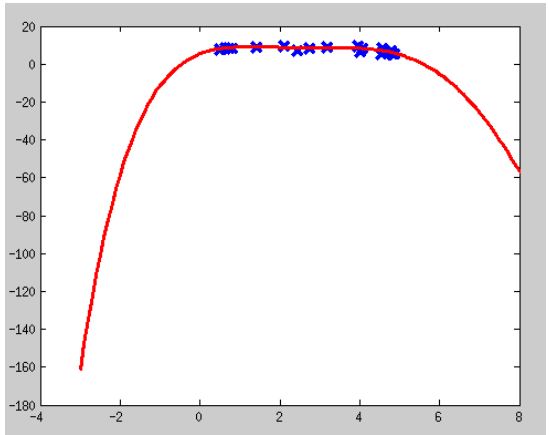
$p = 4$
Residual = 3.4236



$p = 5$
Residual = 3.4217

Generalisation and Overfitting - again

- Strong effect on how to generalise to future data

Tasks

- Next Lab (Week 14): Introduction to Jupyter Notebook II
- Sheet on unit webpage

- Next Problem Class (Wed 9-10): Outliers and least squares
- Prepare your answers in advance [available unit webpage]

Further Reading

- Linear Algebra and its applications
  Lay (2012)
  - Section 6.5
  - Section 6.6
  - Available online
    `http://www.math.usu.edu/powell/pseudoinverses.pdf`