COMP4388: **MACHINE LEARNING**

Linear Regression - Part 2:

- Multivariate Regression
- Normal Equation
- Other forms of regression models

Dr. Radi Jarrar
Department of Computer Science
جامعة بيرزيت
BIRZEIT UNIVERSITY

Linear Regress – Multiple Features

- How to handle the cases in which there is more than one feature?
- Area, Nr. Surrounding Roads, Distance from City Centre, …
- Different features denoted as x_1, x_2, x_3, …
- x^1represents the order of the feature vector
- x^1_3 represents the third feature of the first feature vector

LR – Multiple Features (2)

- With a single feature, the hypothesis was

$$h(x) = a + bx$$

- In the case of multiple features, the hypothesis becomes

$$h(x) = a + bx_1 + gx_2 + \ldots$$

- For simplification, consider $x_0 = 1$ and the parameters as follows

$$h(x) = a_0 x_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

LR – Multiple Features (3)

- where $x=[x_0, x_1, x_2, x_3, \ldots, x_n]$

$$a = [a_0, a_1, a_2, a_3, \ldots, a_n]$$

- so

$$h(x) = a_0 x_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n$$

- which is equivalent to

$$h(x) = a^T x$$

GD for Multiple Features

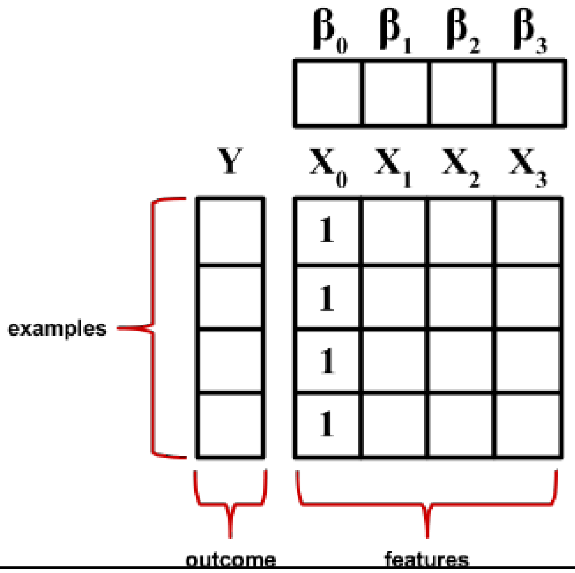
- Hypothesis: $h(x) = a^T x$
- Parameters: a
- Cost function $J(\alpha) = \frac{1}{N} \sum_{n=1}^{N} (h(x_n) - y_n)^2$
- Gradient decsent $a_j = a_j - F \frac{¶}{¶a_j} J(a)$

* Repeat GD and simeltanuosly update for every j=0, 1, ..., n

Learning rate

- When choosing a small value for the learning rate, the cost function has to decrease after every iteration
- If the learning rate is too small, GD can be very slow to converge
- If the learning rate is too large, the cost function may not decrease on every iteration (i.e., may not converge)
- Learning rate can be selected as 0.001, 0.01, 0.1, 1, 10, 100, ...

The Normal Equation



The Normal Equation (2)

- Matrix Algebra can be used to solve for vector β (that minimises the sum of squared errors between the predicted and actual y values)

$$\alpha = \left(X^{T}X\right)^{-1}X^{T}Y$$

The Normal Equation (3)

Price(Y)	Area (m²)	Distance to CC	Nr. of Roads
40000	600	100	2
50000	650	50	2
60000	800	100	3
100000	1000	50	2
35000	600	300	1

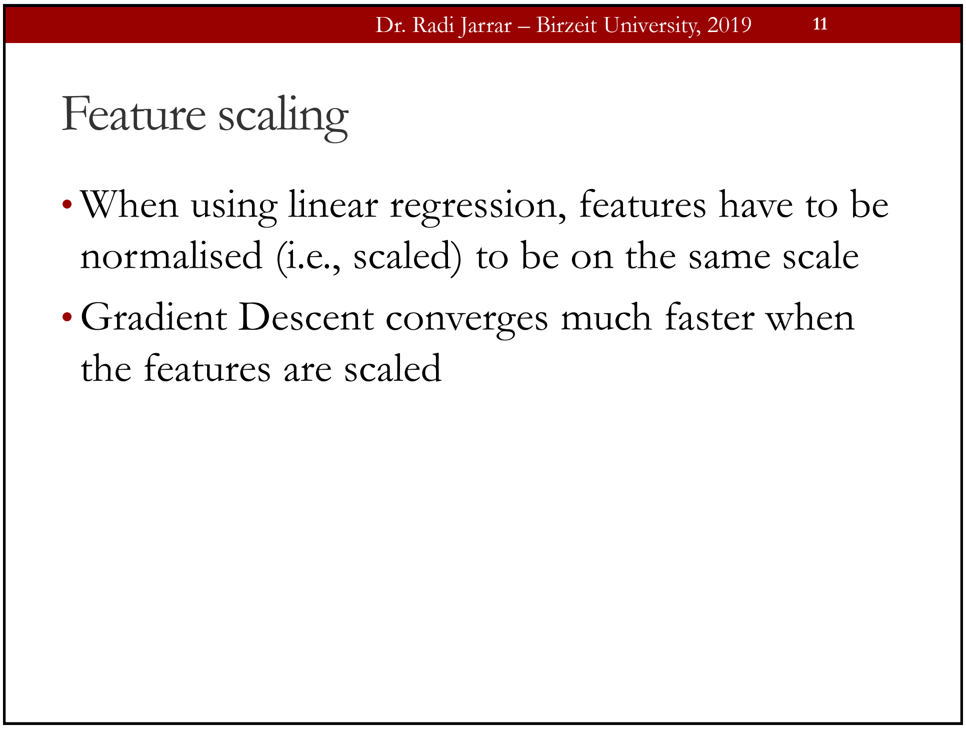
- To represent it using the Normal Method, add $x_0=1$:

$$X = \begin{bmatrix} 1 & 600 & 100 & 2 \\ 1 & 650 & 50 & 2 \\ 1 & 800 & 100 & 3 \\ 1 & 1000 & 50 & 2 \\ 1 & 600 & 300 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 40000 \\ 50000 \\ 60000 \\ 100000 \\ 35000 \end{bmatrix}$$

GD vs. Normal Equation

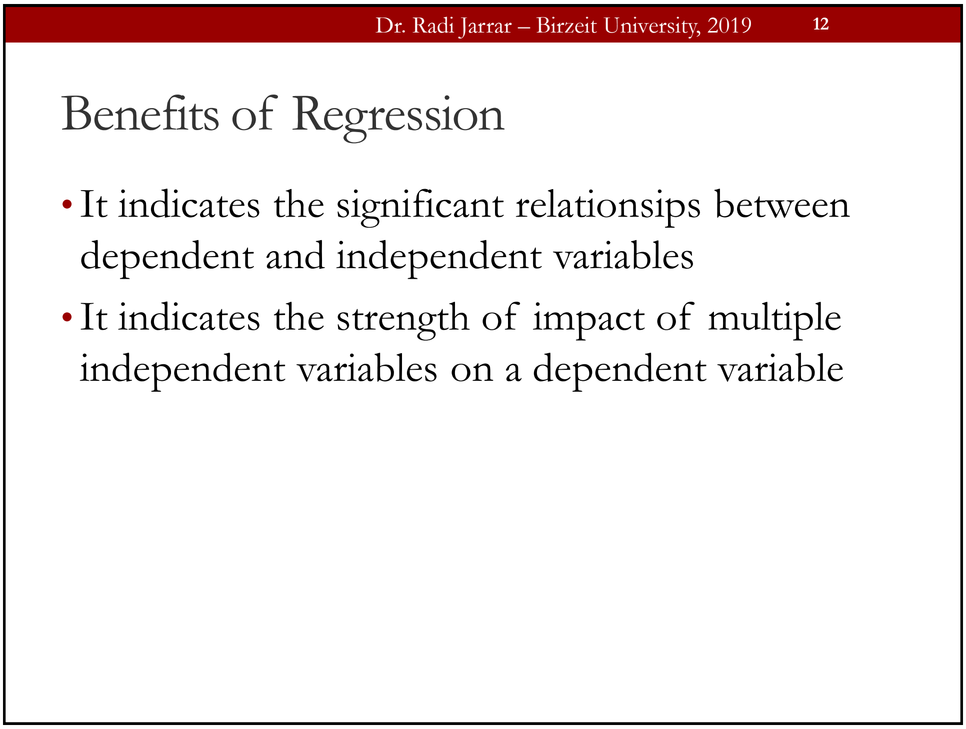
- GD
  - Should choose a value for the learning rate
  - Takes many iterations to find the optimal values
  - Works very well even if the dataset is large
- Normal Equation
  - No learning rate
  - No iterations needed
  - Computing X^TX takes $O(N^3)$
  - Slow (especially with large datasets

→ Use Normal equation if the number of features <1000



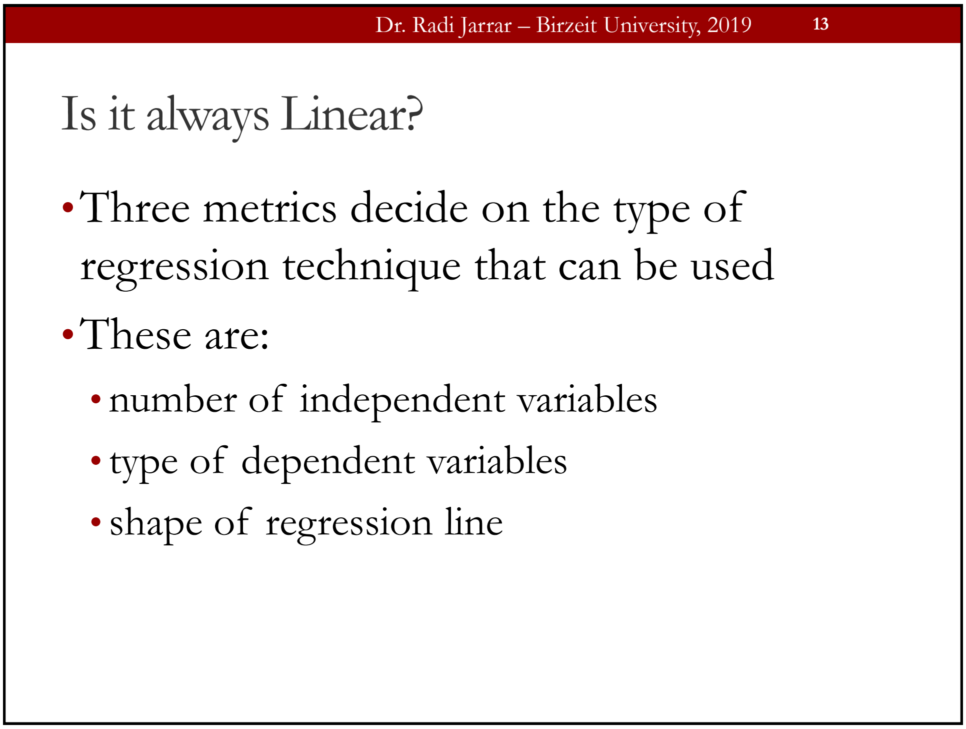
Feature scaling

- When using linear regression, features have to be normalised (i.e., scaled) to be on the same scale
- Gradient Descent converges much faster when the features are scaled



Benefits of Regression

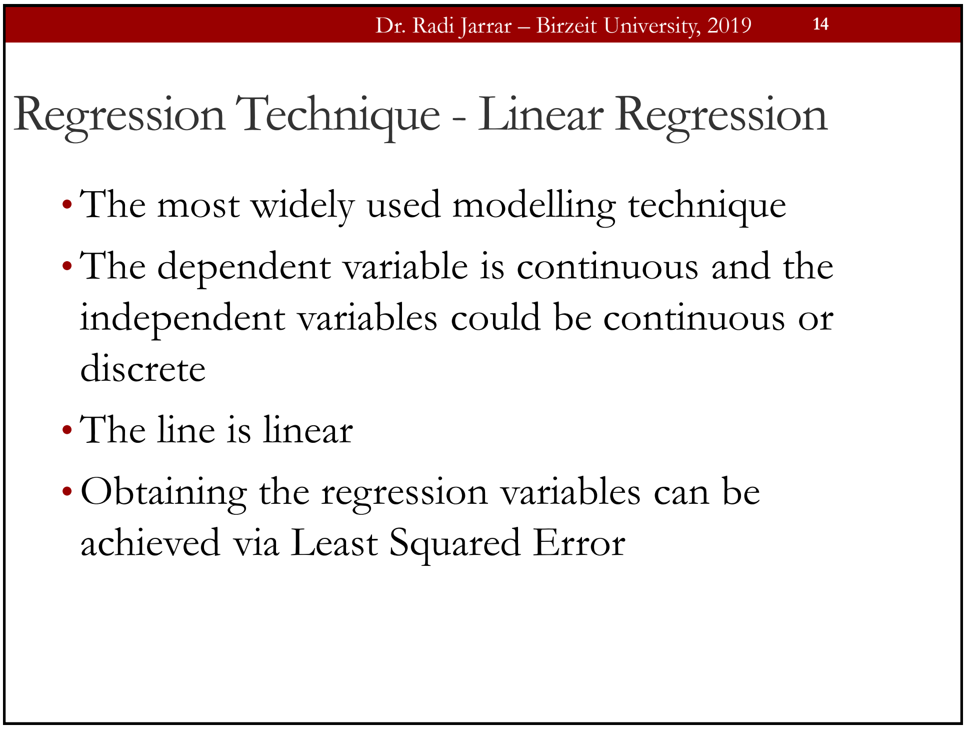
- It indicates the significant relationsips between dependent and independent variables
- It indicates the strength of impact of multiple independent variables on a dependent variable



Dr. Radi Jarrar – Birzeit University, 2019 13

Is it always Linear?

- Three metrics decide on the type of regression technique that can be used
- These are:
  - number of independent variables
  - type of dependent variables
  - shape of regression line



Dr. Radi Jarrar – Birzeit University, 2019 14

Regression Technique - Linear Regression

- The most widely used modelling technique
- The dependent variable is continuous and the independent variables could be continuous or discrete
- The line is linear
- Obtaining the regression variables can be achieved via Least Squared Error

Regression Technique - Linear Regression (2)

- There must be a linear relationship between the dependent variable and the independent variable(s)
- Very sensitive to outliers
- In case of multiple independent variables, feature selection (forward selection, backward elimination, or step-wise approach) can be used to select the most significant independent variables

Regression Technique – Polynomial

- Polynomial
  - Is used when the relation between the independent variables and the dependent variable is not linear
  - The best fit is not a straight line. It is rather a curve that fits into data points
  - 2nd degree

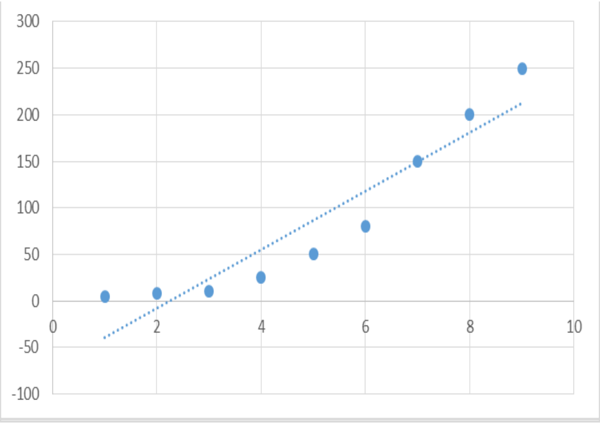
$$h(x) = \alpha + \beta \cdot x^2$$

  - 3rd degree

$$h(x) = \alpha + \beta \cdot x^3$$

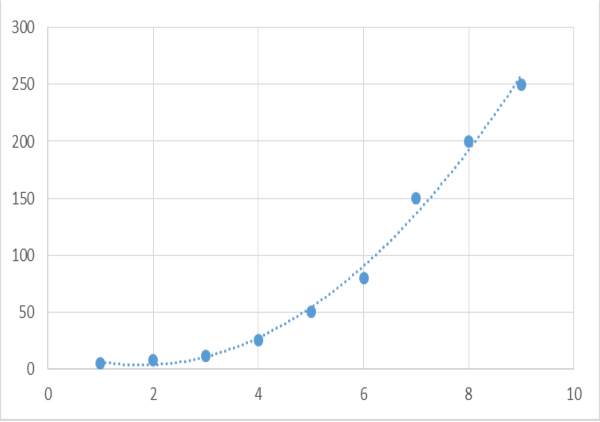
Regression Technique – Polynomial (2)

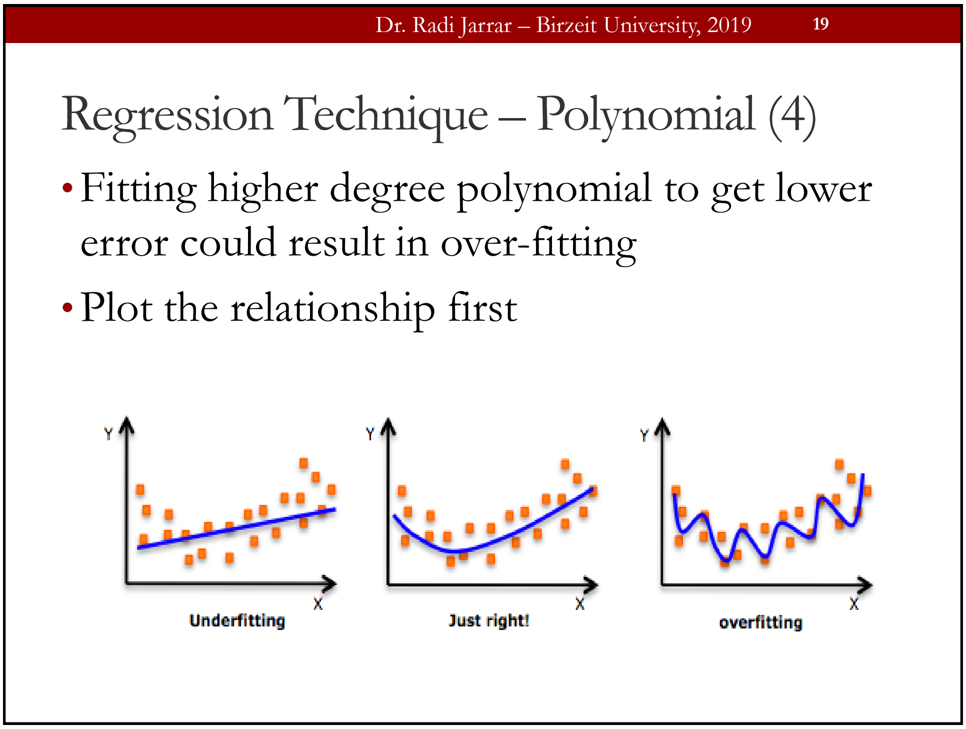
X	Y
1	5
2	8
3	11
4	25
5	50
6	80
7	150
8	200
9	250



Regression Technique – Polynomial (3)

X	Y
1	5
2	8
3	11
4	25
5	50
6	80
7	150
8	200
9	250





Regression Technique – Polynomial (4)

- Fitting higher degree polynomial to get lower error could result in over-fitting
- Plot the relationship first

Underfitting

Just right!

overfitting