MACHINE LEARNING, DATA SCIENCE, BIG DATA MINING

MACHINE LEARNING, DATA SCIENCE, BIG DATA MINING

Machine Learning for Data Mining Applications in Big Data

Prof. Dr. Dewan Md. Farid
Department of Computer Science & Engineering
United International University

https://cse.uiu.ac.bd/profiles/dewanfarid/

ABOUT THE AUTHOR

PROF. DR. DEWAN MD. FARID is a Professor of Computer Science and Engineering at United International University, and IEEE Senior Member. Prof. Farid worked as a Postdoctoral Fellow/Staff at the following research labs/groups: (1) Computational Intelligence Group (CIG), Department of Computer Science and Digital Technology, University of Northumbria at Newcastle, UK in 2013, (2) Computational Modelling Lab (CoMo) and Artificial Intelligence Research Group, Department of Computer Science, Vrije Universiteit Brussel, Belgium in 2015-2016, and (3) Decision and Information Systems for Production systems (DISP) Laboratory, IUT Lumière – Université Lyon 2, France in 2020. Prof. Farid was a Visiting Faculty at the Faculty of Engineering, University of Porto, Portugal in June 2016. He holds a PhD in Computer Science and Engineering from Jahangirnagar University, Bangladesh in 2012. Part of his PhD research has been done at ERIC Laboratory, University Lumière Lyon 2, France by Erasmus-Mundus ECW eLink PhD Exchange Program. His PhD was fully funded by Ministry of Science & Information and Communication Technology, Government of the People's Republic of Bangladesh and European Union (EU) eLink project. Prof. Farid has published 109 peer-reviewed scientific articles, including 30 highly esteemed journals like Expert Systems with Applications, Journal of Theoretical Biology, Journal of Neuroscience Methods, Bioinformatics, Scientific Reports (Nature), Proteins and so on in the field of Machine Learning, Data Mining and Big Data. Prof. Farid re-

ceived the following awards: (1) Dr. Fatema Rashid Best Paper Award (2nd Position) for the paper titled "KNNTree: A new method to ameliorate k-nearest neighbour classification using decision tree" in 3rd International Conference on Electrical Computer and Communication Engineering (ECCE 2023), CUET, Chittagong, Bangladesh, (2) JuliaCon 2019 Travel Award for attending Julia Conference at the University of Maryland, Baltimore, USA, and (3) United Group Research Award 2016 in the field of Science and Engineering. He received the following research funds as Principal Investigator: (1) a2i Innovation Fund of Innov-A-Thon 2018 (Ideabank ID No.: 12502) from a2i-Access to Information Program – II, Information and Communication Technology (ICT) Division, Government of the People's Republic of Bangladesh, and (2) Project Code: UIU/IAR/01/2021/SE/23 received from Institute for Advanced Research (IAR), United International University. Prof. Farid received the following Erasmus Mundus scholarships: (1) LEADERS (Leading mobility between Europe and Asia in Developing Engineering Education and Research) to undertake a staff level mobility at the Faculty of Engineering, University of Porto, Portugal in 2015, (2) cLink (Centre of excellence for Learning, Innovation, Networking and Knowledge) for pursuing Postdoc at University of Northumbria at Newcastle, UK in 2013, and (3) eLink (east west Link for Innovation, Networking and Knowledge exchange) for pursuing Ph.D. at University Lumière Lyon 2, France in 2009. Prof. Farid also received Senior Fellowship I and II awards by National Science & Information and Communication Technology (NSICT), Ministry of Science & Information and Communication Technology, Government of the People's Republic of Bangladesh respectively in 2008 and 2011 for pursuing Ph.D. at Jahangirnagar University. He visited 18 countries for attending international conferences, research and higher education. Prof. Farid delivered several invited/keynote talks including an invited research talk at Data to AI Group (DAI), Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA.

CONTENTS

SYMBOLS

$\mathbb{X}$ The set of training instances/ training data
N Size of $\mathbb{X}$
$(x^{(i)}, y^{(i)})$ The i-th instance pair in $\mathbb{X}$ (supervised learning)
$x^{(i)}$ The input (features) of i-th training instance in $\mathbb{X}$ (unsupervised learning)
$x_j^{(i)}$ The value of feature j in i-th training instance
D Dimension of an instance $x^{(i)}$
K Dimension of a label $y^{(i)}$
$X \in \mathbb{R}^{N \times D}$ Design matrix, where X_i,: denotes $x^{(i)}$
X_i A feature in training data
$\mathbb{F}$ Hypothesis space of functions to be learnt, i.e., a model
$C[f]$ A cost function of $f \in \mathbb{F}$
$C[\theta]$ A cost function of θ parameterising $f \in \mathbb{F}$
(x', y') A testing pair
$\hat{y}$ Label predicted by a function f, i.e., $\hat{y} = f(x')$ (supervised learning)

$P(x, y)$ A data generating distribution

PART I

LEARNING FROM DATA

CHAPTER 1

SUPERVISED LEARNING

All knowledge - past, present, and future - can be derived from data by a single, universal learning algorithm.

—Pedro Domingos, Professor Emeritus of CSE at the University of Washington

Machine intelligence is the last invention that humanity will ever need to make.

—Nick Bostrom, Professor at the University of Oxford

Supervised learning, also known as supervised machine learning, is used to trained models/concepts from labelled datasets. The labelled data includes instances/data points, x_i as well as correct class-labels, y_i that allow the models to learn over time. Supervised learning can be bifurcated into two subcategories: Regression and Classification.

Regression is used to understand the relationship between dependent and independent variables, e.g. linear regression. It works with labelled data when the output/dependent variable is numeric/real-value.

Classification is used to classify the labelled data when the output/class variable is category/nominal, e.g. support vector machines (SVM), decision trees, k-nearest neighbour classifier etc.

1.1 Linear Regression

Linear Regression is a supervised machine learning algorithm where the predicted output, $\hat{y}$ is continuous (e.g. sales, price) and has a constant slope. It's used to predict values, $\hat{y}_i$ based on independent variables $X_1, X_2, \ldots, X_p$ – the kind of relationship between independent variables $X_1, X_2, \ldots, X_p$ and dependent variables $y_1, y_2, \ldots, y_q$. Therefore, the regression models are used to describe relationships between variables by fitting a line from the training data. It can be classified into three types:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Multivariate Linear Regression

1.1.1 Simple Linear Regression

Simple linear regression uses traditional slope-intercept form to predict the dependent variable (y) as continuous output based on a given independent variable (x) as input variable. Fig. 1.1 shows a simple linear regression example where experience is horizontal axis x-axis and salary is vertical axis y-axis. If $\omega_0 = 0$ which mean the experience is zero and ω_1 point indicate the point at y-axis where a person with zero experience salary would be 30K, and a person who has 1 more year of experience is eligible for extra 1K salary. So, the salary hike depends on the slop if slop is a high person may get more than 1K salary or if slop is less salary increment ratio will effect accordingly. Eq. 1.1 shows the simple linear regression function, where x is the input training data, y is the labels to data, ω_0 is y-intercept of the regression line, and ω_1 is slope of the regression line/coefficient of independent variable x (in machine learning we call coefficients weights). Eq. 1.2 and Eq. 1.3 are shows the ω_0 and ω_1 calculations. The correlation coefficient (r), standard deviation of y, and standard deviation of x calculations are shown in Eq. 1.4 to Eq. 1.6 respectively.

$$y = \omega_0 + \omega_1 . x \tag{1.1}$$

$$\omega_0 = \bar{y} - \omega_1 . \bar{x} \tag{1.2}$$

$$\omega_1 = r \frac{Sy}{Sx} \tag{1.3}$$

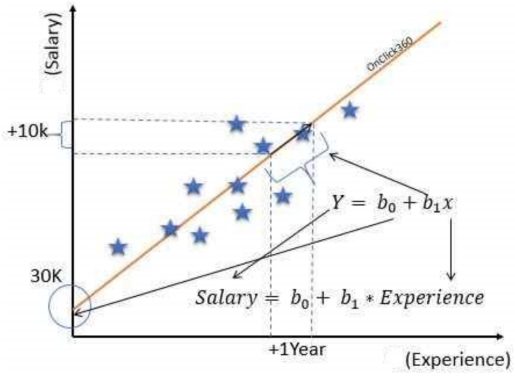

Figure 1.1: Simple linear regression where Experience = x- axis and salary = y- axis.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \tag{1.4}$$

$$Sy = \sqrt{\frac{\sum(y-\bar{y})^2}{N-1}} \tag{1.5}$$

$$Sx = \sqrt{\frac{\sum(x-\bar{x})^2}{N-1}} \tag{1.6}$$

1.1.1.1 *Illiterate an Example of Simple Linear Regression*

In a given regression problem that shown in Table 1.1, we need to find the best ω_0 and ω_1 values for getting the best fitted line. So, when we are finally using the simple linear regression model for prediction, it will predict the value of y for the input value of x.

Table 1.1: Simple linear regression example.

Input: x	Output: y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	150	-2	-100	200	4	10000
2	200	-1	-50	50	1	2500
4	300	1	50	50	1	2500
5	350	2	100	200	4	10000
$\bar{x} = 3$	$\bar{y} = 250$			$\sum = 500$	$\sum = 10$	$\sum = 25000$

So, the correlation coefficient, $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}} = 1$, standard deviation of $Sy = \sqrt{\frac{\sum(y-\bar{y})^2}{N-1}} = 52.704$ (where, $N = 4$, total number of training instances), standard deviation of $Sx = \sqrt{\frac{\sum(x-\bar{x})^2}{N-1}} = 1.054$. So, $\omega_1 = r\frac{Sy}{Sx} = 50.003$, and

$\omega_0 = \bar{y} - \omega_1.\bar{x} = 99.991$.

Now, we can predict the value of y for the input value of x, e.g., $x = 3$ then $y = \omega_0 + \omega_1.x = 99.991 + 50.003 \times 3 = 250$.

1.1.2 Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is an extension of simple linear regression. It is used when we want to predict the value of a dependent variable (the outcome/target/forecasting value) based on the value of two or more independent variables (explanatory variables). We use MLR to know the relationship is between two or more independent variables and one dependent variable, and the value of the dependent variable at a certain value of the independent variables. MLR formula is shown in Eq. 1.7, where y is the dependent variable, ω_0 is the y-intercept (value of y when all other parameters are set to 0), $\omega_1.X_1$ is the regression coefficient ω_1 of the first independent variable X_1, and $\omega_N.X_N$ the regression coefficient ω_N of the last independent variable X_N.

$$\begin{aligned} y &= \omega_0 + \omega_1.X_1 + \omega_2.X_2 + \cdots + \omega_n.X_n \\ &= \omega_0 + \sum_{i=1}^{n} \omega_i.X_i \end{aligned} \tag{1.7}$$

1.1.2.1 Illiterate an Example of MLR Let's consider the Table 1.2 where we have two independent variables (or input features), X_1 and X_2 and one dependent variable, y. Now, we will build a MLR model to predict the value of y based on the input variables of X_1 and X_2. So, the MLR formula will be $y = \omega_0 + \omega_1.X_1 + \omega_2.X_2$.

Table 1.2: MLR example: Training Data.

X_1	X_2	y
3	8	-3.7
4	5	3.5
5	7	2.5
6	3	11.5
2	1	5.7

We can apply Eq. 1.8, Eq. 1.9 and Eq. 1.10 to find the value of ω_0, ω_1, and ω_2 respectively. If we know the ω_0, ω_1, and ω_2 values then we can easily find the value of y based on the values of X_1 and X_2. Eq. 1.11 to Eq. 1.13 shows the $\sum {x_i}^2$, $\sum x_i y$, and $\sum x_1 x_2$ calculations.

$$\omega_0 = \bar{y} - \omega_1.\bar{X_1} - \omega_2.\bar{X_2} \tag{1.8}$$

$$\omega_1 = \frac{(\sum {x_2}^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum {x_1}^2)(\sum {x_2}^2) - (\sum x_1 x_2)^2} \tag{1.9}$$

$$\omega_2 = \frac{(\sum {x_1}^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum {x_1}^2)(\sum {x_2}^2) - (\sum x_1 x_2)^2} \tag{1.10}$$

Table 1.3: Multiple linear regression example.

X_1	X_2	y	$(X_1)^2$	$(X_2)^2$	$X_1 y$	$X_2 y$	$X_1 X_2$
3	8	-3.7	9	64	-11.1	-29.6	24
4	5	3.5	16	25	14	17.5	20
5	7	2.5	25	49	12.5	17.5	35
6	3	11.5	36	9	69	34.5	18
2	1	5.7	4	1	11.4	5.7	2
$\sum = 20$	$\sum = 24$	$\sum = 19.5$	$\sum = 90$	$\sum = 148$	$\sum = 95.8$	$\sum = 45.6$	$\sum = 99$
$\bar{X_1} = 4$	$\bar{X_2} = 4.8$	$\bar{y} = 3.9$					

$$\sum {x_i}^2 = \sum {X_i}^2 - \frac{(\sum X_i)^2}{N} \tag{1.11}$$

If $i = 1$, then $\sum {x_1}^2 = \sum {X_1}^2 - \frac{(\sum X_1)^2}{N} = 90 - \frac{(20)^2}{5} = 10$
If $i = 2$, then $\sum {x_2}^2 = \sum {X_2}^2 - \frac{(\sum X_2)^2}{N} = 148 - \frac{(24)^2}{5} = 32.8$

$$\sum x_i y = \sum X_i y - \frac{(\sum X_i)(\sum y)}{N} \tag{1.12}$$

If $i = 1$, then $\sum x_1 y = \sum X_1 y - \frac{(\sum X_1)(\sum y)}{N} = 95.8 - \frac{(29 \times 19.5)}{5} = 17.8$
If $i = 2$, then $\sum x_2 y = \sum X_2 y - \frac{(\sum X_2)(\sum y)}{N} = 45.6 - \frac{(24 \times 19.5)}{5} = -48$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N} \tag{1.13}$$

So, $\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 99 - \frac{(20 \times 24)}{5} = 3$
$\omega_1 = \frac{(\sum {x_2}^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum {x_1}^2)(\sum {x_2}^2) - (\sum x_1 x_2)^2} = \frac{(32.8 \times 17.8) - (3 \times -48)}{(10 \times 32.8) - (3)^2} = 2.28$
$\omega_2 = \frac{(\sum {x_1}^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum {x_1}^2)(\sum {x_2}^2) - (\sum x_1 x_2)^2} = \frac{(10 \times -48) - (3 \times 17.8)}{(10 \times 32.8) - (3)^2} = -1.67$
$\omega_0 = \bar{y} - \omega_1 . \bar{X_1} - \omega_2 . \bar{X_2} = 3.9 - (2.28 \times 4) - (-1.67 \times 4.8) = 2.796$
Therefore, $y = \omega_0 + \omega_1 . X_1 + \omega_2 . X_2 = 2.796 + (2.28 \times X_1) + (-1.67 \times X_2)$

1.1.3 Multivariate Linear Regression

Multivariate Linear Regression (MvLR), also called multivariate multiple regression, is a technique that estimates a single regression model with multiple independent variables and multiple dependent variables that is shown in Fig. 1.3. Eq. 1.14

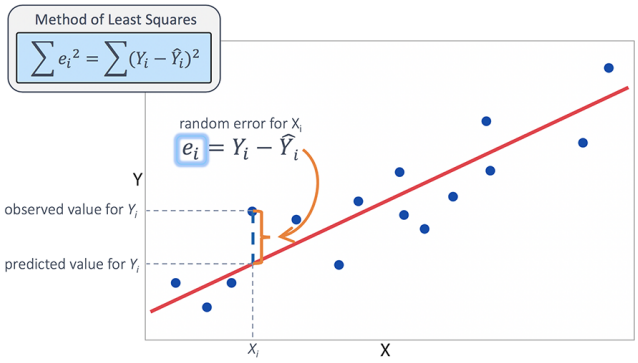

Figure 1.2: Error calculation in linear regression.

shows the estimation of parameters of MvLR. Please note that we can apply MLR: $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_q$ to solve the MvLR: $\hat{b} = (X^T X)^{-1} X^T y$ that shown in Eq. 1.15.

$$y_{n\times q} = X_{n\times(p+1)}\omega_{(p+1)\times q} + \varepsilon_{n\times q} \tag{1.14}$$

$$\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_q \rightarrow \hat{b} = (X^T X)^{-1} X^T y \tag{1.15}$$

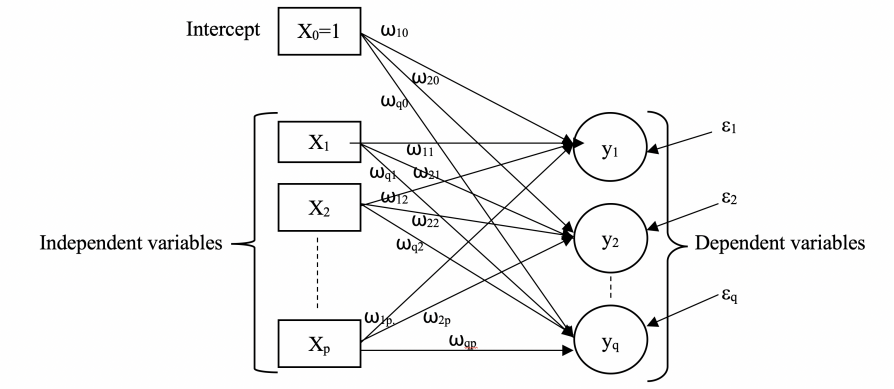

Figure 1.3: Multivariate linear regression example.

By applying multiple linear regression (MLR) we can have the following conceptual model for the multivariate multiple regression (MvLR), where y is the dependent variables, X is the independent variables, ε is the errors, and b is the regression cofficients:

$y_1 = \omega_{10} + \omega_{11}.X_1 + \omega_{12}.X_2 + \cdots + \omega_{1p}.X_p + \varepsilon_1$

$$y_2 = \omega_{20} + \omega_{21}.X_1 + \omega_{22}.X_2 + \cdots + \omega_{2p}.X_p + \varepsilon_2$$
$$y_q = \omega_{q0} + \omega_{q1}.X_1 + \omega_{q2}.X_2 + \cdots + \omega_{qp}.X_p + \varepsilon_q$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{bmatrix}_{q\times 1} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}_{p\times 1} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_q \end{bmatrix}_{q\times 1} \quad b = \begin{bmatrix} \omega_{10} & \omega_{20} & \dots & \omega_{q0} \\ \omega_{11} & \omega_{21} & \dots & \omega_{q1} \\ \vdots & \vdots & \vdots & \vdots \\ \omega_{1p} & \omega_{2p} & \dots & \omega_{qp} \end{bmatrix}_{(p+1)\times q}$$

Conceptual model data, where $y_{n\times q}$ is the dependent variables, and $X_{n\times(p+1)}$ is the design matrix with constant & independent variables:

$$y_{n\times q} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1q} \\ y_{21} & y_{22} & \dots & y_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ y_{i1} & y_{i2} & \dots & y_{iq} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nq} \end{bmatrix}_{n\times q} \quad X_{n\times(p+1)} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{i1} & X_{i2} & \dots & X_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}_{n\times(p+1)}$$

$$y_{i1} = \omega_{10} + \omega_{11}.X_{i1} + \omega_{12}.X_{i2} + \cdots + \omega_{1p}.X_{ip} + \varepsilon_{i1}$$
$$y_{i2} = \omega_{20} + \omega_{21}.X_{i1} + \omega_{22}.X_{i2} + \cdots + \omega_{2p}.X_{ip} + \varepsilon_{i2}$$
$$\vdots$$
$$y_{ik} = \omega_{k0} + \omega_{k1}.X_{i1} + \omega_{k2}.X_{i2} + \cdots + \omega_{kp}.X_{ip} + \varepsilon_{ik}$$
$$\vdots$$
$$y_{iq} = \omega_{q0} + \omega_{q1}.X_{i1} + \omega_{q2}.X_{i2} + \cdots + \omega_{qp}.X_{ip} + \varepsilon_{iq}$$

Where, $i = 1, 2, \dots, n$ observations. So, we can have the following general equation in terms of data for MvLR: $y_{n\times q} = X_{n\times(p+1)}\omega_{(p+1)\times q} + \varepsilon_{n\times q}$, which is Eq. 1.14. Please note that MLR $\rightarrow \varepsilon_{n\times 1}$ and MvLR $\rightarrow \varepsilon_{n\times q}$.

1.1.3.1 *Illiterate an Example of MvLR*

Given a data that is shown in Tabel 1.4 where we have two independent variables X_1, X_2, and two dependent variables y_1, y_2; Fig 1.4.

Table 1.4: MvLR example with two dependent variables.

X_1	X_2	y_1	y_2
9	62	10	100
8	58	12	110
7	64	11	105

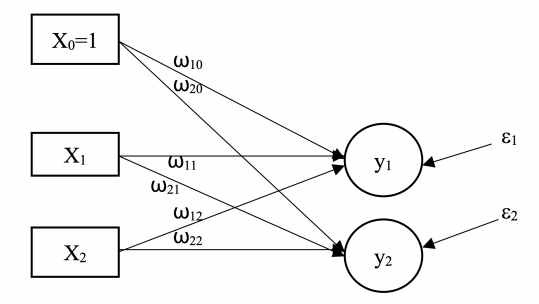

Figure 1.4: MvLR example with two IVs and two DVs.

$$y_{n\times q} = \begin{bmatrix} 10 & 100 \\ 12 & 110 \\ 11 & 105 \end{bmatrix}_{3\times 2} \qquad X_{n\times p} = \begin{bmatrix} 9 & 62 \\ 8 & 58 \\ 7 & 64 \end{bmatrix}_{3\times 2} \rightarrow X_{n\times(p+1)} = \begin{bmatrix} 1 & 9 & 62 \\ 1 & 8 & 58 \\ 1 & 7 & 64 \end{bmatrix}_{3\times 3}$$

Now, we need to find $\hat{b} = (X^T X)^{-1} X^T y$

Step 1: Compute $X^T X$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 9 & 8 & 7 \\ 62 & 58 & 64 \end{bmatrix} \begin{bmatrix} 1 & 9 & 62 \\ 1 & 8 & 58 \\ 1 & 7 & 64 \end{bmatrix} = \begin{bmatrix} 3 & 24 & 184 \\ 24 & 194 & 1470 \\ 184 & 1470 & 11304 \end{bmatrix}$$

Step 2: Compute $(X^T X)^{-1}$

$$(X^T X)^{-1} = \frac{1}{|X^T X|} abj(X^T X) = \begin{bmatrix} 320.76 & -8.16 & -4.216 \\ -8.16 & 0.56 & 0.06 \\ -4.16 & 0.06 & 0.06 \end{bmatrix}$$

Step 3: Compute $X^T y$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 \\ 9 & 8 & 7 \\ 62 & 58 & 64 \end{bmatrix} \begin{bmatrix} 10 & 100 \\ 12 & 110 \\ 11 & 105 \end{bmatrix} = \begin{bmatrix} 33 & 315 \\ 263 & 2515 \\ 2020 & 19300 \end{bmatrix}$$

Step 4: Compute b

$$b = \omega_1 \omega_2 = (X^T X)^{-1} X^T y = \begin{bmatrix} 35.8 & 229 \\ -0.8 & -4 \\ -0.3 & -1.5 \end{bmatrix}$$

$$y = Xb + \varepsilon \rightarrow \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{bmatrix} \hat{b}_{10} & \hat{b}_{20} \\ \hat{b}_{11} & \hat{b}_{21} \\ \hat{b}_{12} & \hat{b}_{22} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

$$y_1 = \hat{b}_{10} + \hat{b}_{11}X_1 + \hat{b}_{12}X_2 + \varepsilon_1 = 35.8 - 0.8X_1 - 0.3X_2 + \varepsilon_1$$
$$y_2 = \hat{b}_{20} + \hat{b}_{21}X_1 + \hat{b}_{22}X_2 + \varepsilon_2 = 229 - 4X_1 - 1.5X_2 + \varepsilon_2$$

1.2 Polynomial Regression

Polynomial Regression is a form of regression analysis in which the complex non-linear relationship between the independent variables and dependent variables are modelled in the n^{th} degree polynomial. The n^{th} order polynomial model in one variable is shown by Eq. 1.16. Polynomial regression models are usually fit with the method of least squares. For example: $y = \omega_0 + \omega_1 X + \omega_2 X^2$ is a polynomial regression model in one variable and is called a *second-order model* or *quadratic model*. The coefficients ω_1 and ω_2 are called the *linear effect parameter* and *quadratic effect parameter*, respectively.

$$y = \omega_0 X^0 + \omega_1 X^1 + \cdots + \omega_n X^n + \varepsilon \tag{1.16}$$

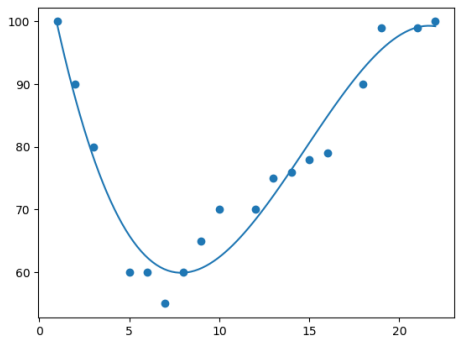

Figure 1.5: Polynomial regression.

1.2.1 Illiterate an Example of Polynomial Regression

Given a regression problem that shown in Table 1.5, we need to find the ω_0, ω_1, and ω_2 values for getting the best fitted line: $y = \omega_0 + \omega_1 x + \omega_2 x^2$.

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i.x \\ \sum y_i.x^2 \end{bmatrix}$$

Table 1.5: Polynomial regression example.

x	y	x^2	x^3	x^4	$y.x$	$y.x^2$
80	6.47	6400	512000	40960000	517.6	41408
40	6.24	1600	64000	2560000	249.6	9984
-40	5.72	1600	-64000	2560000	-228.8	9152
-120	5.09	14400	-1728000	207360000	-610.8	73296
-200	4.30	40000	-8000000	1600000000	-860	172000
-280	3.33	78400	-21952000	6146560000	-932.4	261072
$\sum x =$	$\sum y =$	$\sum x^2 =$	$\sum x^3 =$	$\sum x^4 =$	$\sum y.x =$	$\sum y.x^2 =$
-520	31.15	1.424×10^5	-3.117×10^7	8×10^9	-1.865×10^3	5.669×10^5

$$\begin{bmatrix} 6 & -520 & 1.424 \times 10^5 \\ -520 & 1.424 \times 10^5 & -3.117 \times 10^7 \\ 1.424 \times 10^5 & -3.117 \times 10^7 & 8 \times 10^9 \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{bmatrix} = \begin{bmatrix} 31.15 \\ -1.865 \times 10^3 \\ 5.669 \times 10^5 \end{bmatrix}$$

Therefore, $\omega_0 = 6.013, \omega_1 = 6.424 \times 10^{-3}$, and $\omega_2 = -1.113 \times 10^{-5}$
So, $y = 6.013 + 6.424 \times 10^{-3}.x - 1.113 \times 10^{-5}.x^2$
And, $y(70) = 6.013 + 6.424 \times 10^{-3}(70) - 1.113 \times 10^{-5}(70)^2 = 6.408$

1.3 Evaluation Matrix for Regression Model

After building a linear regression model from the training data, we need to test how well the model fits the data. There are several key goodness-of-fit techniques for regression analysis, e.g. R-squared (R^2), Mean Squared Error (MSE) or Mean Squared Deviation (MSD), and Mean Absolute Error (MAE) etc.

R-squared (R^2) is the coefficient of determination that presents the proportion of the variance for a dependent variable, y that's explained by an independent variable, x or variables $X_1, X_2, \ldots, X_p$ in a regression model. It evaluates the scatter of the data points around the fitted regression line. R^2 values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. An R^2 of 0% represents a model that does not explain any of the variation in the response variable around its mean, and 100% represents a model that explains all the variation in the response variable around its mean. A high (R^2), between 85% and 100%, indicates that the model performance moves relatively in line of best fit. A low (R^2), at 70% or less, indicates the model does not generally follow the movements of the line of best fit. The formula for R^2 is $(1 - \frac{Unexplained\ Variation}{Total\ Variation})$ that is shown in Eq. 1.17, where $\bar{y}$ is the mean of the observed data: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the residuals as $e_i = y_i - \hat{y}_i$

$$
\begin{aligned}
R^2 &= 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} \\
&= 1 - \frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - \hat{y_i})^2} \\
&= 1 - \frac{\sum_i (y_i - \bar{y})^2}{\sum_i e_i^2}
\end{aligned} \tag{1.17}
$$

Mean Squared Error (MSE) or mean squared deviation (MSD) measures the average of the squares of the errors that's the average squared difference between the predicted values, $\hat{y}$ and the actual value, y. The MSE is the mean $\left(\frac{1}{n}\sum_{i=1}^{n}\right)$ of the squares of the errors $(y_i - \hat{y_i})^2$ that's shown in Eq. 1.18 where n = number of data points/instances, y_i = observed values, and $\hat{y}_i$ = predicted values.

$$
MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{1.18}
$$

Mean Absolute Error (MAE) measures accuracy for continuous variables by finding the average of the absolute errors $|e_i| = |\hat{y_i} - y_i|$, where $\hat{y_i}$ is the prediction and y_i the true value. The absolute error is the absolute value of the difference between the predicted/forecasted value and the actual value. MAE tells us how big of an error we can expect from the forecast on average. The MAE is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$
MAE = \frac{\sum_{i=1}^{n} |\hat{y_i} - y_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{1.19}
$$

1.4 Logistic Regression

Logistic Regression (or logit regression) is a binary-class classification method used in supervised learning to predict the dichotomous (binary) dependent variable (such that the dependent variable is categorical) on a given set of nominal, ordinal, interval or ratio-level independent variables. It uses a logistic function to model the probability of a binary dependent variable, e.g., a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used for solving binary-class classification problems. The natural logarithm of the odds/odds ratio $= \frac{p}{1-p}$ is equivalent to a *linear* function of the independent variables. The probability, $P = \frac{odds}{1+odds}$ ranges from 0 to 1, however Log Odds from $-\alpha$ to $+\alpha$. The antilog of the logit function allows us to find the estimated regression equation, $\hat{p} = \frac{e^{y^*}}{1+e^{y^*}}$ that shown in Eq. 1.20. Logistic regression can't handle non-linearities in data.

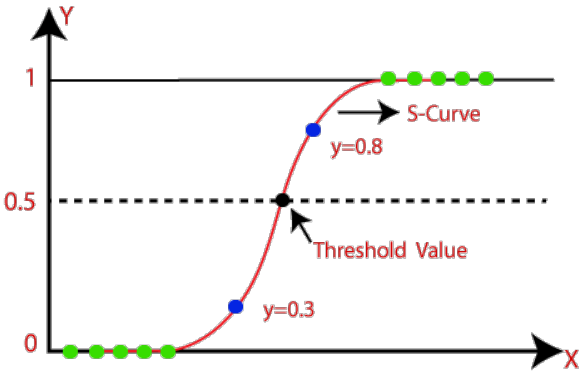

Figure 1.6: Logistic regression.

$$
\begin{aligned}
y^* = logit(p) = ln\Big(\frac{p}{1-p}\Big) &= \omega_0 + \omega_1.X_1 \\
Antilog \Rightarrow \frac{p}{1-p} = exp^{\omega_0+\omega_1.X_1} &= e^{\omega_0+\omega_1.X_1} \\
p &= e^{\omega_0+\omega_1.X_1}(1-p) \\
p &= e^{\omega_0+\omega_1.X_1} - e^{\omega_0+\omega_1.X_1} \times p \\
p + e^{\omega_0+\omega_1.X_1} \times p &= e^{\omega_0+\omega_1.X_1} \\
p(1 + e^{\omega_0+\omega_1.X_1}) &= e^{\omega_0+\omega_1.X_1} \\
\hat{p} = \frac{e^{\omega_0+\omega_1.X_1}}{1+e^{\omega_0+\omega_1.X_1}} = \frac{exp^{y^*}}{1+exp^{y^*}} &= \frac{e^{y^*}}{1+e^{y^*}}
\end{aligned}
\tag{1.20}
$$

Table 1.6: Linear Regression vs Logistic Regression

	Linear Regression	Logistic Regression
Method	Predicting continuous variable	Predicting categorical variable
Dependent variable type	Continuous	Categorical
Estimation method	Least square estimation	Maximum like-hood estimation
Equation	$y = \omega_0 + \omega_1.x + \varepsilon$	$log\Big(\frac{y}{1-y}\Big) = \sum_{i=1}^{n} \omega_i.X_i + \varepsilon$
Best fitted line	Straight line	Curve
Linear Relationship $(x^{(i)}, y^{(i)})$	Needed	Not mandatory
Output	Predict integer value	Predict binary value
Applications	Forecasting problems	Classification problems

1.4.1 Sigmoid Function

Sigmoid is a mathematical function that takes any real number and maps it to a probability between 1 and 0. The sigmoid function forms an S shaped graph, which means as x approaches infinity, the probability becomes 1, and as x approaches negative infinity, the probability becomes 0. A common example of a sigmoid function is the logistic function that shown in Fig. 1.21.

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x) \tag{1.21}$$

1.4.2 Multinomial Logistic Regression

Multinomial Logistic Regression is a multi-class classification method that generalises logistic regression to multi-class tasks, i.e. with more than two possible discrete outcomes of dependent variable. We can solve the multi-class classification tasks using logistic regression by employing **One-vs-all** (or one-vs-rest) technique.

1.4.2.1 One-vs-All Classification One-vs-All (or One-vs-Rest) classification is used for using binary classification algorithms, e.g., logistic regression and SVM for multi-class classification tasks. It splits the multi-class classification dataset into multiple binary classification datasets and fits a binary classification model on each binary classification dataset (Fig. 1.7).

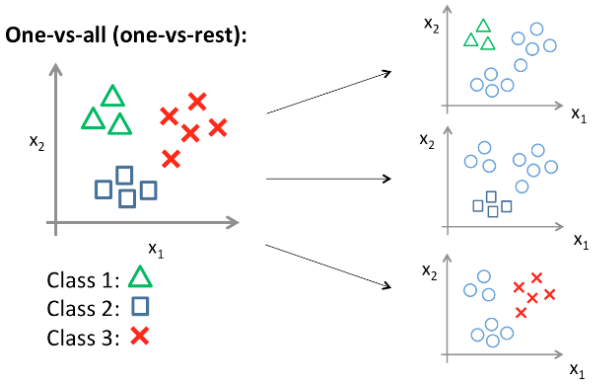

Figure 1.7: One-vs-All classification.

1.5 Ensemble Learning

1.5.1 Adaptive Boosting

Table 1.7: Weight updating process in adaptive boosting technique.

Instance, x	Weight, w	Prediction	Accuracy	Error	$w_{old} \times \frac{Error}{Accuracy}$	$w_{new} \times \frac{\sum w_{old}}{\sum w_{new}}$
x_1	$w_1 = 1$	Classified	1	0	$1 \times 0.66 = 0.66$	$0.66 \times 1.25 = 0.825$
x_2	$w_2 = 1$	Misclassified	0	1	1	$1 \times 1.25 = 1.25$
x_3	$w_3 = 1$	Classified	1	0	$1 \times 0.66 = 0.66$	$0.66 \times 1.25 = 0.825$
x_4	$w_4 = 1$	Misclassified	0	1	1	$1 \times 1.25 = 1.25$
x_5	$w_5 = 1$	Classified	1	0	$1 \times 0.66 = 0.66$	$0.66 \times 1.25 = 0.825$
	$\sum w_{old} = 5$		3/5= 0.6	2/5=0.4	$\sum w_{new} = 3.98$	**Updated weights**

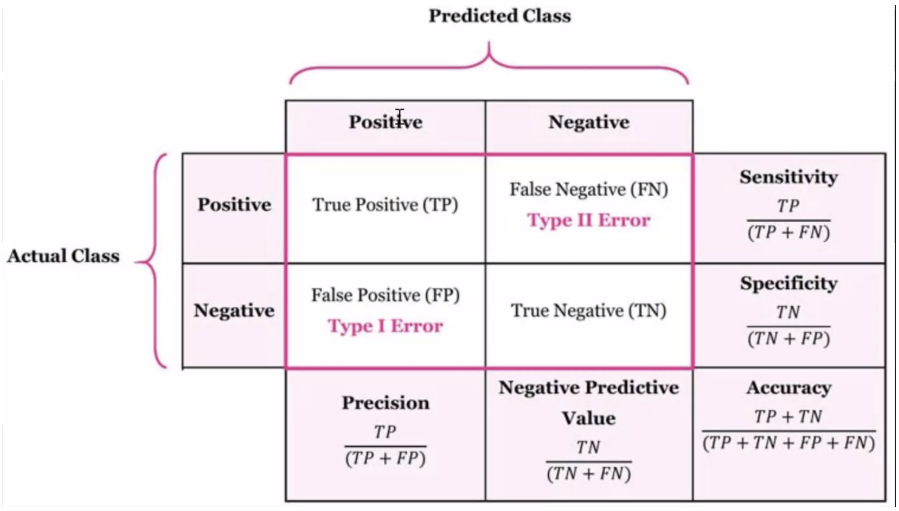

Figure 1.8: Confusion Matrix.