Applied Machine Learning

Unit-2: Regression Techniques

1

Unit 2: Syllabus

Unit-II: Regression Techniques:

- Basic Concepts and applications of Regression,
- Simple Linear Regression – Gradient Descent and Normal Equation Method,
- Multiple Linear Regression,
- Non-Linear Regression,
- Linear Regression with Regularization, Hyper-parameters tuning, Loss Functions,
- Decision Tree Regression, Evaluation Measures for Regression Techniques

Unit 2: Definition of Probability

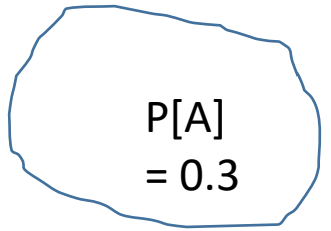
- When an experiment is performed many times, many outcome events are possible.
- Then the probability P[A] is possibility of a some event, e.g. A, occurring in each experimental trial.
- Definition of Probability: P[A] = f/N, where f is frequency of occurrence of event A and N is number of times the experiment is performed.
- This is known as relative frequency approximation to probability.
- In this case, the experiment must be repeatable. Each experiment should be independent of previous experiment and unbiased.
- The value of P[A] is approximate as it depends on the value of N, the number of times of experiment.
- As N -> ∞, P[A] converges asymptotically to true probability value.
- Example 1: Tossing of a well balanced coin. As N -> ∞, P[A] -> 0.5 (A: Head/Tail)
- Example 2: Throwing of a dice. In this case, P[A] = 1/6 = 0.166… (A: Any Face of Dice)

Unit 2: Definition of Probability

- **Classical formula of probability:**

  P[A] = n[A]/n[S], where n[A] = no of ways event A can occur and n[S] = total no of times experiment S can proceed.

- **If outcomes are truly equally likely, then P[A] is not approximate, but it is an accurate description of occurring of event A.**
- **Some events do not require experimentation. Probability can be estimated theoretically. Consider the Permutation and combination experiment.**
- **Permutation is number of all possible ways in which a set of objects (m) are arranged in sequences from n objects; denoted by nPm =n!/(n-m)!**
- **Combinations is the non-repetitive ways of arranging a set of m objects in sequences from n objects; denoted by nCm = n!/{m! (n-m)!}**
- **In permutations and combinations, the number of ways (Probability of occurrence) a pattern can occur is known beforehand. In this probability is known a prior i.e. before hand.**
- **P[] can range from 0 to 1 (or 0 to 100 %). Conceptually, the probability is represented as Venn Diagram.**
- **Venn diagram is used to represent the probability.**



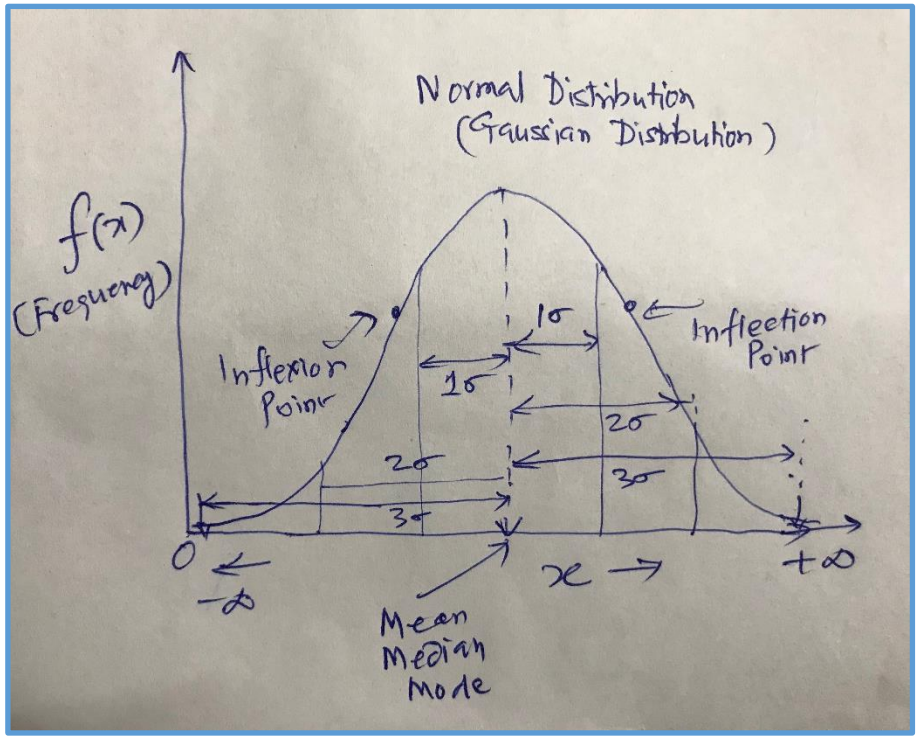
Venn Diagram.

Unit 2: Normal Distribution

- **Mean, Variance and Standard Deviation:**
  - The variability in sample data are expressed by term known as Variance.
  - When an outcome of a random event is measured multiple times, the outcomes show spread around a certain value known as a central value.
  - It is almost similar to the arithmetic mean of outcomes of events.
  - Larger the number of multiple measurements (i.e. events), the more the well defined is the central value.
  - This value is known as the mean of the random variable.
  - The spread in the measurements around the central tendency value (i.e. mean) is described by a term known as sample variance, and Standard Deviation.
  - It is defined as $S^2 = \frac{\Sigma (x_i - \overline{X})^2}{n-1}$, where i = 1 to n. The $\overline{X}$ is mean of x_i
  - The Standard Deviation S is the positive square root of Variance S^2.
  - Here X is variable name and x_i are the values the variable X assumes.

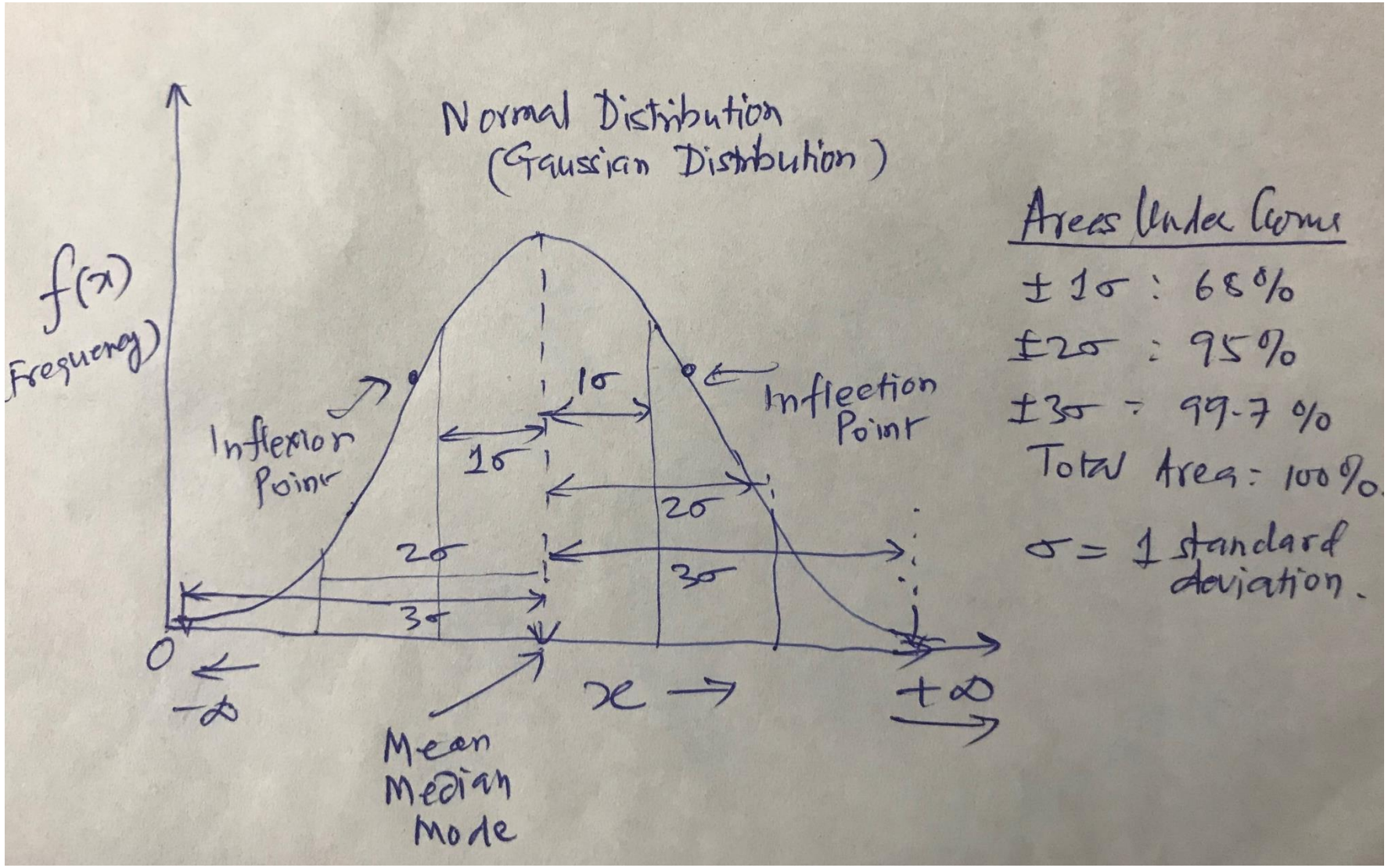
Unit 2: Normal Distribution

- The Normal distribution is also known as Gaussian distribution.
- If f(x) is the probability of X variable taking a x, then the probability distribution of x is given by:
- N.f(x) = $\frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)^2/\sigma 2]}$
- It can be verified that $\int_{-\infty}^{+\infty} f(x)dx = 1$.
- In fact, the $\frac{1}{\sigma\sqrt{2\pi}}$ factor is a normalizing factor to make integration equation = 1.



- Properties of Normal Distribution Function.
  - It is symmetrical both sides of mean value (μ)
  - It has bell shaped curve
  - It extends from –∞ to + ∞ of x values.
  - The area under the entire curve from -∞ to +∞ is equal to 1.
  - There two inflection points (points of change in sign of slope) at 0.75μ and 1.25μ.

Unit 2: Normal Distribution



Unit 2: Normal Distribution

Variance and Standard Deviation:

- The integration $\int_{-\infty}^{+\infty} f(x)dx = 1.$
- The f(x) is probability density function that a random number taking a value of x. It is the probability of a random variable takes value of x.
- Equality to 1 Indicates that the random number will definitely assume a value in range of -∞ to + ∞
- For normal distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)^2/\sigma 2]}$
- Similar equation can be written for sample mean $\overline{X}$ and sample standard deviation S.
- The area under the curve from x_1 to x_2 gives the probability the f(x) will assume a values in this interval.

Unit 2: Normal Distribution

- **Properties of Normal Distribution**:
  - It is symmetrical both sides of μ
  - It has bell shaped curve
  - The area under the entire curve from -∞ to +∞ is equal to 1.
  - Area within -1σ to +1σ is 68 %
  - And -2σ to +2σ it is 95 %
  - And -3σ to +3σ it is 99.7 %.
  - There two inflection points (points of change in sign of slope) at 0.75μ and 1.25μ.
  - The mean, mode and median all are same.

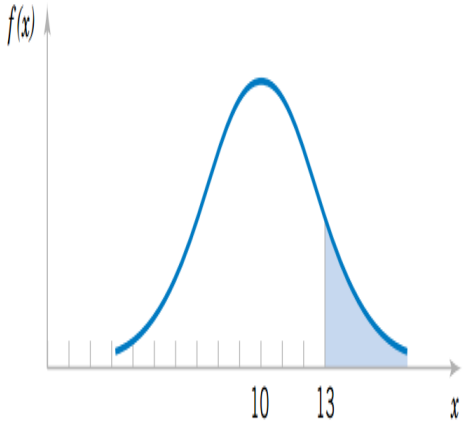


Figure 4-11 Probability that $X > 13$ for a normal random variable with $\mu = 10$ and $\sigma^2 = 4$.

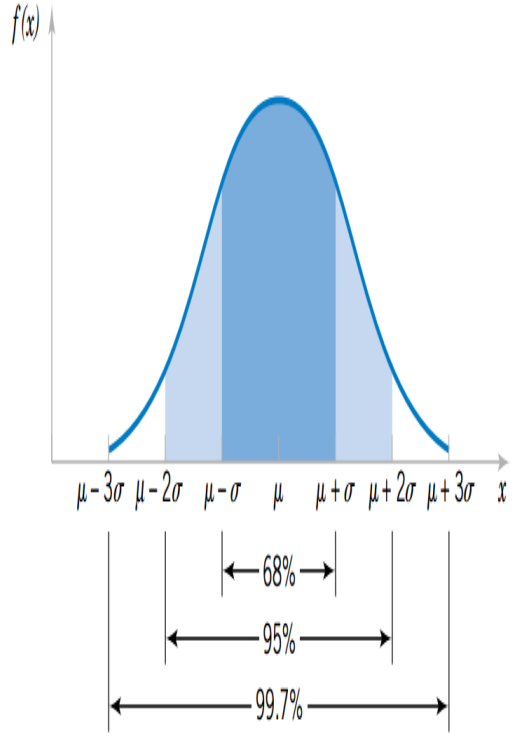


Figure 4-12 Probabilities associated with a normal distribution.

Unit 2: Normal Distribution

- It is mostly commonly followed distribution of many processes.
- Many statistical models of data analysis are based on this distribution.
- It is a limiting form of Binomial distribution, as number of Bernoulli trials increases to ultimately to become infinity.
- This distribution is also known as Gaussian Distribution. But commonly known as Normal Distribution.
- Definition: A random variable X with density function given by

  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)^2/\sigma 2]}$ and number of events with X= x and total number of events N_0 is given by $N(x) = N_o f(x)$.
- It is known as Normal Distribution with $\mathrm{mean}\ \mu$ and standard deviation σ parameters.
- Both of them range from $-\infty$ to $+\infty$.
- By definition σ is positive root of variance.

Unit 2: Standard Normal Distribution

- It is a normal distribution with $\mu = 0$ and $\sigma = 1$.
- Then, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x)^2}$ This is Standard Normal Distribution.
- Here, the $\frac{1}{\sqrt{2\pi}}$ is a normalizing factor to make integration equation = 1.
- The statistical lookup tables for normal distribution are based this standard normal distribution.
- The variable given by $(x-\mu)/\sigma$ is known as Standard Normal variable and is denoted by z.
- Normal Approximation to Binomial Distribution:
  - If X is a Binomial Variable with parameters n number of Bernoulli trials and p is the probability of the Bernoulli event, then for large n, X shows approximately normal distribution with mean = np and variance = np(1-p).
  - Largely True, if $p \leq 0.5$ and $np > 5$ or $p > 0.5$ and $n(1-p) > 5$

Unit 2: Log-Normal Distribution

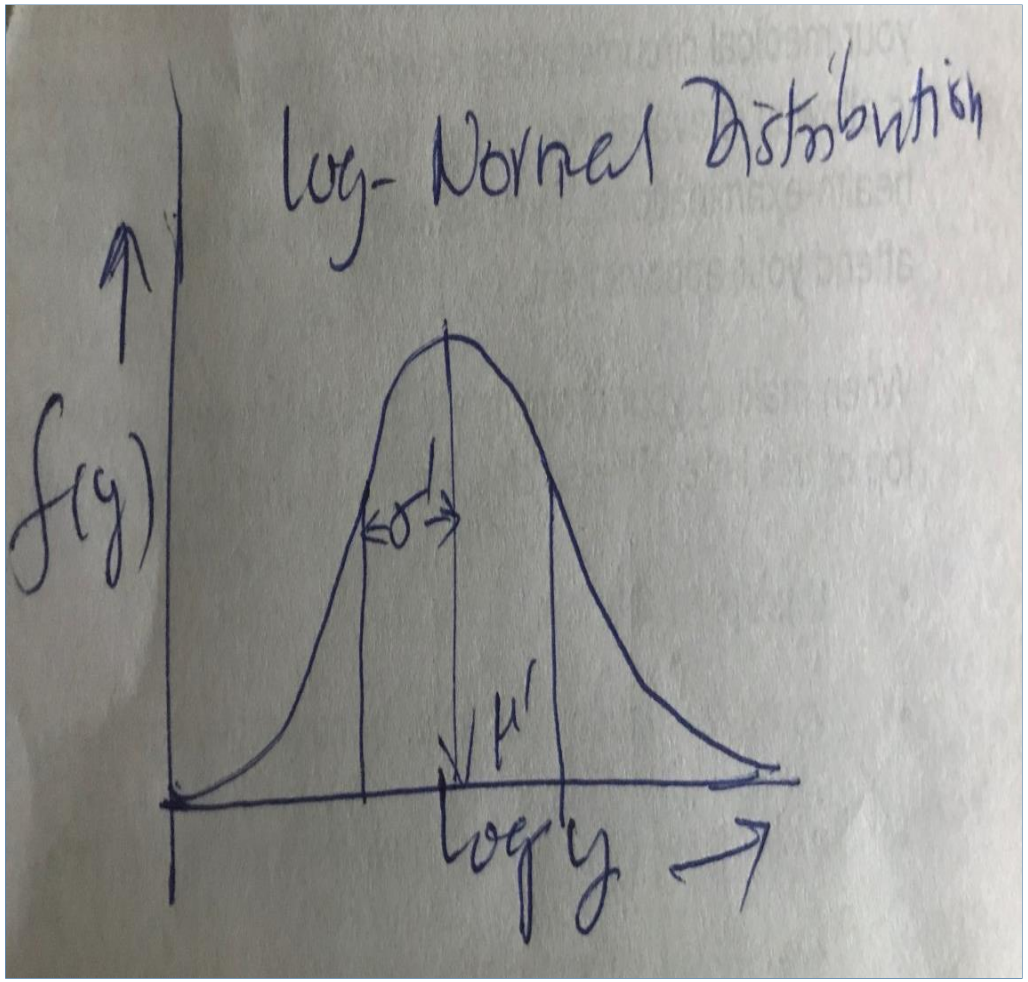
- We know that the normal distribution with mean μ and variance σ^2 is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[(x-\mu)^2/\sigma 2]}$$

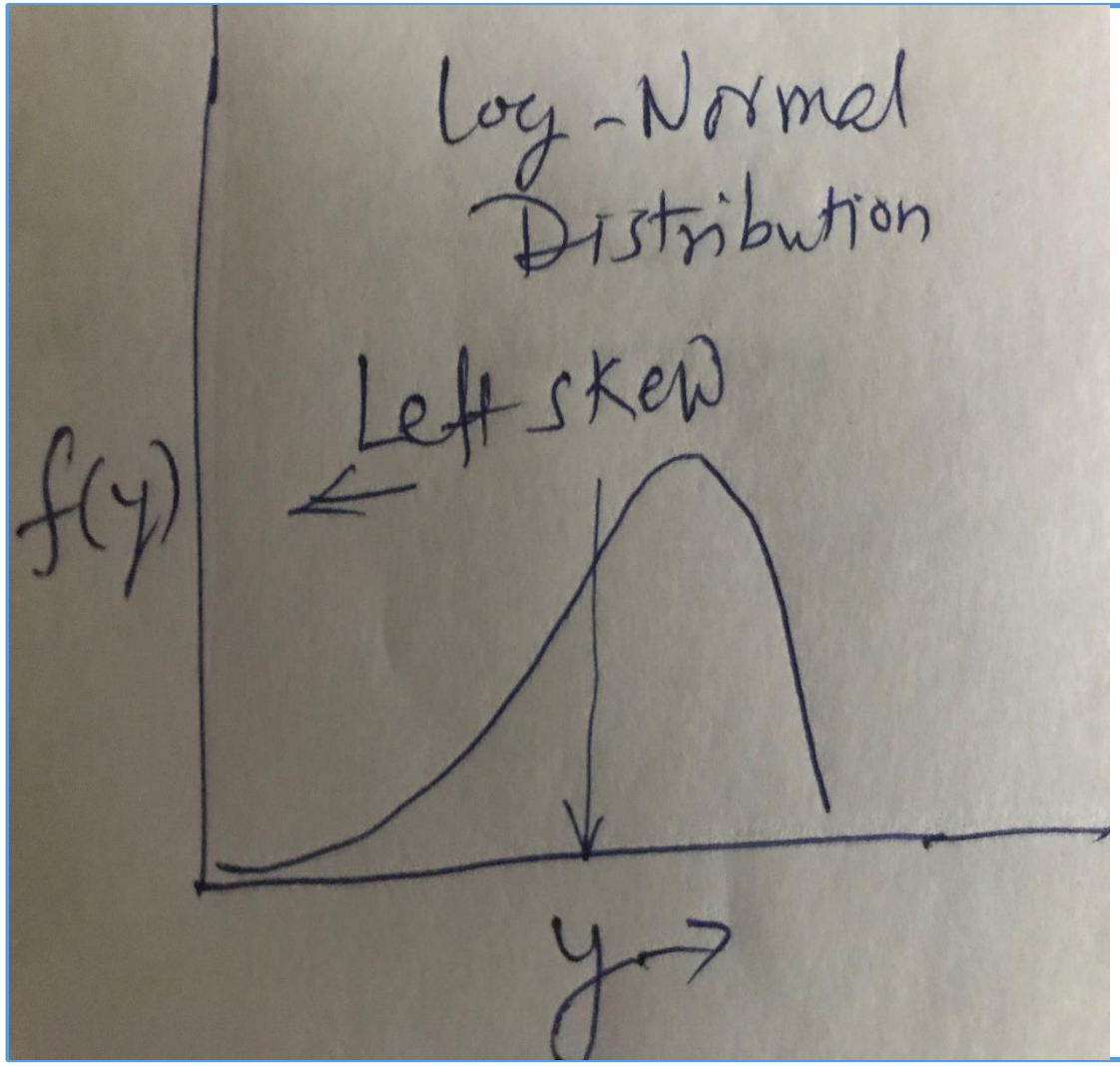
- If **x = log (y),** then the Normal Distribution in y can be written with different mean μ′ and variance σ′2 as:

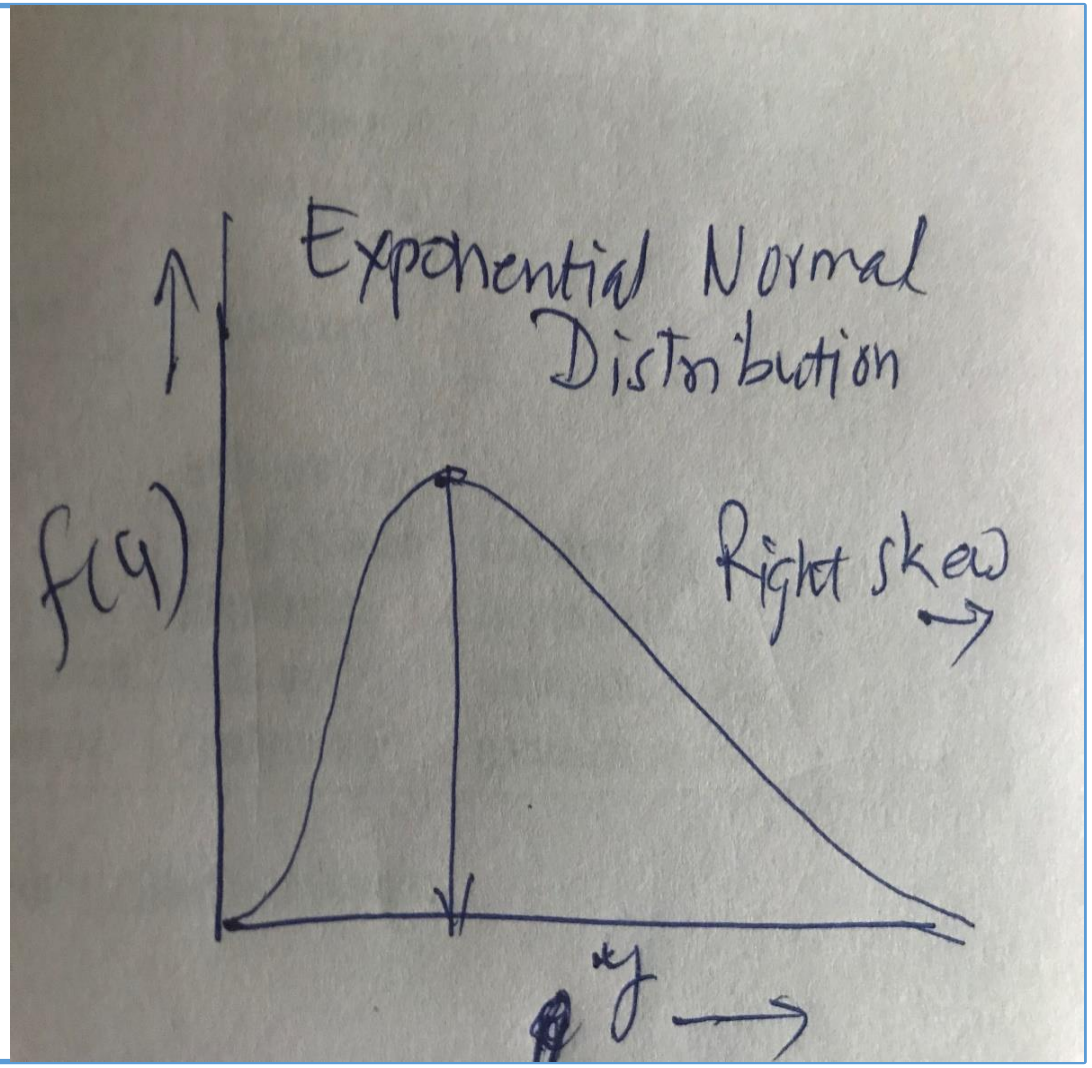
$$f(y) = \frac{1}{\sigma'\sqrt{2\pi}} e^{-\frac{1}{2}[(\log(y)-\mu')^2/\sigma' 2]}$$

- This distribution is known as Log-Normal Distribution of x.
- If plotted against y, it looks like left side skewed curve.
- If **x = e^y**, then similar curve will show right side skewness.



Unit 2: Log-Normal & Exponential Distribution





Applied Machine Learning

Unit-2: Regression Techniques

2

Unit 2:Bivariate Normal Distribution

- We have seen the normal distribution with one independent variable.
- There can be normal distribution with 2 or more variables.
- The density function of, say two variables, is represented by f(x,y). And that of multiple variables represented by f(x,y,z,…).
- If the density function f(x,y) of x and y, which are both independent variables, then f(x,y) will appear as bell shaped surface, with both x and y as horizontal axes and f(x,y) as vertical axis.

Unit 2:Bivariate Normal Distribution

- **The f(x,y) distribution is known as bivariate normal distribution.**
- **It is defined as:**

$$f(x,y) = \frac{e^{\{-\left[\frac{1}{2(1-\rho^2)}\right]*[X_N - YN]^2\}}}{2\pi*\sqrt{1-\rho^2}*\sigma_x*\sigma_y}$$

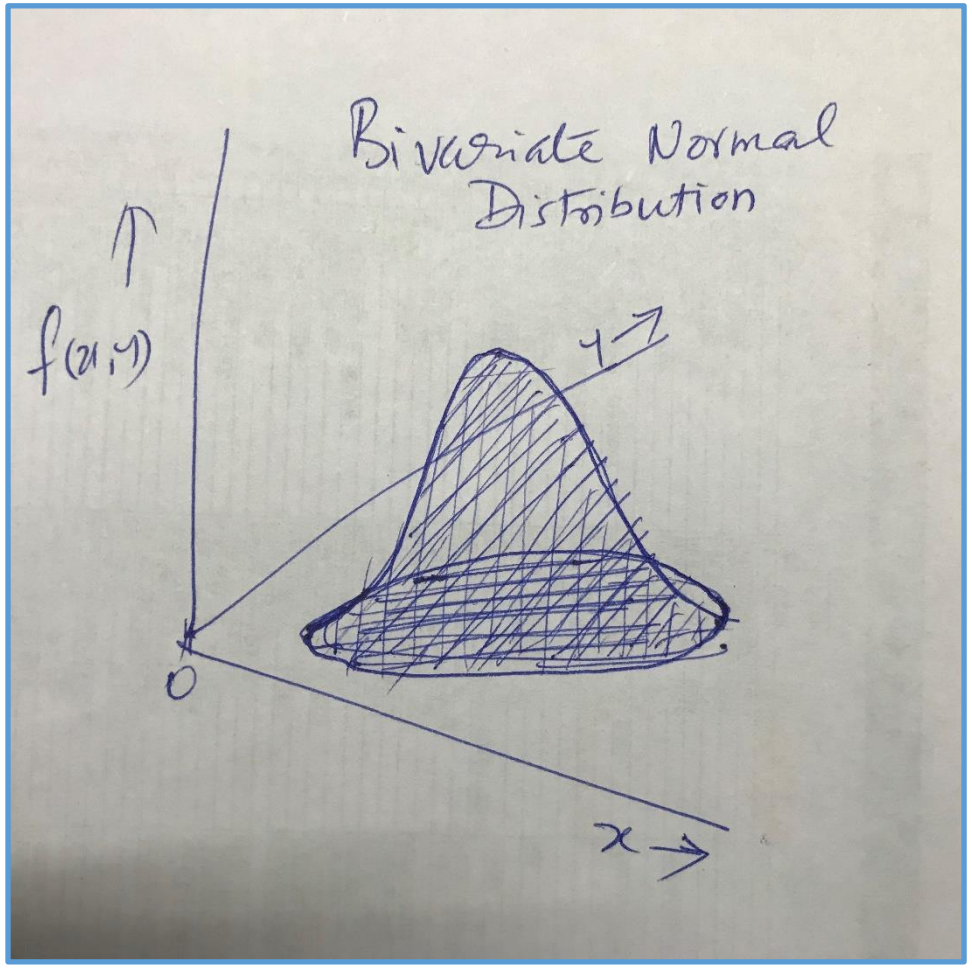
where $x_N = \frac{x-\mu_x}{\sigma_x}$, $y_N = \frac{y-\mu_N}{\sigma_y}$ **and**

ρ **= correlation coefficient between the bivariates x and y.**

- **If** ρ **= 0, then x and y are independent, i.e. they are not correlated, and**

If ρ **= 1, then they are perfectly correlated**

If $0 < |\rho| < 1$, **the they are partially correlated.**



Unit 2: Co-variance and Correlation

- Sometimes, it is of interest to know whether one or two random variables are linearly related.
- That is, one of the variable (independent one) is the cause of changes in the other variable (dependent variable).
- Therefore, they are called co-related to each other,

  i.e. correlated with each other.
- One measure to quantify this correlation is **Pearson Correlation Coefficient, ρ**.

Unit 2: Co-variance and Correlation

- **Definition of ρ (correlation coefficient):**
  IF X and Y are random variables with their means μ_x and μ_y, and variance σ^2 and σ'^2, respectively, then the Pearson Correlation Coefficient ρ_{xy} is defined as:

$$\rho_{xy} = \frac{cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

- **The cov is co-variance between X and Y, and is given by:**

$$cov(X,Y) = \frac{\sum_{all\ x,y} [(x_i - \mu_x) * (y_i - \mu_y)]}{n}$$ **where n is number of observations.**

- **And the var(X) and var(Y) are given by:**

$$var(X) = \frac{\sum_x [(x_i - \mu_x)^2]}{n}$$ **and similarly** $$var(Y) = \frac{\sum_y [(y_i - \mu_y)^2]}{n}$$

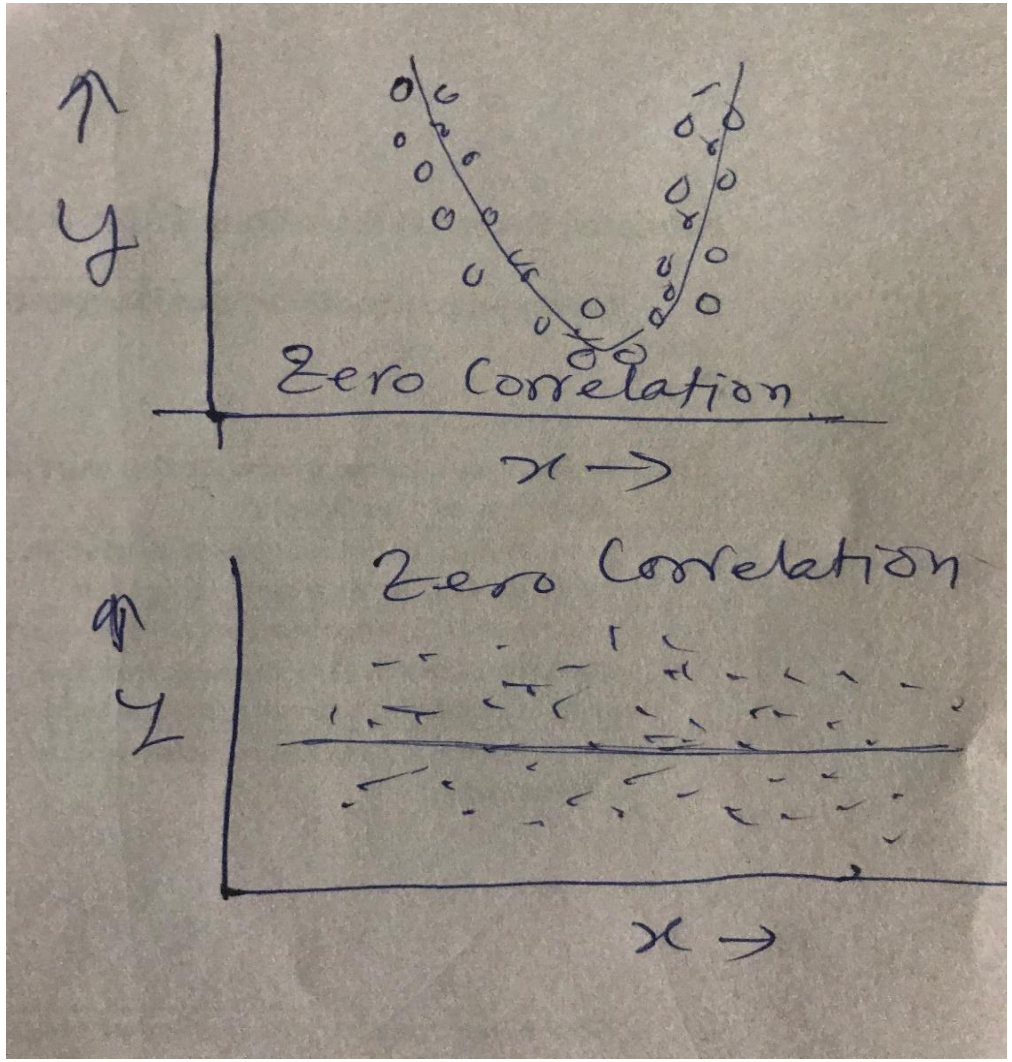
- **Therefore,** $$\rho_{xy} = \frac{\sum_{all\ x,y} [(x_i - \mu_x) * (y_i - \mu_y)]}{\sqrt{\{\sum_x [(x_i - \mu_x)^2] * \sum_y [(y_i - \mu_y)^2]\}}}$$

- **Note: The square root in the denominator encompasses all terms**

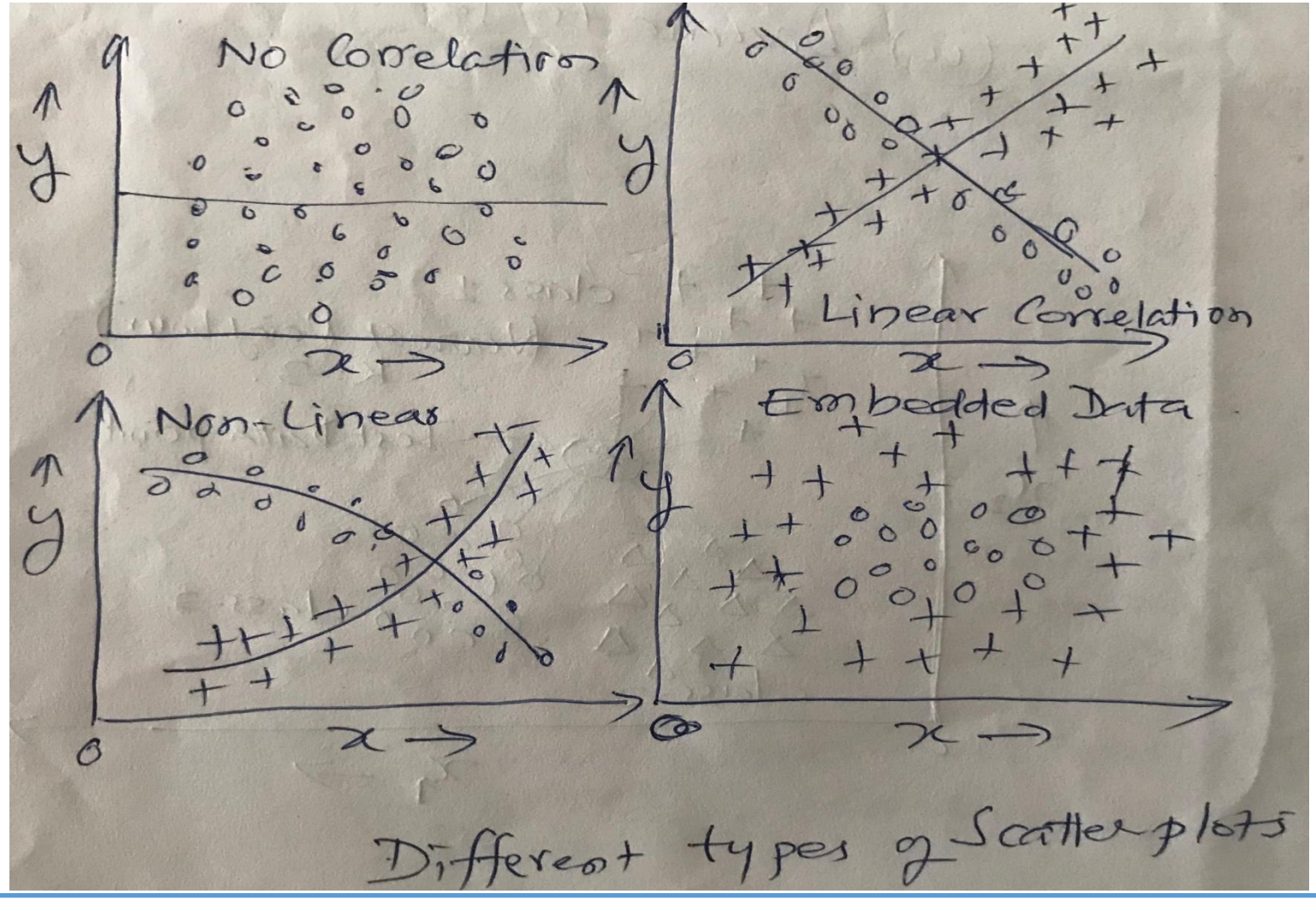
Unit 2: Co-variance and Correlation

Properties of Correlation Coefficient:

- The ρ_{xy} for any two random variables X and Y lies between -1 to +1 , through 0.
- ρ_{xy} = -1, then they are perfectly inversely correlated. $Y = C_0 - C_1 * X$
- ρ_{xy} = +1, then they are perfectly directly correlated. $Y = C_0 + C_1 * X$
- ρ_{xy} = 0, then they are NOT at all correlated/related. $Y = C_0$



Unit2: Scatterplots



Unit 2: Regression Techniques

- In the discussion on Pearson Correlation so far, both X and y variables were independent random variables.
- But what if they are not independent random variables?
- There would exit cause-effect relationship between them.
- For Example, The temperature of lake water and depth of lake (or even ocean). There would be relation between them.
- If several temperature measurements are made at any depth, such measured data will vary in values.

Unit 2: Regression Techniques

- In the above example, for a given depth x, we are dealing with conditional random variable y, denoted by y|x.
- Such a set of measurements at depth x will have a mean $\mu_{y|x}$.
- The plot of this function $\mu_{y|x}$ vs x is called the Curve of Regression of Y on X.
- Here, y is an dependent or response variable on X. X being an independent variable or the regressor variable.

Unit 2: Regression Techniques

- When the relation between y and X is linear, it is called Linear Regression in single variable.
- The Simple Linear regression equation is given by:

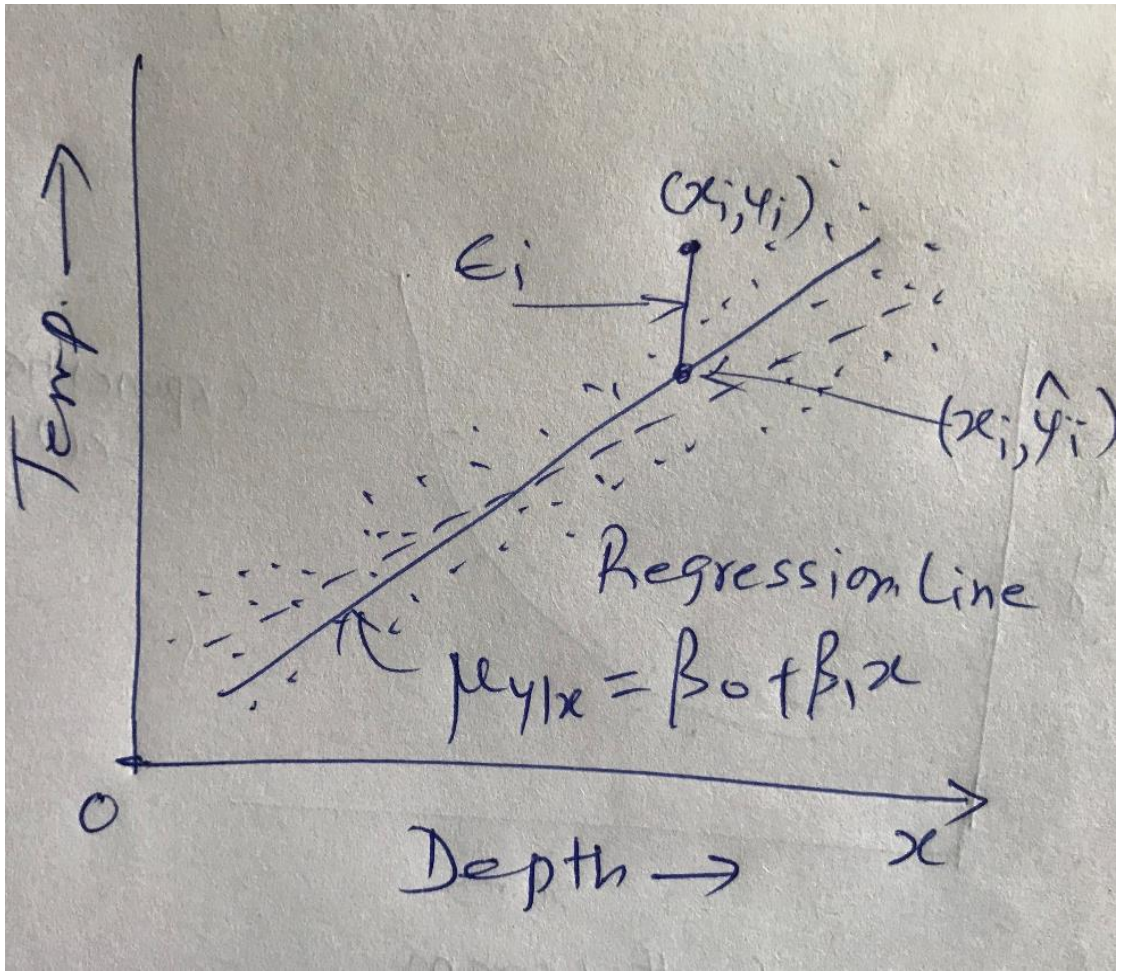
  $\mu_{y|x} = \beta_0 + \beta_1 X$, where β_0 and β_1 are real numbers and they are called regression coefficients.
- In regression study, X stands for x_i attributes, and i = 1 to m.
- Very important, if X are assumed to be known without error, such a study is known as Controlled Study.
- In controlled study, there are errors only in the $\mu_{y|x}$, as it has random errors.

Unit 2: Regression Techniques

- Therefore, the simple Linear Regression Model would be:

  $y_i = \beta_0 + \beta_1 x_i + e_i$

  where e_i is random error term in y values.
- It is a component with mean of e = 0.
- The e_i are known as Residual Errors.
- A plot of y_i vs x_i is known as scatter plot.
- Such plots show the trend in the relationship between y and X.



Applied Machine Learning

Unit-2: Regression Techniques

3

Unit 2: Regression Techniques (R)

- When the relation between y and X is linear, it is called Linear Regression in single variable.
- The Simple Linear regression equation is given by:

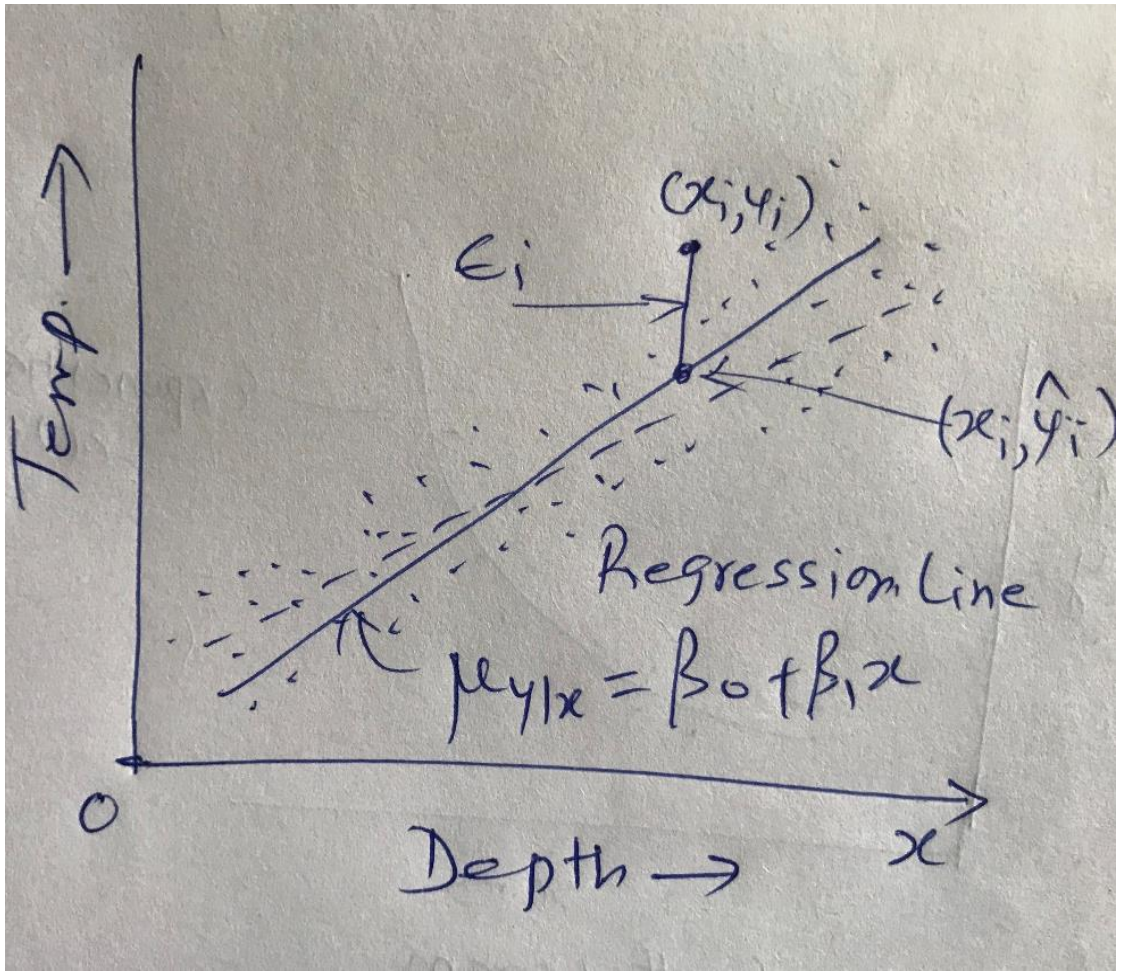
  $\mu_{y|x} = \beta_0 + \beta_1 X$, where β_0 and β_1 are real numbers and they are called regression coefficients.
- In regression study, X stands for x_i attributes, and i = 1 to m.
- Very important, if X are assumed to be known without error, such a study is known as Controlled Study.
- In controlled study, there are errors only in the $\mu_{y|x}$, as it has random errors.

Unit 2: Regression Techniques (R)

- Therefore, the simple Linear Regression Model would be:

  $y_i = \beta_0 + \beta_1 x_i + e_i$

  where e_i is random error term in y values.
- It is a component with mean of e = 0.
- The e_i are known as Residual Errors.
- A plot of y_i vs x_i is known as scatter plot.
- Such plots show the trend in the relationship between y and X.



Unit 2: Regression Techniques

Determination of Regression Line: Least Square Fitting Method.

- The parameters or coefficients β_0 and β_1 are estimated by the Least Square Fitting method.
- The Least Square Fitted line is the best fit line which gives the Least Sum of Squares of deviations/errors of estimated point value from the actual data point value.
- For each example/sample, we can write

  $y_i = \beta_0 + \beta_1 x_i + e_i$. Here index i goes from 0 to n or 1 to n.
- Then the error term is expressed as $e_i = y_i - \beta_0 - \beta_1 x_i$
- Here e_i will have both +ve as well as –ve values.
- If e_i are summed over all examples, then this sum is likely to be near zero.
- This type of sum does not gives the clear picture of deviation of regression line. Other way would be to use absolute value of e_i.
- Therefore, the Sum of Squares of Errors (SSE) are considered, as it has some useful mathematical properties.
- Note: SSE is biased towards outliers, whereas the absolute sum is not.

Unit 2: Regression Techniques

- The SSE is then given by:

  $SSE = \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

- The Least Square Fit minimises the SSE by finding such coefficient values of β_0 and β_1.
- The standard minimisation method is to equate first partial derivatives of SEE with respect to β_0 and β_1 and equate them to zero.

  $\delta(SSE)/\delta(\beta_0) = -2 \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$ and

  $\delta(SSE)/\delta(\beta_1) = -2 \sum_{i=0}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0$

- These equations reduce to: (By replacing $\sum_{i=0}^{n}$ by just Σ)

  $\Sigma y_i - n\beta_0 - \beta_1 \Sigma x_i = 0$

  $\Sigma x_i y_i - \beta_0 \Sigma x_i - \beta_1 \Sigma x_i^2 = 0$

- These above equations are known as **NORMAL EQUATIONS** which have to be solved for two unknowns β_0 and β_1.

Unit 2: Regression Techniques

- After solving the two Normal Equations for the unknowns β_0 and β_1, we get:

  $\beta_0 = (\Sigma y_i/n) - \beta_1\overline{X} = \overline{Y} - \beta_1\overline{X}$ and

  $\beta_1 = (n\Sigma x_i y_i - \Sigma x_i \Sigma y_i)/(n\Sigma x_i^2 - (\Sigma x_i)^2)$

- Note: Even though the regression equation actually estimates the mean value of Y for a given X, it is used extensively to estimate value of Y itself.
- Therefore, $y = \mu_{y|x} = \beta_0 + \beta_1 x$
- In such case, the random errors e_i have mean = 0 and variance σ^2,

  i.e. $e_i \sim N(0, \sigma^2)$ and variance is given by: $\sigma^2 = SSE/(n-2)$.

Unit 2: Regression Techniques

Some points to remember:

- The random variable y_i are independent and are normally distributed.
- The mean of y_i is given by $\beta_0 + \beta_1 x_i$
- The variance of y_i is σ^2 given by SSE/(n-2).
- The correlation coefficient R between Y and X is given by:

$$R = \left(\Sigma(xi - \bar{x})*(y_i - \bar{y})\right) / \sqrt{\left(\Sigma(xi - \bar{x})^2 * \Sigma(y_i - \bar{y})^2\right)}$$

$$= \frac{\Sigma(xi - \bar{x})*(y_i - \bar{y})}{\sqrt{\left[\left(\Sigma(xi - \bar{x})^2 * \Sigma(y_i - y)^2\right)\right]}}$$

- The coefficient of determination is related to R^2 by

  Coeff. of Determination (of mean of y) = $1 - R^2$
- The Standard Error of Estimation of mean (of y) is: $SSE*\sqrt{(1-R^2)}$

Unit 2: Multiple Regression Equation

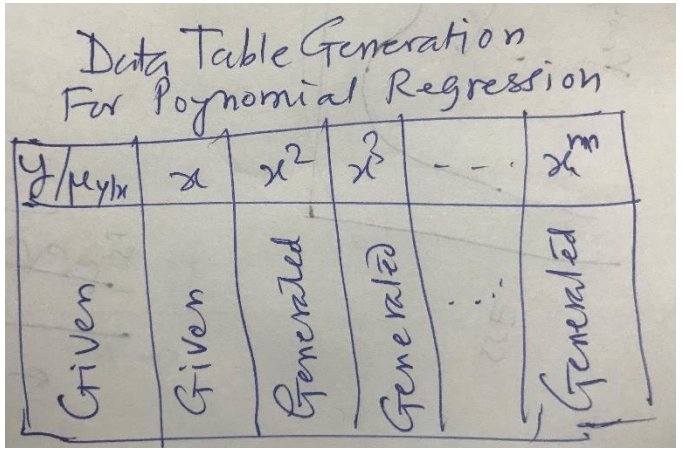
- In Simple Linear Regression, a response variable Y depends on the value assumed by a single predictor variable X.
- This concept is extended to multiple predictor variables X.
- Two basic models of Multiple Linear Regression are:
  - Polynomial Model: A single predictor variable can appear to powers greater than 1.
  - Multilinear Model: in which more than one distinct predictor variables appear.

Unit 2: Polynomial Regression Equation

Polynomial Model of degree p:

- The general model of degree p expresses the mean response Y as a polynomial function of one predictor.

  $\mu_{y|x} = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \ldots\ldots + \beta_p X^p$



- Two special cases are:

  (i) Quadratic Model is given by: $\mu_{y|x} = \beta_0 + \beta_1 X^1 + \beta_2 X^2$, and

  (ii) The Cubic Model is given by: $\mu_{y|x} = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3$

- If we let $X_1 = x$, $X_2 = x^2$, $X_3 = x^3$, ……., $X_p = x^p$, then the model is written in general form as:

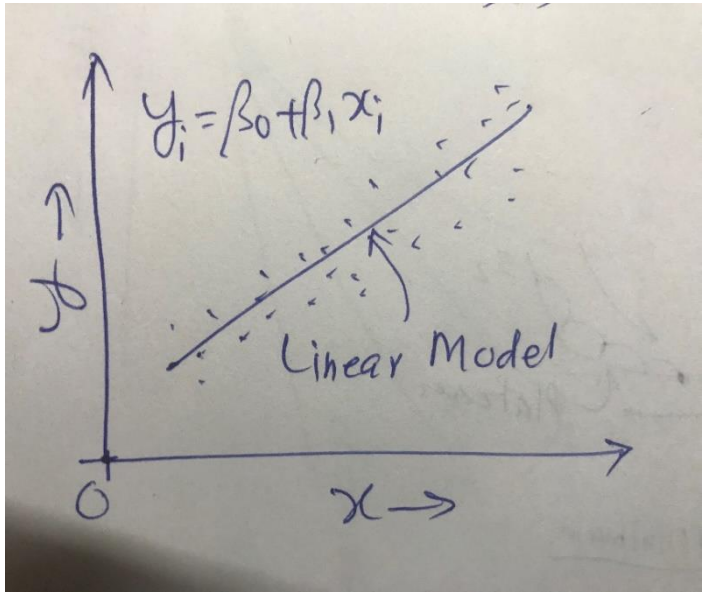
  $\mu_{y|x} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_p X_p$

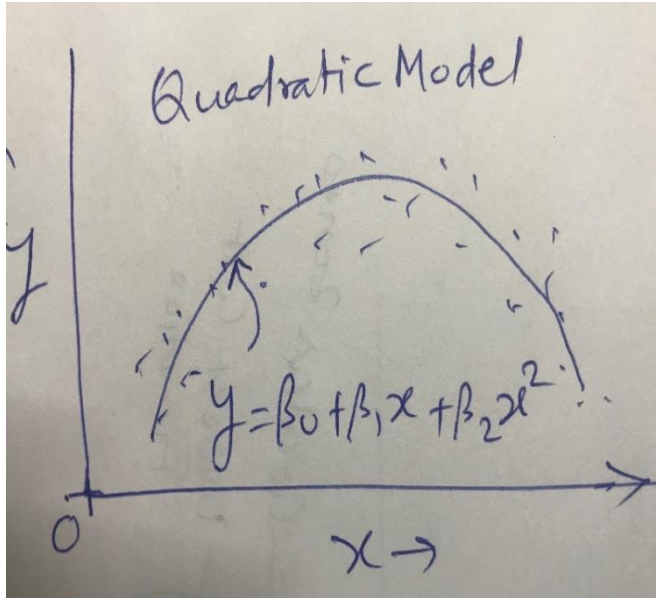
- If we replace $\mu_{y|x}$ by y for given X, then $y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \ldots\ldots + \beta_p X^p + E_i$ is the Polynomial Regression equation model with error term.

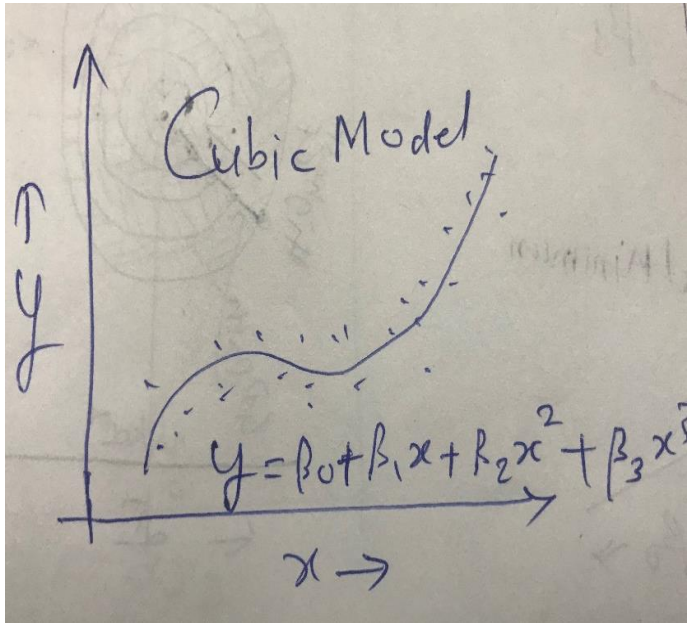
  $E \sim N(0, SEE/(n-2))$

Unit 2: Scatterplots

- The scatterplots of y vs X helps to decide about which polynomial model is appropriate to the given data.







Unit 2: Polynomial Regression Equation

- If y is a function of powers of independent variables X, which stands for x_j, j = 1 to m, then the eqn. is known as Polynomial Linear Regression equation and is given by:

$$y_i = \mu_{y|x} = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \ldots\ldots + \beta_m X^m + E_i$$
$$= \beta_0 + \sum_{j=1}^{m} \beta_j X_i^j + E_i$$
$$= \sum_{j=0}^{m} \beta_j X_i^j + E_i \qquad \text{with } X_0 = 1 \text{ always}$$

- The sum of E_i^2 is minimised for least square fitting:

$$\sum_{j=1}^{m} (E_i^2) = \sum_{j=1}^{m} (yi - \beta_0 - \sum_{j=1}^{m} \beta_j X_i^j)^2$$

- The (m+1) Normal Equations will be:

$$\sum_{i=0}^{n} y_i - n\beta_0 - \sum_{i=0}^{n} (\sum_{j=1}^{m} \beta_j X_i^j) = 0 \qquad \text{for } \beta_0$$
$$\sum_{i=0}^{n} y_i x_i^j - \beta_0 x_i^j - x_i^j \sum_{i=0}^{n} (\sum_{j=1}^{m} \beta_j X_i^j) = 0 \qquad \text{for } \beta_j,\ j = 1 \text{ to } m$$

Unit 2: Multiple Regression Equation

- If y is a function of several independent variables X, which stands for x_j, j = 1 to m, then the eqn. is known as Multiple Linear Regression equation and is given by:

$$y_i = \mu_{y|x} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_m X_m + E_i$$
$$= \beta_0 + \sum_{j=1}^{m} \beta_j X_{ji} + E_i$$
$$= \sum_{j=0}^{m} \beta_j X_{ji} + E_i \qquad \text{with } X_0 = 1 \text{ always.}$$

- The sum of E_i^2 is minimised for least square fitting:

$$\sum_{j=1}^{m} (E_i^2) = \sum_{j=1}^{m} (y_i - \beta_0 - \sum_{j=1}^{m} \beta_j X_{ji})^2$$

Unit 2: Multiple Regression Equation

- The First Normal Equations is given by:

$$\sum_{i=0}^{n} y_i - n\beta_0 - \sum_{i=0}^{n} \left(\sum_{j=1}^{m} \beta_j X_{ji}\right) = 0; \quad \text{for } \beta_0$$

- The (m+1)th Normal Equations will be:

$$\sum_{i=0}^{n} y_i x_{ji} - \beta_0 \sum_{i=0}^{n} x_{ji} - x_{ji} \sum_{i=0}^{n} \left(\sum_{j=1}^{m} \beta_j X_{ji}\right) = 0 ; \text{ for } \beta_j, j = 1 \text{ to } m$$

- These above Normal Equations have to be solved for (m+1) coefficients following either:
  - Successive Elimination method,
    or
  - Matrix inversion method.

Applied Machine Learning

Unit-2: Regression Techniques

4

Unit 2: Multiple Regression Equation (R)

- If y is a function of several independent variables X, which stands for x_j, j = 1 to m, then the eqn. is known as Multiple Linear Regression equation and is given by:

$$y_i = \mu_{y|x} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_m X_m + E_i$$
$$= \beta_0 + \sum_{j=1}^{m} \beta_j X_{ji} + E_i$$
$$= \sum_{j=0}^{m} \beta_j X_{ji} + E_i \qquad \text{with } X_0 = 1 \text{ always.}$$

- The sum of E_i^2 is minimised for least square fitting:

$$\sum_{j=1}^{m} (E_i^2) = \sum_{j=1}^{m} (y_i - \beta_0 - \sum_{j=1}^{m} \beta_j X_{ji})^2$$

Unit 2: Multiple Regression Equation (R)

- The First Normal Equations is given by:

$$\sum_{i=0}^{n} y_i - n\beta_0 - \sum_{i=0}^{n} \left(\sum_{j=1}^{m} \beta_j X_{ji}\right) = 0; \quad \text{for } \beta_0$$

- The (m+1)th Normal Equations will be:

$$\sum_{i=0}^{n} y_i x_{ji} - \beta_0 \sum_{i=0}^{n} x_{ji} - x_{ji} \sum_{i=0}^{n} \left(\sum_{j=1}^{m} \beta_j X_{ji}\right) = 0 ; \text{ for } \beta_j, j = 1 \text{ to } m$$

- These above Normal Equations have to be solved for (m+1) coefficients following either:
  - Successive Elimination method,
    or
  - Matrix inversion method.

Unit 2: Writing Normal Equations

- The Polynomial Regression equations is given by:

$$y_i = \beta_0 + \beta_1 X^1_i + \beta_2 X^2_i + \beta_3 X^3_i + \ldots\ldots + \beta_m X^m_i + E_i$$

- The normal equations are given by:

$$n\beta_0 + \beta_1 \Sigma x_i^1 + \beta_2 \Sigma x_i^2 + \beta_3 \Sigma x_i^3 + \ldots\ldots + \beta_m \Sigma x_i^m = \Sigma y_i$$

$$\beta_0 \Sigma x_i^1 + \beta_1 \Sigma x_i^2 + \beta_2 \Sigma x_i^3 + \beta_3 \Sigma x_i^4 + \ldots\ldots + \beta_m \Sigma x_i^{m+1} = \Sigma x_i^1 y_i$$

$$\beta_0 \Sigma Xi^2 + \beta_1 \Sigma Xi^3 + \beta_2 \Sigma xi^4_i + \beta_3 \Sigma x_i^5 + \ldots\ldots + \beta_m \Sigma x_i^{m+2} = \Sigma x_i^2 y_i$$

.....

.....

$$\beta_0 \Sigma X_i^m + \beta_1 \Sigma X_i^{m+1} + \beta_2 \Sigma x_i^{m+2} + \beta_3 \Sigma x_i^{m+3} + \ldots\ldots + \beta_m \Sigma x_i^{m+m} = \Sigma x_i^m y_i$$

- First eqn. is obtained by multiplying by coeff. of β_0 , i.e.1.
- Next eqns. are obtained by sequentially multiplying by coefficients of β_j
- This above statement is known as the rule for writing normal equations.

Unit 2: Solving Normal Equations

- **The Multilinear Regression equations is given by:**

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots + \beta_m X_{mi} + E_i$$

- **The normal equations are given by:**

$$n\beta_0 + \beta_1 \Sigma X_{1i} + \beta_2 \Sigma X_{2i} + \beta_3 \Sigma X_{3i} + \ldots\ldots + \beta_m \Sigma X_{mi} = \Sigma y_i$$

$$\beta_0 \Sigma X_{1i} + \beta_1 \Sigma X_{1i} X_{1i} + \beta_2 \Sigma X_{1i} X_{2i} + \beta_3 \Sigma X_{1i} X_{3i} + \ldots\ldots + \beta_m \Sigma X_{1i} X_{mi} = \Sigma X_{1i} y_i$$

$$\beta_0 \Sigma X_{2i} + \beta_1 \Sigma X_{2i} X_{1i} + \beta_2 \Sigma X_{2i} X_{2i} + \beta_3 \Sigma X_{2i} X_{3i} + \ldots\ldots + \beta_m \Sigma X_{2i} X_{mi} = \Sigma X_{2i} y_i$$

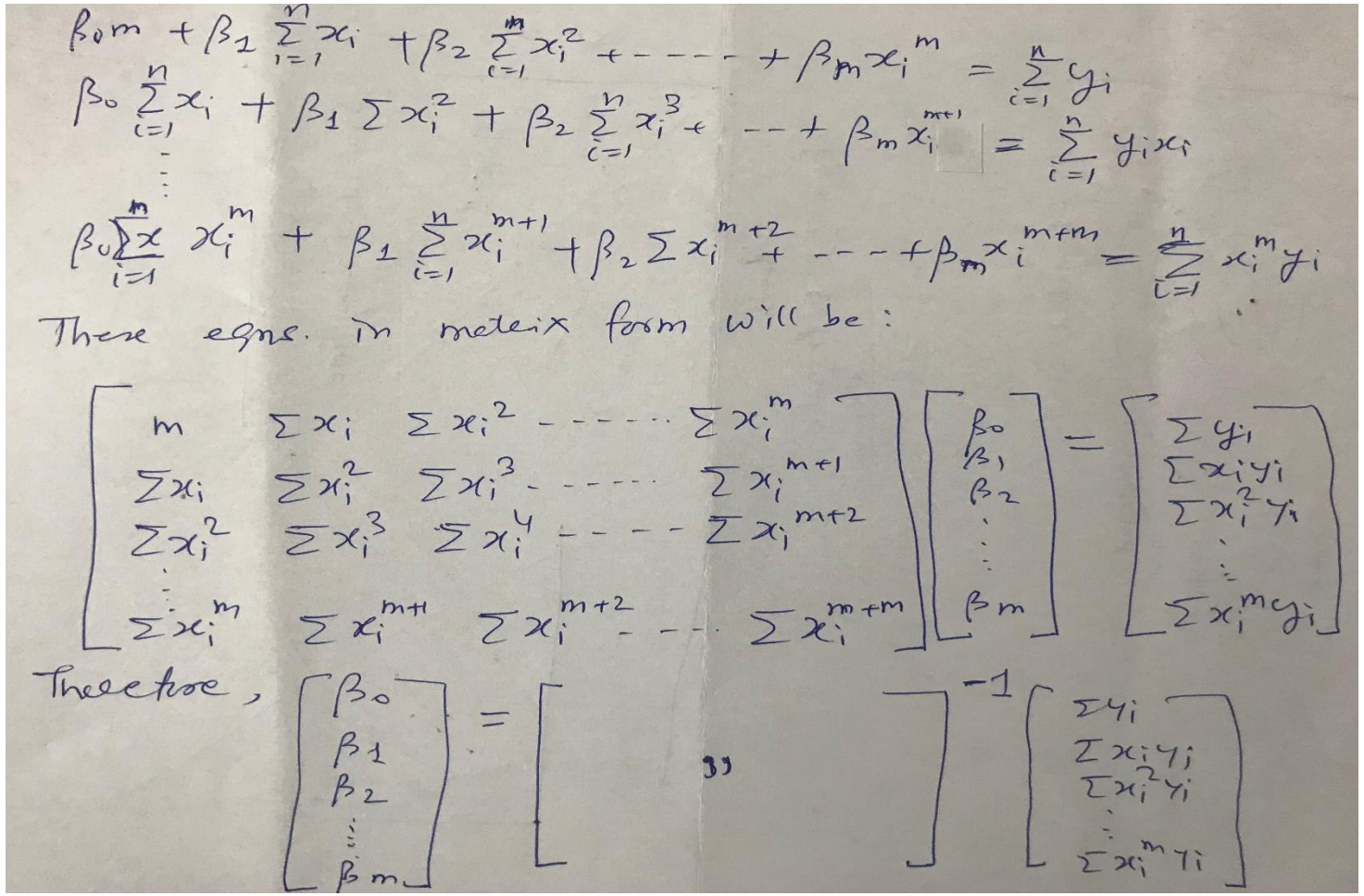
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\beta_0 \Sigma X_{mi} + \beta_1 \Sigma X_{mi} X_{1i} + \beta_2 \Sigma X_{mi} X_{2i} + \beta_3 \Sigma X_{mi} X_{3i} + \ldots\ldots + \beta_m \Sigma X_{mi} X_{mi} = \Sigma X_{mi} y_i$$

- **These Normal Equations can be written in Matrix notation: $VxB = R$**
  - V is a matrix of m rows and m column formed by n size matrix of functions of independent variables
  - B is matrix of m rows and 1 column of regression coefficients and R is also matrix of functions of responses

Unit 2: Polynomial Regression Equation

In matrix form: The (m+1) Normal Equations will be:



Unit 2: Solving Normal Equations

- The vector of coefficients is obtained by

  $B = V^{-1}xR$, where V^{-1} is inverse of matrix V and R is a response vector.
- As V is square matrix, its inverse does exit. **Note: Non Square matrix can not be inverted.**
- However, to get unique solution, the determinant of V, i.e. det(V) should be significantly different from zero.
- Small value or near zero value of Determinant causes instability in inverting the matrix, as it appears in denominator in expression for inverse matrix.
- The instability due to small value determinant can cause problem in getting unique solution. **The residual digits in float numbers are not significant.**
- The zero or near zero value of a determinant can arise due to strongly correlated attributes of some samples/examples.
- Such strongly correlated or repeated with nearly same attribute values for samples are difficult to locate if number of samples/examples is very large.

Unit 2: More about Multilinear Regression

Multiple/Polynomial Regression Techniques:

- The explanatory variables are expected to be independent among themselves.
- They represent different causes for variability in the dependent/response variable.
- In case of any two or more variables are linearly correlated among themselves or multiple of one another, then such variables do not add any information.
- Thus one of them is redundant.

Unit 2: Computational Complexity

Good Aspects of Normal Equation:

- The Normal Equations are linear with respect to m, complexity is O(m).
- So it handles large data sets efficiently, provided they fit in memory.
- Once fit with Normal Equations, then the prediction is fast, the computational complexity is O(m).
- Valid in case solving them by successive elimination method.

Computational Complexity in Case of Matrix Inversion Method:

- The normal equations compute $X^T.X$, which is a m by m matrix. M is number of explanatory variables or features.
- This value can go in to 100s to 1000s or more in some problems.
- The computational complexity of inverting a matrix X is $O(m^{2.4})$ to $O(m^3)$ depending on implementation method.
- This means for every doubling of m, the complexity time increases by $2^{2.4}$ (= 5.3) or 2^3 (= 8) factor.

Unit 2: Computational Complexity

Computational Complexity:

- Consequently, the Normal Equations solution gets very slow when the no. of features grow large (~ 10^6).
- In this situation, we need a different way to Least Square Fitting or Train to Linear regression model, when no. of features is very large, too large to fit in computer memory.
- One generation of such models is known as **Gradient Descent method**.
- **Gradient Descent method** is a generic optimization algorithm/optimization of mean minimizing error function/ or a cost function (SSE, MSE, RMSE).

Unit 2: More about Multilinear Regression

Multiple/Polynomial Regression Techniques:

- In case of redundancy, this is known as the variables are collinear or there exists multi-collinearity.
- The determinant of functions of X variable is near zero, if not exactly zero.
- If so, then there does not exist a single stable solution or there exist multiple solutions.
- In this situation, a procedure called Gradient Descent Method is used.
- This method is not based on solving Normal Equations either by Successive elimination or matrix inversion.
- The Gradient Descent method yields biased estimates with low variance so that the SSE is usually small.
- Most accurate solution comes from Normal Equations method.

Applied Machine Learning

Unit-2: Regression Techniques

5

Unit 2: Gradient Descent Method of Optimization

- The general Multiple Linear Regression model is written as:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots + \beta_m X_{mi} + e_i$$

$$= \beta_0 + \sum_{j=1}^{m} \beta_j X_{ji} + e_i = \sum_{j=0}^{m} (\beta_j X_{ji}) + e_i \text{ with } X_0 = 1 \text{ always.}$$

- In the Least Square Fitting, the equations are solved by Normal Equations method to determine the β_j coefficients.
- The normal equation method is such that the cost function/SSE is minimised or optimised.
- This is known as Optimisation Technique.

Unit 2: Gradient Descent Method of Optimization

- The **Gradient Descent Method** (known as GD method) is one such method of optimisation used when:
- Number of features is very large, too large to fit in computer memory, and also too large to use matrix inversion method of solving Normal Equations.
- **General Outline of the Gradient Descent Method :**
  - In this method, first start with assigning **some random values** to β_j and evaluate SSE=$\sum_{i=0}^{n} e_i^2$ after computing ei for each sample.
  - Use this value of cost function to sequentially change the β_j values, so that cost function reduces sequentially by evaluating slope of SSE with respect to β_j, in the direction of fastest descent.
  - Repeat the above steps until the cost function reaches the minimum value and remains so thereafter. This is why this method is known as “Gradient Descent Method”.
  - Then the final values of β_j are the solution to the regression model equation.

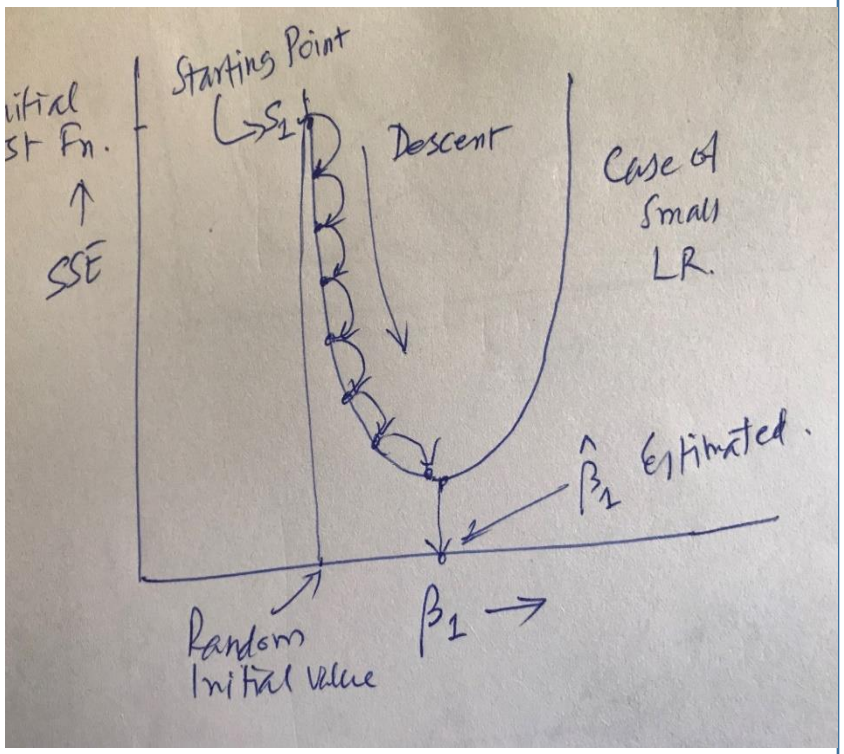
Unit 2: Gradient Descent Method of Optimization

- But in actual practice, the Gradient Descent method implementation is not easy or simple.
- The time taken depends highly on the initial values chosen and method of introducing changes in the coefficients in subsequent steps.
- Let us consider a case of linear regression with a single independent variable.

  $y_i = \beta_0 + \beta_1 X_i + e_i$
- Then, SSE = $\sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (\beta_0 + \beta_1 X_i - yi)^2$
- Here β_0 is bias (also known as intercept) and β_1 is slope of linear regression model equation and line.

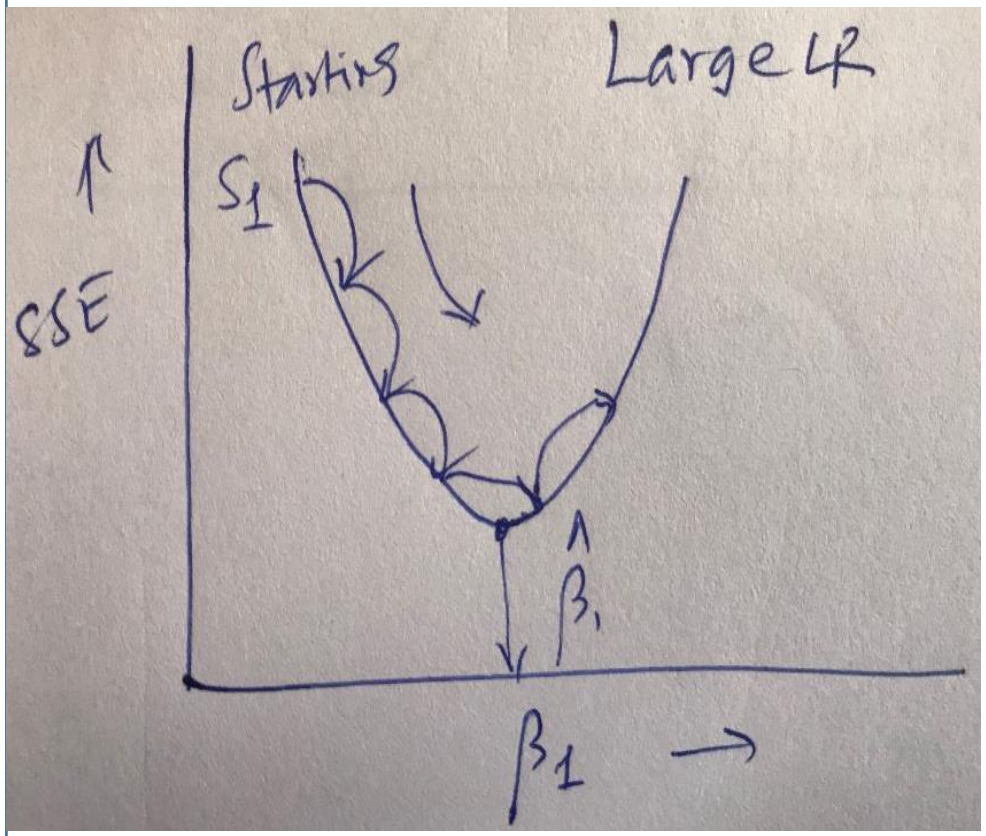
Unit 2: Gradient Descent Method of Optimization

- First let us consider the theoretical SSE dependence on β_1. Obviously, there is a certain value of β_1 when SSE results in a minimum value.
- Note that SSE always positive. Therefore, the curve is nearly U shaped.
- If we start with some random initial value of β_1, then the next move is to change β_1 in steps in the direction of decreasing SSE (or Descending SSE).
- This is obtained by finding slope at that point of SSE.
- Change the β_1 in the direction of decreasing SSE in steps until it reaches a point of minimum SSE, i.e. the point of $\widehat{\beta_1}$.
- The step size chosen is called Learning rate (LR).



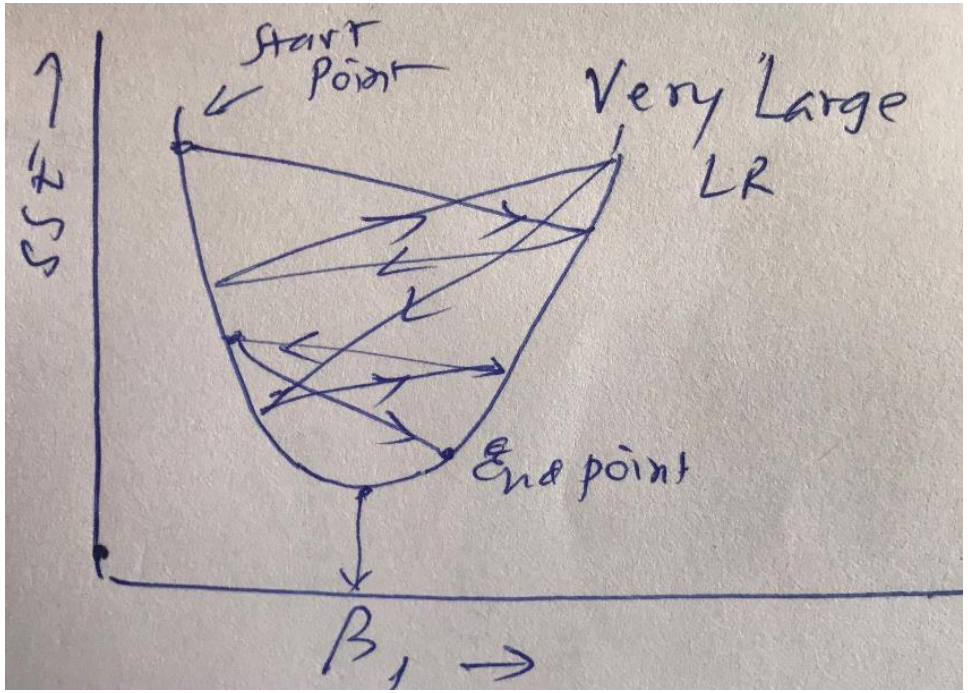
Unit 2: Gradient Descent Method of Optimization

- After the first step, the next step will result in decrease in value of SSE in steps of size decided by the set value of LR.
- If the Learning Rate (LR) is small, then the process has to go through large number of steps and it becomes highly time consuming.
- But, this can result in reaching to a point very close to the real optimum value of SSE and $\hat{\beta}_1$.



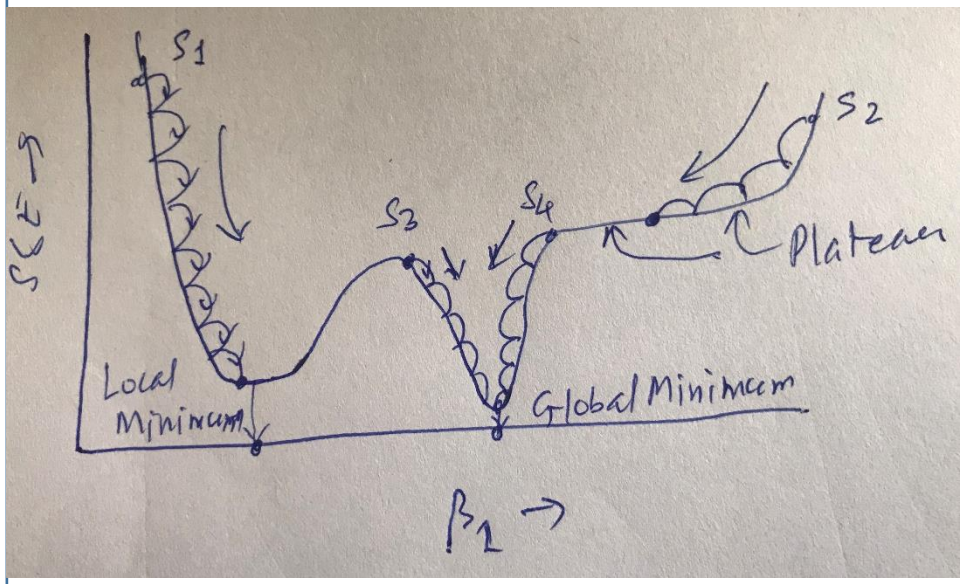
Unit 2: Gradient Descent Method of Optimization

- On the contrary, if the step size (LR) is too large, then a convergence point close minimum may be reached in a fewer steps.
- But the error in estimated $\widehat{\beta}_1$ may be quite large.
- Also, a very large LR may result into the point jumping across the valley and end up on the other side, possibly even higher up in subsequent steps.
- This might make the algorithm to diverge with larger and larger values and failing to reach a good solution.



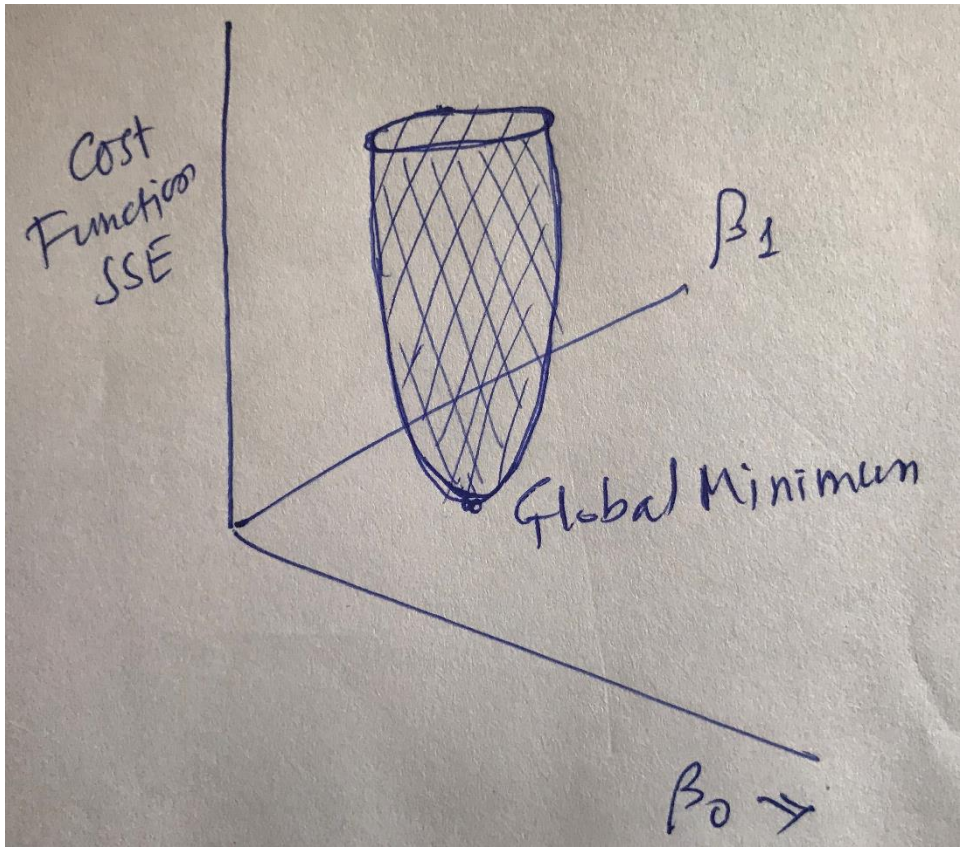
Unit 2: Gradient Descent Method of Optimization

- The cost function shape may not always of regular bowl or U shape.
- They may be like irregular terrains *(shown in figure)* making the convergence more difficult.
- In such cases, there are risks of settling GD in a false local minimum, depending on the starting point of GD process.
- When started from left (e.g. S_1), the GD ends up in a local minimum L_1.
- And, if started from right (S_2 point), it ends up on a plateau (L_2 Point).
- The global minimum is not reached in either case, i.e. in both the cases.
- Possible to reach global minimum if started from S3 or S4 points.



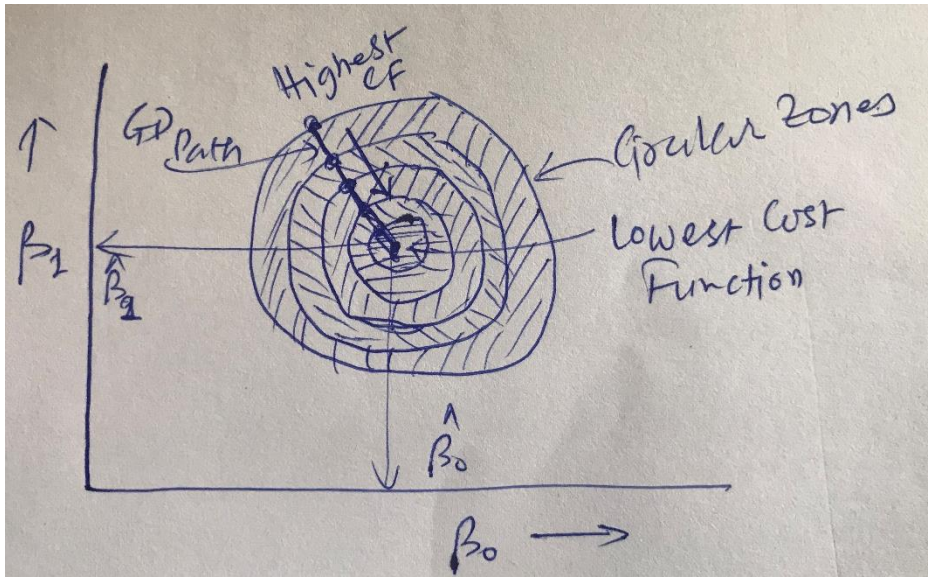
Unit 2: Gradient Descent Method of Optimization

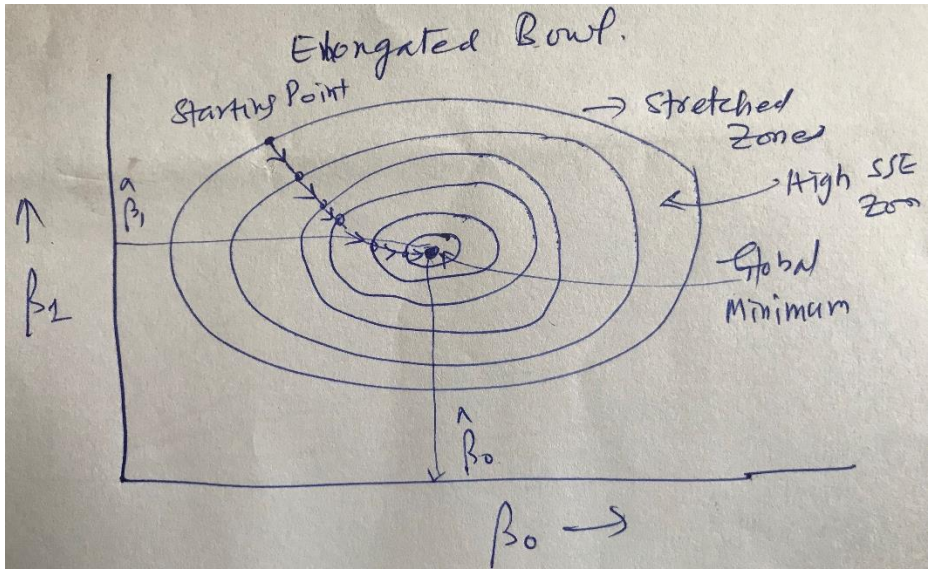
- Fortunately, the SSE cost function for **linear regression model** happens to be Convex Bowl Shaped function.
- If one picks two points, the lone line segment joining them never crosses the curve.
- Therefore, there are no local minima, but just one Global minimum exists. Also, the slope is a continuous function.
- The GD method promises to approach Global minimum, if waited for long enough time and LR is not too high.
- The cost function in two variables (2-D) looks like a basket or bowl as shown in the 3-D figure.
- The next figure also shows the view from top projected on the β-planes.



Unit 2: Gradient Descent Method of Optimization

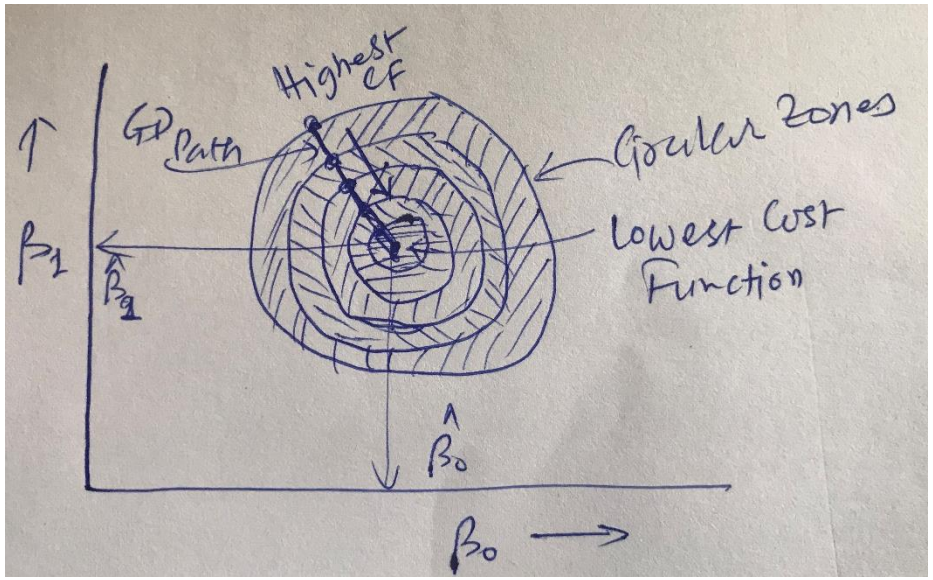
- The zones of the cost function are circular in shape if the scales of are β_0 and β_1 are the same (top fig).
- But, they are not on the same scale, the bowl gets elongated in shape (as shown in lower figure).
- The SSE components will be dominated by those variables who have large values as well as large range.
- Hence, the SSE will be determined by the variable with large ranges.
- Fortunately, the cost function is convex in the cases of linear regression.
- Hence the global point (search point) lies at the bottom of the bowl.

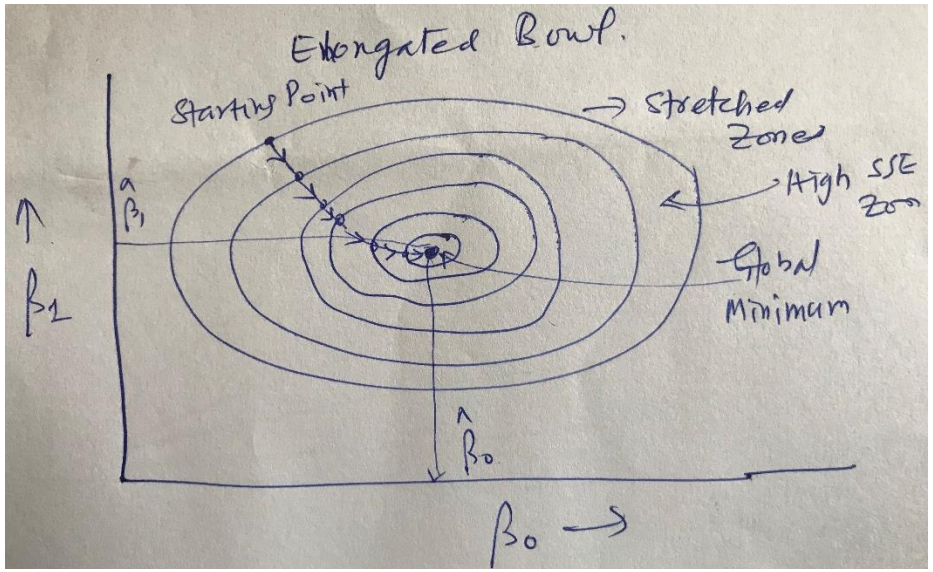




Unit 2: Gradient Descent Method of Optimization

- **The consequence of the elongated are:**
  - In both cases, whether circular or elongated bowl shapes, the global minimum reached is the same.
  - The GD curve follows a curved path and take longer time to converge to reach a point near to the global minimum
  - Therefore, it is necessary to have all x_j the same scale.
  - Usually, all variables are scaled to 0 to 1 scale.
  - After the regression performed, they are scaled back, if required, by dividing by the scaling factor.





Applied Machine Learning

Unit-2: Regression Techniques

6

Unit 2: Gradient Descent Method of Optimization

- In Gradient Descent implementation, the gradient of cost function has to be computed as a function of each parameter β_j.
- This is called Partial Derivative method.
- The partial derivative is given by:

$$\frac{\partial}{\partial \beta}[SSE(\beta)] = (2/n)\sum(\beta^T \cdot X - Y)xj$$

Here β is a vector of $[\beta_0, \beta_1, \beta_2, \ldots, \beta_m]$

- The total derivative will be:

$$\nabla_\beta [SSE(\beta)] = \left[\frac{\partial}{\partial \beta}[SSE(\beta_0)], \frac{\partial}{\partial \beta}[SSE(\beta 1)], \ldots\ldots, \frac{\partial}{\partial \beta}[SSE(\beta m)]\right]^T$$

$$= (2/n)\sum(\beta^T \cdot X - Y)xj$$

Unit 2: Gradient Descent Method of Optimization

Batch Gradient Descent Method:

- This method of Gradient Descent uses the full data set at each gradient descent step.
- Therefore, this is known as Batch-Gradient Descent.
- Due to this, this is a very very slow and time consuming method.
- But, this method is found to be faster than the Normal Equations Solving method
- This is especially so when there are large number of features and also examples.

Unit 2: Batch Gradient Descent Method

The Learning Rate (LR):

- It is used to decide the next β value depending on GD value of the SSE:

  $\beta(\text{next step}) = \beta(\text{present step}) - \eta * \nabla_{\beta}(SSE(\beta))$

- Here the η is known as the Learning Rate (LR).
- The value of the Gradient Descent is to determine the β value for the next step.
- The next β value is in the direction of decreasing SSE.
- Note: The Gradient is always referred to direction of decreasing direction.

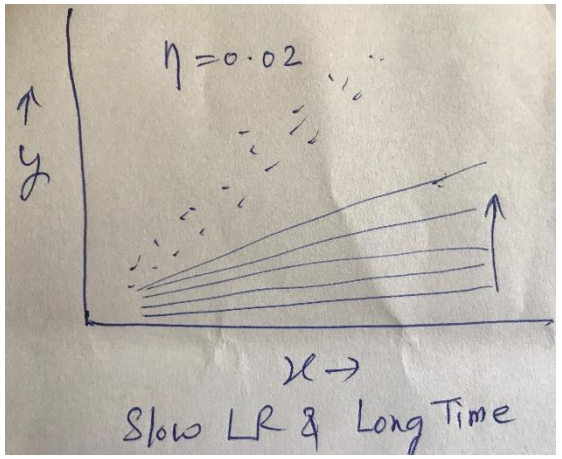
Unit 2: Batch Gradient Descent Method

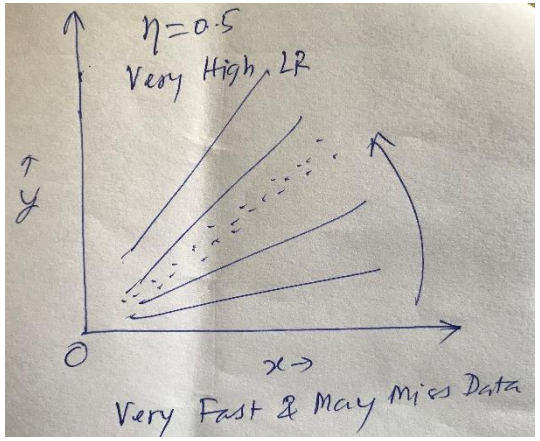
The Learning Rate (LR):

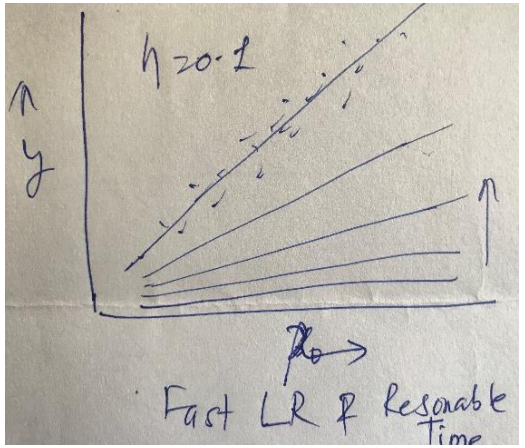
- The ultimate optimal point is reached and the SSE value is same/nearly same to that which could be obtained by Solving Normal Equations, but in a faster way.
- But, the convergence and final value of regression eqn. depends very much on vale of the Learning Rate η set.
- Large learning rate converges faster, but there can be large difference in position of final optimum point reached and the actual optimum point.
- Small learning rate converges slower, but it will end up very close to the final optimum point.
- In case of Batch Gradient Descent method, a fixed η has convergence rate of O(1/no. of iterations).

Unit 2: Gradient Descent Method of Optimization

- But, the convergence and final value of regression eqn. depends very much on vale of the Learning Rate η set.
- Large learning rate converges faster, but there can large difference in position of final optimum point reached and the actual optimum point.
- Small learning rate converges slower, but it will end up very close to the final optimum point.
- In case of Batch Gradient Descent method, a fixed η has convergence rate of O(1/no. of iterations).





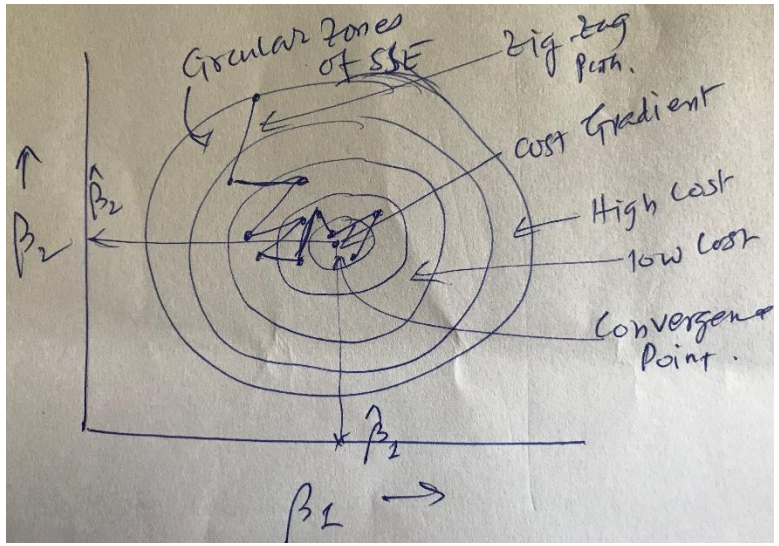


Unit 2: Batch Gradient Descent Method

- Important Limitation of this Gradient Descent Method:
  - This method uses full data set at each gradient descent step.
  - Therefore, it is known as **Batch Gradient Descent** algorithm.
  - Since it uses whole data set at every epoch, it is much slower and time consuming.
  - Even if it is slow, it is found that this Batch GD is still faster than Solving Normal Equations method.
  - It is so especially when there are large number of features and examples/samples both.
  - To find a good Learning Rate, the grid search method is used.
- What should be best strategy:
- To achieve faster convergence and good accuracy:
  - Set a large value for number of iterations (~ 100 or greater),
  - Start with large LR and slowly keep decreasing it, and
  - Finally, exit the iterations when the change in SSE from previous step to present step is very small and less than a pre-set value, called Tolerance value.

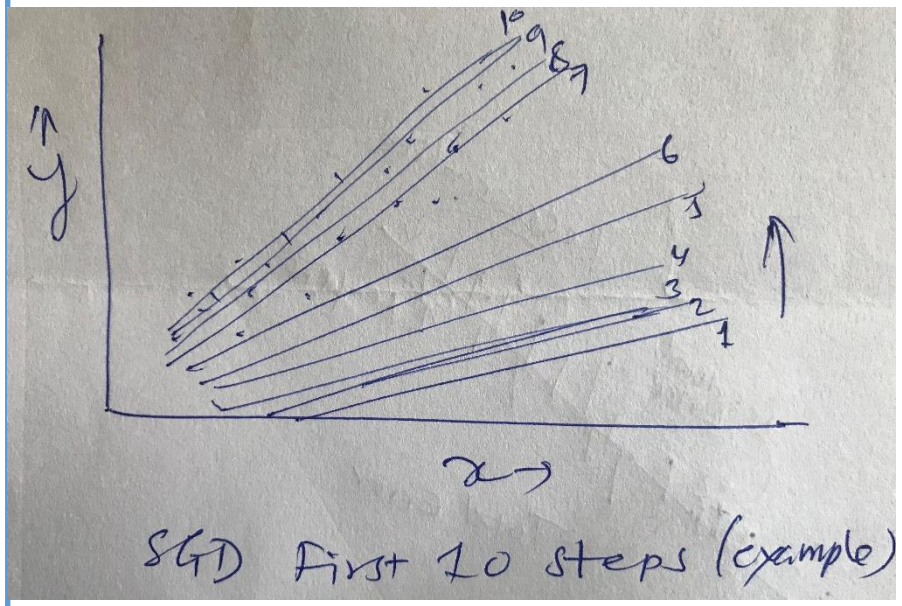
Unit 2: Stochastic Gradient Descent Method

- The Stochastic Gradient Descent (SGD) is a variant of the Gradient Descent algorithm overcomes the problem of long time requires in case of the Batch Gradient Descent method.
- In batch gradient descent method the whole data set is used in every step (every epoch).
- Instead of using total data set, the Stochastic Gradient Descent method picks up randomly one sample or example at every step for computation of gradient.
- Thus the execution time speeds up by large factor, as only one sample data is in memory at each iteration or epoch.
- But the execution speed up has a drawback.
- The cost function bounces up and down and comes at the cost of decreasing only on average.
- Although over time, this will end up very close to the global minimum, but once it gets there it continues to bounce around, but never settles down.
- Thereby, once the algorithm stops, the final parameter values are Good, but not Optimal.



Unit 2: Stochastic Gradient Descent Method

- When the cost function is highly irregular (as shown in irregular shape of cost function), the SGD method can help the algorithm to jump out of the local minima.
- That is, the SGD method has better chance of finding the global minimum point than that of Batch Gradient Descent method does.
- One solution to make SGD to reach to global minimum point is to reduce the Learning rate gradually in the process.
- Start with large learning rate η and decrease gradually allowing the algorithm to settle at the global minimum point.
- This method of slowing changing η is known as “Simulated Annealing”.



Unit 2: Stochastic Gradient Descent Method

- If the LR is reduced too quickly, the algorithm may get stuck in a local minimum.
- If the LR is reduced too slowly, it may jump around the global minimum for long time and end up with sub-optimal solution, it the algorithm reached end of number of iterations set.
- Note each round of iteration is called an EPOCH.
- Since some examples are picked randomly, there are chances of those samples picked up several times.
- To avoid this repeated picking up some samples, the data set is shuffled at the beginning of each epoch.
- This shuffling method further slows down the convergence process.

Applied Machine Learning

Unit-2: Regression Techniques

7

Unit 2: Stochastic Gradient Descent Method (R)

- If the LR is reduced too quickly, the algorithm may get stuck in a local minimum.
- If the LR is reduced too slowly, it may jump around the global minimum for long time and end up with sub-optimal solution, it the algorithm reached end of number of iterations set.
- Note each round of iteration is called an EPOCH.
- Since some examples are picked randomly, there are chances of those samples picked up several times.
- To avoid this repeated picking up some samples, the data set is shuffled at the beginning of each epoch.
- This shuffling method further slows down the convergence process.

Unit 2: Mini-Batch Gradient Descent Method

Mini-Batch Gradient Descent:

- In the Stochastic Gradient Descent, in every epoch, one random example is chosen for computing SSE, its descent gradient and to determine learning rate LR.
- The MINI-BATCH Gradient Descent algorithm uses a small data set of examples to compute the GD and Learning rate. Therefore, known as Mini-Batch GD algorithm.
- This method gives a performance boost in optimisation of matrix operations.
- The mini-batch GD algorithm progresses in parameter space in less erratic compared to SGD.
- Specially so, when batch size is fairly large.
- But it may be harder to escape from local minima.

Unit 2: Mini-Batch Gradient Descent Method

The comparison of paths taken by the three GD methods in parameter space show that:

- The Batch-GD's path actually stops very very close to the minimum.
- The paths of GD and SGD methods continue to walk around.
- The disadvantage is that the Batch GD takes more time than SGD.
- It is found that the SGD and Mini-Batch GD can also perform equally well, if good learning rates are used.

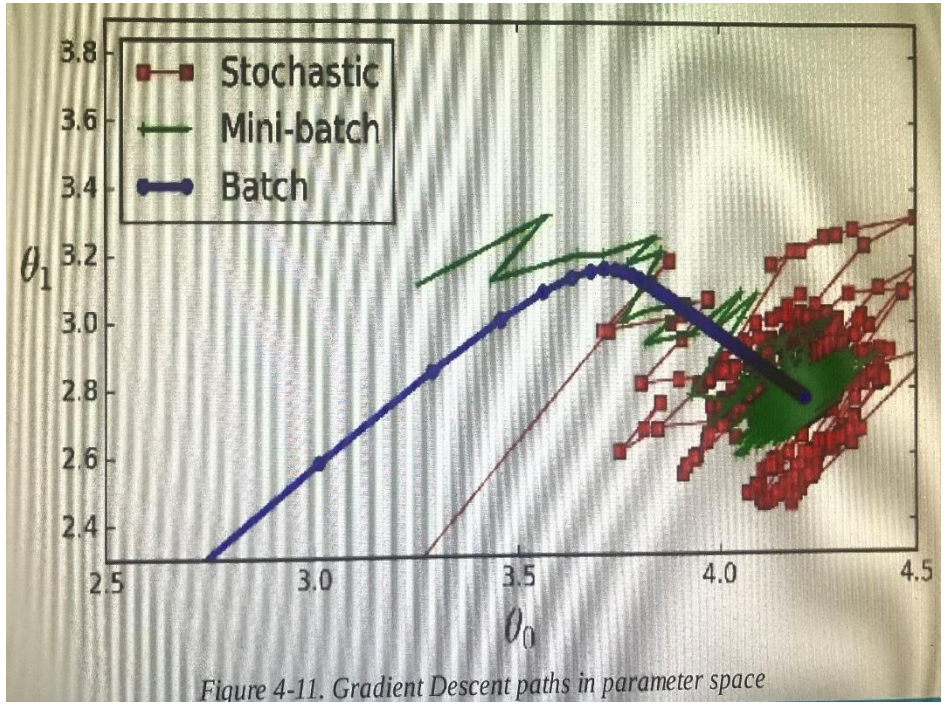
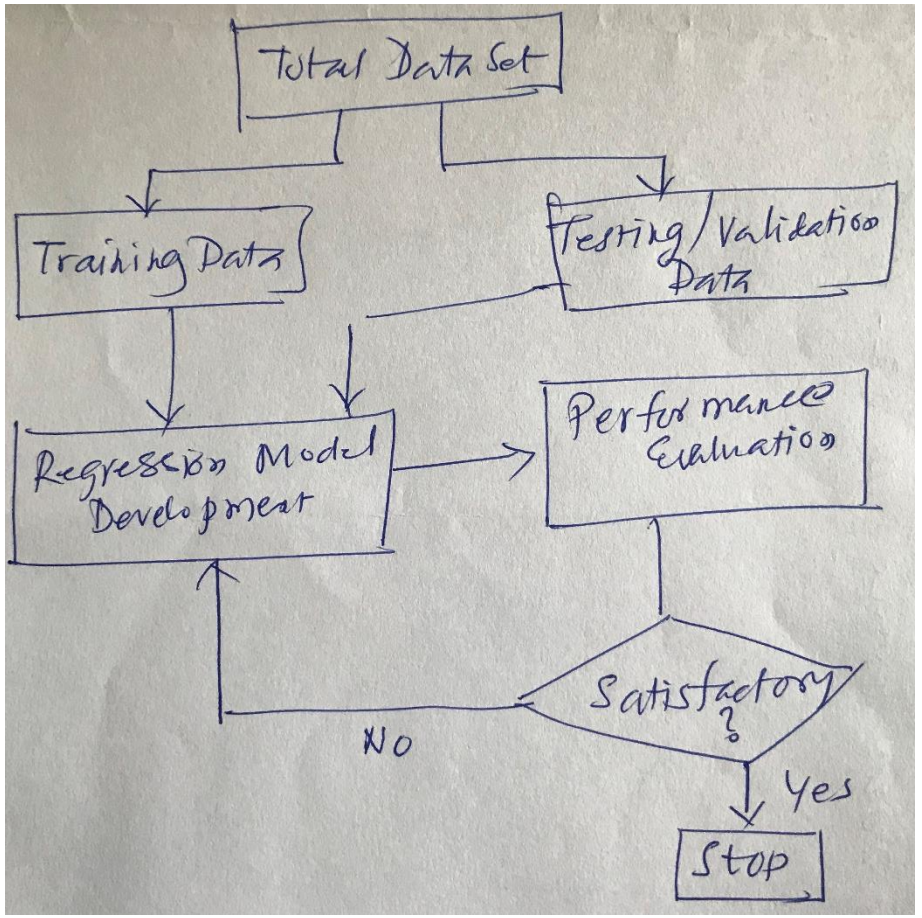


Figure 4-11. Gradient Descent paths in parameter space

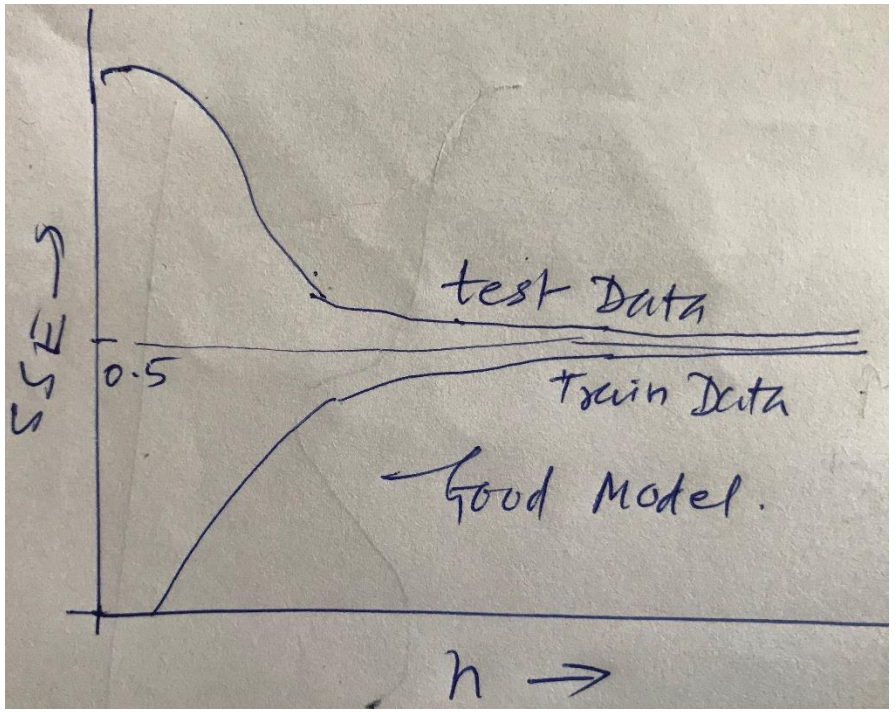
Unit 2: Splitting Data into Training and Testing Sets

- The major goal of regression modelling is to use it for model prediction for unknown independent variables/samples/features.
- The fitting of LSF regression model developed is measured by Least/Optimum value of SSE and Correlation Coefficient (R) , R^2 or coefficient of determination etc.
- To test how good is the model, the sample data is divided in two parts, viz. training data set and testing data set.
- Training data set is used to develop model and the testing data set is used for validation/prediction ability.
- Some % of data are selected randomly from total data set as training data set.
- The model developed using the training data set is used to predict dependent variable and these are then compared with the actual values by computing the squared errors (SE).



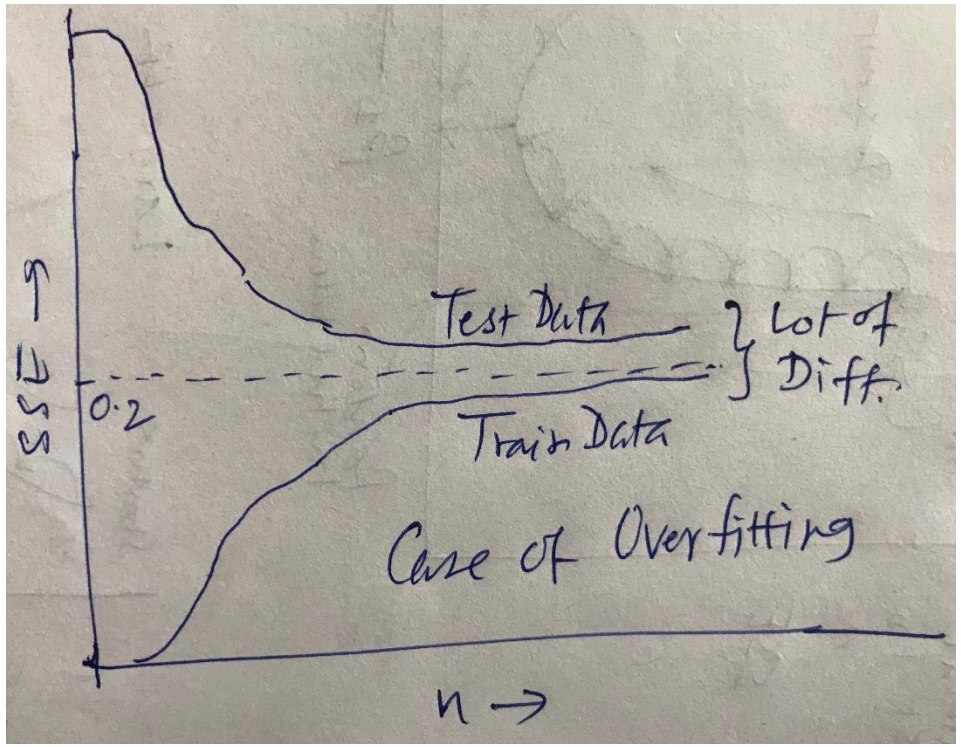
Unit 2: Under fitting and Overfitting

- If the regression model under fits the training data, it is advised to increase the data size.
- The figure on right shows the effect of increasing training data on the SSE of both the training and testing data sets.
- As the training data size increases from 1 to 80 (in the example shown), the SSE increases rapidly from zero at the data size of 1 (single independent variable) or 2 (two variables) to a certain value and then steadily increases to attain a constant value.
- But, the SSE of testing data decreases and slowly attains a value close to SSE of training data.
- This is the case of good model wherein the LR model generalises very well.



Unit 2: Underfitting and Overfitting

- In the second case (fig on the right), the training model does not generalize very well even with n = 80. Needs to increase n further.
- This is a case of overfitting the training data.
- As n is increased, the model may behave better. But there may not exist sufficiently large no. of examples.
- But in case of this second situation, the overfitting situation, even with the available data size, the model needs to be Regularized.
- Regularization means to reduce the overfitting or moderation of overfitting and improve generalization (i.e. prediction on the test data).
- **Ridge Regression method** is one such method meant for Regularization.



Unit 2: Ridge Regression

- As stated above the Ridge Regression moderates the fitting.
- Moderation decreases correlation coefficient or increases the SSE of fitting of regression model.
- This process of reducing the correlation coefficient is known as “Annealing process” of the model.
- A good way to reduce overfitting is to regularise the models (i.e. to constrain model equation).
- The fewer the DoF a model has, the more difficult it is to overfit to the data.

Unit 2: Ridge Regression

- One simple way to regularise a polynomial model is to reduce the number of polynomial degree.
- In case of linear regression model, regularization is achieved by constraining the weights or coefficients of the model.
- There are three ways to constrain LR model: (i) Ridge Regression, (ii) Lasso Regression, and (iii) Elastic Net model.
- The Ridge Regression and LASSO introduces an additive term to the SSE.

Unit 2: Ridge Regression

- The Ridge Regression is a regularized version of the Linear Regression.
- The Ridge Regression introduces an additive term to the SSE.
- A regularization term equal to $\alpha \sum_{j=1}^{m} \beta_j^2$ is added to the cost function, i.e.

  SSE → SSE + $\alpha \sum_{j=1}^{m} \beta_j^2$
- Here α is known as "Hyperparameter" of the regularization of the Ridge Regression.
- The hyperparameter always takes a +ve value, because regularization needs to increase optimal SSE so as to reduce overfitting or reduce Correlation Coefficient

Unit 2: Ridge Regression

Important Points to Note about Ridge Regression:

(i) The regularization term is added to the cost function ONLY during the training process, i.e. the model development process ONLY.

(ii) The regularization term is NOT added during testing or validation phases using the testing or validation data.

(iii)The performance of the trained Ridge Regression model is evaluated against the performance of that of un-regularized model.

(iv) The extent of regularization is determined by value of the hyperparameter α.

(v) Larger the value of α, faster the regularization. Usually smaller values are preferred.

(vi) Larger values of α may overshoot the optimal value of the regularized SSE.

Applied Machine Learning

Unit-2: Regression Techniques

8

Unit 2: Ridge Regression

- The hyperparameter α controls how much the model is regularized.
- When α = 0, no regularization is done.
- It is found that as α value is increased, then all coefficients β_j end up close to zero.
- In this situation, this results in the LR to be flat line, no relationship with features.
- The Ridge Regression cost function is:

  $J(\beta) = SSE + \alpha \sum_{j=1}^{m} \beta_j^2$

- Note that the β_0 term is not regularized, as j starts from 1, not from 0.

Unit 2: Ridge Regression

- The Ridge Regression is a regularized version of the Linear Regression.
- The Ridge Regression introduces an additive term to the SSE.
- A regularization term equal to $\alpha \sum_{j=1}^{m} \beta_j^2$ is added to the cost function, i.e.

  SSE → SSE + $\alpha \sum_{j=1}^{m} \beta_j^2$
- Here α is known as "Hyperparameter" of the regularization of the Ridge Regression.
- The hyperparameter always takes a +ve value, because regularization needs to increase optimal SSE so as to reduce overfitting or reduce Correlation Coefficient

Unit 2: Ridge Regression

- The hyperparameter α controls how much the model is regularized.
- When α = 0, no regularization is done.
- It is found that as α value is increased, then all coefficients β_j end up close to zero.
- In this situation, this results in the LR to be flat line, no relationship with features.
- The Ridge Regression cost function is:

  $J(\beta) = SSE + \alpha \sum_{j=1}^{m} \beta_j^2$

- Note that the β_0 term is not regularized, as j starts from 1, not from 0.

Unit 2: Hyper-parameter Regularization

- The Ridge Regression model would be (with single independent variable X):

  $y_i = \beta_0 + \sum_{j=1}^{m} \beta_j^2 + e_i$

  $J(\beta) = SSE + \alpha \sum_{j=1}^{m} \beta_j^2 = \sum_{i=1}^{n} e_i^2 + \alpha \sum_{j=1}^{m} \beta_j^2$

  $= \sum_{i=1}^{n} (\beta_0 + \sum_{j=1}^{m} \beta_j x_{ji} - y_i)^2 + \alpha \sum_{j=1}^{m} \beta_j^2$

  Or in simple form:

  $J(\beta) = \Sigma_i(\beta_0 + \Sigma_j \beta_j x_{ji} - y_i)^2 + \alpha \Sigma_j \beta_j^2$

- The j Normal Equations are given by: $\frac{\partial J(\beta)}{\partial \beta_j} = \frac{\partial (SSE)}{\partial \beta_j} = 0$,

  $2xji(\beta_0 + \Sigma_j \beta_j x_{ji} - y_i) + 2\alpha \Sigma_j \beta_j = 0$

  which includes the second term as the Regularisation term.

Unit 2: Ridge Regression

Ridge Regression:

- If $\sum_{j=1}^{m} \beta_j^2$ is represented as $(||w||)^2$, the $(||\cdot||)^2$ represents the l_2-norm (Euclidean Norm or based on Euclidean distance).
- In Euclidean norm, Euclidean distance is used in finding error in the fitted value from the measured value.
- The Ridge Regression works better when least square estimates have high variance. Overfitting is due to high variance of the model.
- The Ridge Regression includes all attributes in the final model. It performs better when the response is a function of all attributes with almost equal coefficients.

Unit 2: Ridge Regression

LASSO Regression:

- Contrary to Ridge Regression, the Lasso regression follows the l_1-norm (based on the definition of the Manhattan distance).
- LASSO: Least Absolute Shrinkage and Selector Operator
- In LASSO regression, the regularization term is $\sum_{j=1}^{m} |\beta_j|$
- This term is known as Manhattan Distance.
- The LASSO regression forces some of the coefficients to zero value. Thus it yields sparse model, which are simpler and more interpretable.
- This performs better if a smaller number of attributes have substantial influence than others.

Unit 2: Hyper-parameter Regularization

- The regularisation process involves solving the normal equations with increasing value of the hyper-parameter α.
- Halt the search of the parameter α, when the training models generalizes the validation model very well.
- Note: It is very important to scale the data linearly (e.g. 0 to 1 range) before performing Ridge Regression, as it is sensitive to the scale of the input data.
- Hyper-parameter Tuning: There are two methods of finding the best value of the hyper-parameter α:

  (i) Grid Search Method

  (ii) Random Search Method.

Unit 2: Hyper-parameter Tuning

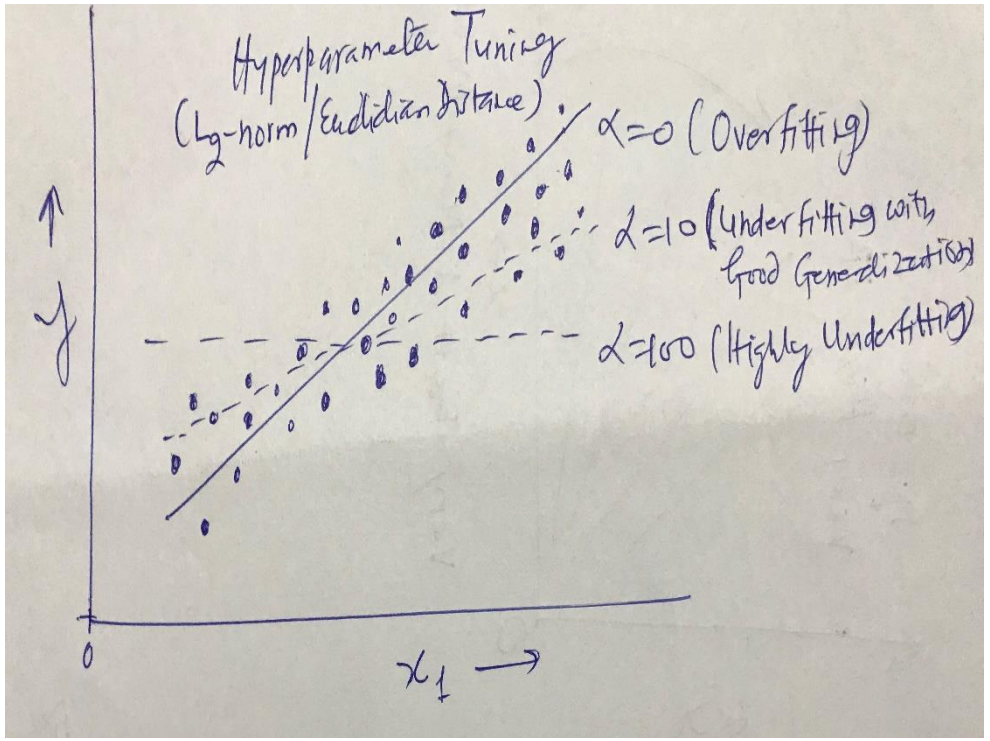
- Hyper-parameter Tuning: Grid Search Method
  - Step1: Divide the range of α into interval segments and evaluate for the RR for each interval mean value of α. For some of the mean α values of the grids, the RR model shows good generalisation. Stop at that value of α.
  - Step2: Next divide that particular range of the grid into subsections and repeat the above step 1.
  - Step3: Continue until satisfactory generalisation is achieved.
- How to check for Generalisation:
  - For each α, compute the SSE for training and validation sets without adding the Regularisation term and compare them for equality.
  - If they are nearly same order of magnitude, then the overfitting is avoided.
  - If not satisfied then try further regularisation hyper parameter by Grid Search method.

Unit 2: Hyper-parameter Tuning

- Hyper-parameter Tuning: Random Search Method
  - Step1: Choose some random values of α and for each one check the regularisation of the RR model.
  - Step2: Find the two consecutive random α values between which good generalisation occurs.
  - Step3: Choose some random α value within this range and repeat the above steps.

Note:

- If the range is small, then grid search method for that particular range may tried.
- It may happen that the perfect value of α may never be reached. Always there is a tolerance range of α. Thus, some uncertainty of the LR model remains.
- The RR model equation may be determined either by method of (i) Normal Equations solving, or (ii) Gradient Descent Method.



Applied Machine Learning

Unit-2: Regression Techniques

9

Unit 2: Non-Linear Regression

- We have seen how to estimate a curve/line of regression when a linear relationship exists between y and X.
- Also seen how to use scattergrams to know the nature of the relationship between y and X.
- But, there can be cases where linear relationships do not exist and linear/multilinear assumption may not be appropriate.
- This is nor easy as there are many types of non-linear relationships exist.
- But there are some useful trick-of-the trade solutions based on scattergrams.
- Inspection of scatter grams are used to determine whether a non-linear model is required for fitting.

Unit 2: Non-Linear Regression

- There are three main types of non-linear models which fit most of the non-linear data sets.
- They are (i) Exponential model, (ii) Power Law model, and (iii) Reciprocal model.

(i) Exponential Model is give by:

$$\mu_{y|x} = \beta_{0*} e^{\beta 1 x} \quad \text{(for } x > 0\text{)}$$

$$y_i = \beta_0 * \exp(\beta_1 x) * e_i$$

(ii) Power law model is given by:

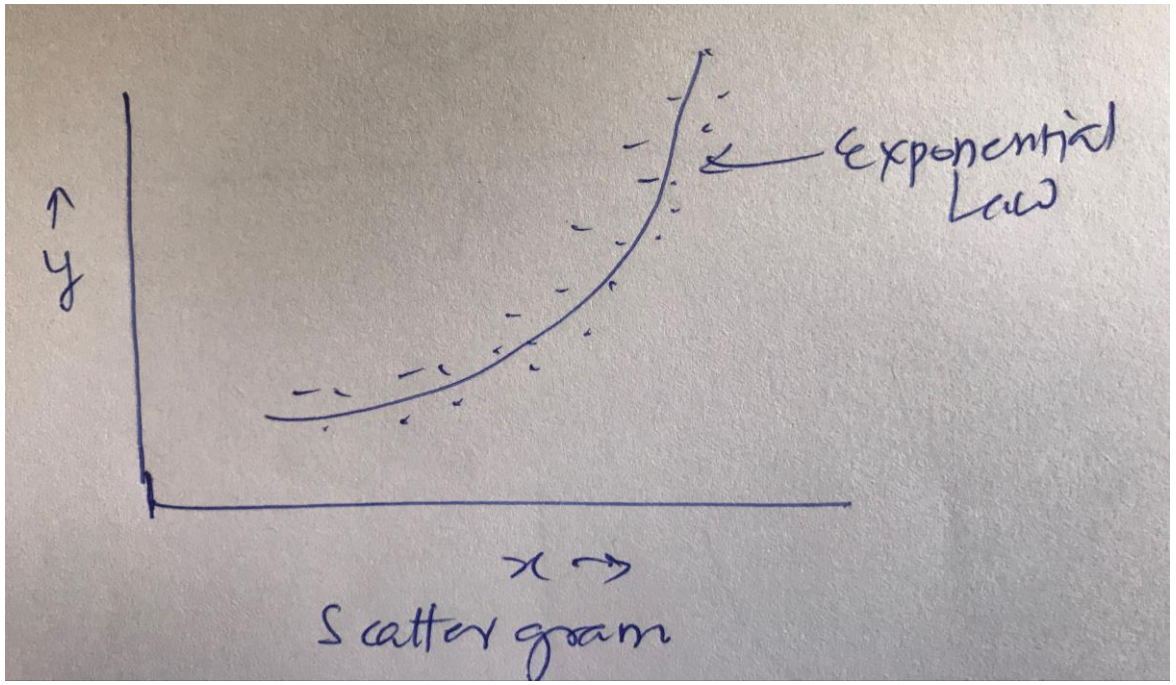
$$\mu_{y|x} = \beta_0 x^{\beta 1} \quad \text{(for } x > 0\text{)}$$

$$y_i = \beta_0 * x^{\beta 1} * e_i$$

(iii) Reciprocal Model:

$$\mu_{y|x} = \beta_0 + \beta_1 * (1/x) \quad \text{(for } x > 0\text{)}$$

$$y_i = \beta_0 + \beta_1 * (1/x_i) + e_i$$



- Note in the above first two models, the random error is not added, but multiplied.
- In third model, it is added.

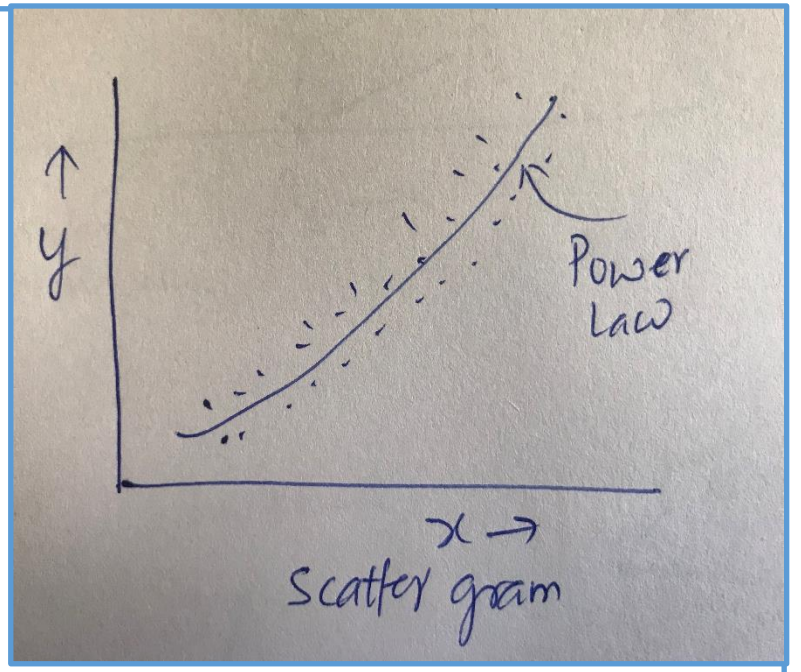
Unit 2: Non-Linear Regression

- Note that although the first two models (viz. Exponential and Power laws) are non-linear.
- But they are intrinsically Linear, i.e. they can be transformed or rewritten in an equivalent Linear form. This process is called Linearization.
- **Linearization of EXPONENTIAL Model**:
  - By natural logarithmic transformation, the exponential model can be written as:
    $$\ln(y_i) = \ln(\beta_0) + \beta_1 \ln(x_i) + \ln(e_i)$$
  - Using the following definitions:
    $$y_i^* = \ln(y_i),\ \beta_0^* = \ln(\beta_0),\ \beta_1^* = \beta_1, \text{ and } e_i^* = \ln(e_i)$$
  - We can write the model as:
    $$y_i^* = \beta_0^* + \beta_1^* x_i + e_i^*$$

  This is a single variable linear regression model.
- Note: Both β_0 and y_i have to be positive for log transformation to work. And as $x > 0$, the y_i increases with x in this model form.
- After the linearized form of the regression equation, the estimates of β_0 and y have to be obtained by taking anti-log (to natural basis e) of β_0^* and y^* .

Unit 2: Non-Linear Regression

- **Linearization of POWER LAW Model**:
  - By natural logarithmic transformation,
    the exponential model can be written as:
    $\ln(y_i) = \ln(\beta_0) + \beta_1 \ln(x_i) + \ln(e_i)$
  - Using the following definitions:
    $y_i^* = \ln(y_i)$, $\beta_0^* = \ln(\beta_0)$, $\beta_1^* = \beta_1$, $x_i^* = \ln(x_i)$,
    and $e_i^* = \ln(e_i)$
  - We can write the model as:
    $y_i^* = \beta_0^* + \beta_1^* x_i^* + e_i^*$

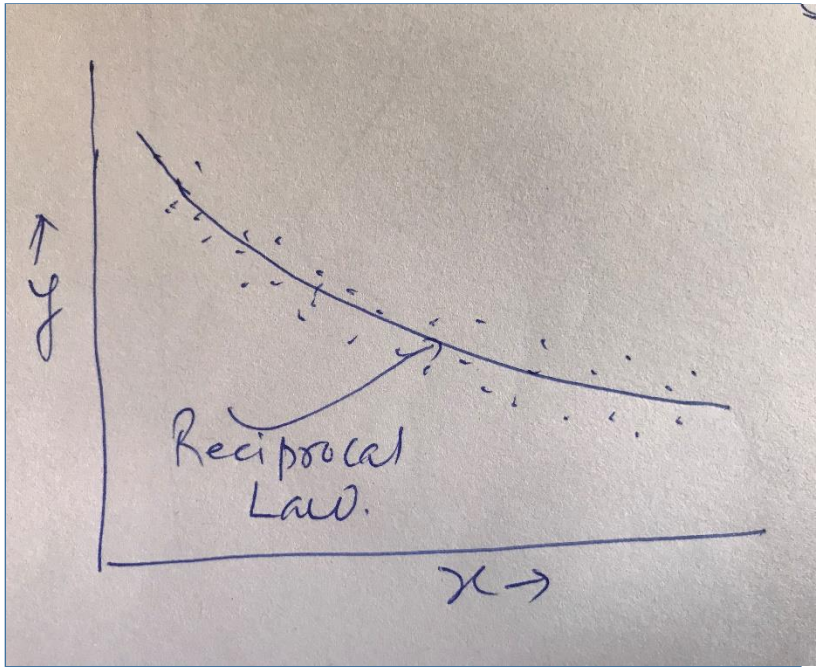


This is also a single variable x* linear regression model.

- **Note:**
  - Both β_0 , x_i and y_i have to be positive for log transformation to work.
    And as $x > 0$, the y_i increases with x in this model form.
  - After the linearized form of the regression equation, the estimates of β_0, x and y have to be obtained by taking anti-log (to natural basis e) of β_0^*, x* and y* .
  - In both the methods described above, β_0^*, x* ,y* and e* are estimates by Least Square Fitting method by solving Normal Equations.

Unit 2: Non-Linear Regression

- **Linearization of Reciprocal Model**:



- The Reciprocal Model is given by:

  $\mu_{y|x} = \beta_0 + \beta_1 * (1/x)$ (for $x > 0$)

  $y_i = \beta_0 + \beta 1 * (1/x_i) + e_i$

- This model is already linear model, but with 1/x as independent variable rather than x itself.

- We substitute z for 1/x for Linearization:

  $y_i = \beta_0 + \beta_1 z_i + e_i$

  This is also a single variable z linear regression model.

- After the linearized form of the regression equation, the estimates of β_0, β_0 are estimates by List Square Fitting method by solving Normal Equations.

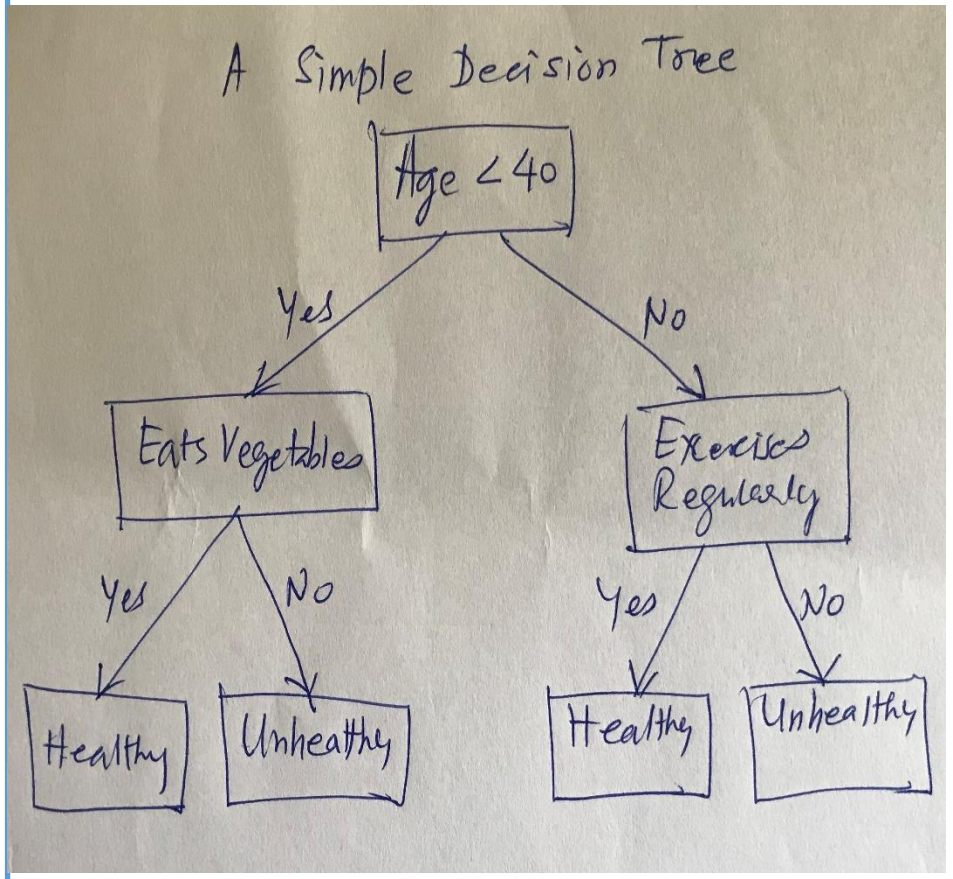
Unit 2: Non-Linear Regression

Note:

- One of the basic assumptions of linear regression is that of equality of variance over the large range of the data of x-values.
- In case this is not valid (by as seen in scatter gram), then take the logarithmic transformation of y to help stabilize the variance.
- Then, replace y by ln(y), which is = y* and then regress against X.
- Then the model takes the form $\mu_{y|x} = \beta_0 + \beta_1 X$. The $\mu_{y|x}$ is the mean value of y* for a given x.
- Once β_0 and β_1 are obtained by Least Square Fit method, then from Y* we can get Y from $Y = e^{Y^*}$

Unit 2: Decision Tree

- A Decision Tree is a machine learning supervised classification algorithm for both the Classification and Regression problems.
- Independent variable is continuous number or a categorical variable.
- A Decision Tree is a inverted treelike structure representing the variables in the form of nodes.
- A Simple Decision Tree shown in the adjoining figure.
- Variables are Age, Eats Vegetables, Exercises Regularly. And the dependent variables are Healthy and Unhealthy.
- Age is Root Node, Eats Vegetables and Exercises Regularly are Decision Nodes, and lastly Healthy and Unhealthy are the Leaf Nodes/ Terminal Nodes.
- Dependent variable occupies the terminal node. The most important independent variable occupies the Root Node
- All other independent variables occupy the positions of Decision Nodes.



Unit 2: Decision Tree Regression

- Criteria for Splitting Decision Tree:
  - The most commonly applied metric for generation of Decision Tree are: (i) Gini impurity, (ii) Entropy (a measure of impurity), (iii) RMSE, (iv) Information Gain.
  - The variable that gives maximum information helps in classification.
  - The lesser the Gini Impurity or entropy or RMSE, better the classification accuracy.
  - The most commonly used technique for a Decision Tree generation is CART (i.e. Classification And Regression Technique).
- CART uses Gini impurity index.

General Points Relating to Decision Tree:

(i) Not all independent variables form part of the Decision Tree.

(ii) Of all independent variables, the most significant variable occupies the Root Node position.

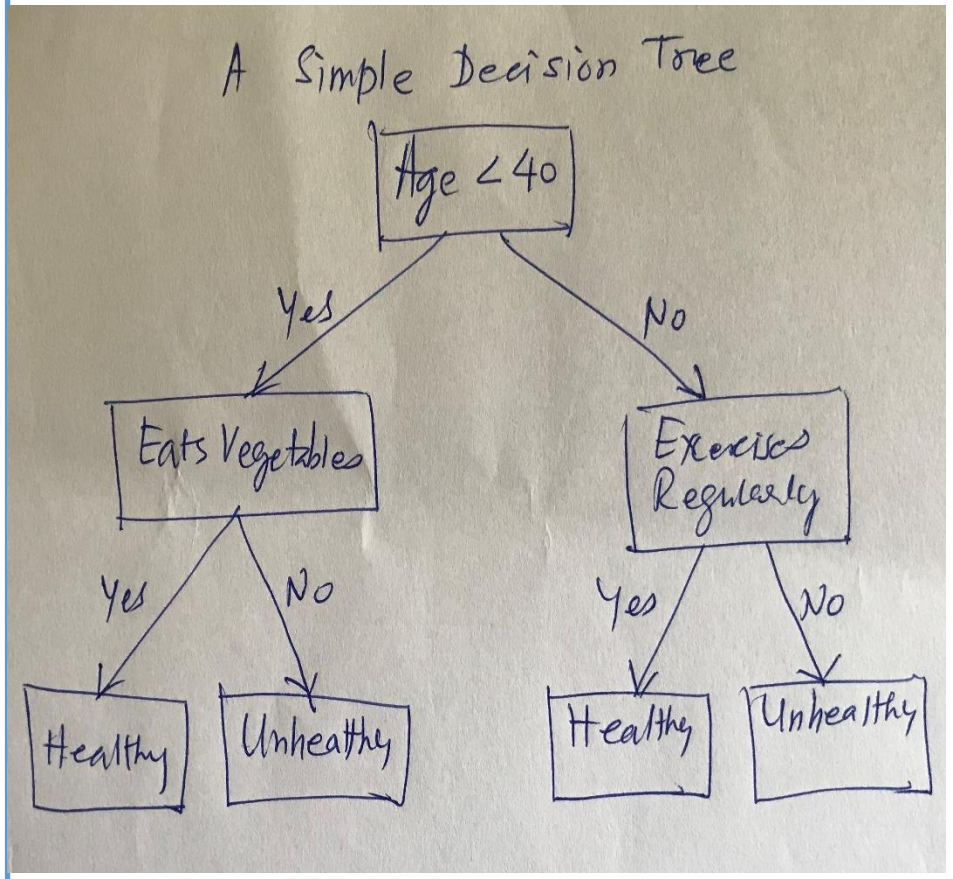
(iii) The Root Node/Decision Nodes have outputs of Yes or No. Yes is shown always on left side and No is shown on right side.

Unit 2: Decision Tree Regression

- The Decision Trees are very versatile Machine Learning Algorithms for both classification and regression.
- And also for multi-output tasks.
- They can even handle fitting of complex data sets.
- They are also fundamental components of Random Forest Algorithms.
- One of the many qualities of the Decision Trees is that they require very little data preparation. Also, they do not require features scaling or centralising at all.
- The Decision Tree regression is based on the numerical data with numerical values in the Main Node, Decision Nodes at all levels/depth and the Leaf Nodes.

Unit 2: Decision Tree

- A Decision Tree is a machine learning supervised classification algorithm for both the Classification and Regression problems.
- Independent variable is continuous number or a categorical variable.
- A Decision Tree is a inverted treelike structure representing the variables in the form of nodes.
- A Simple Decision Tree shown in the adjoining figure.
- Variables are Age, Eats Vegetables, Exercises Regularly. And the dependent variables are Healthy and Unhealthy.
- Age is Root Node, Eats Vegetables and Exercises Regularly are Decision Nodes, and lastly Healthy and Unhealthy are the Leaf Nodes/ Terminal Nodes.
- Dependent variable occupies the terminal node. The most important independent variable occupies the Root Node
- All other independent variables occupy the positions of Decision Nodes.



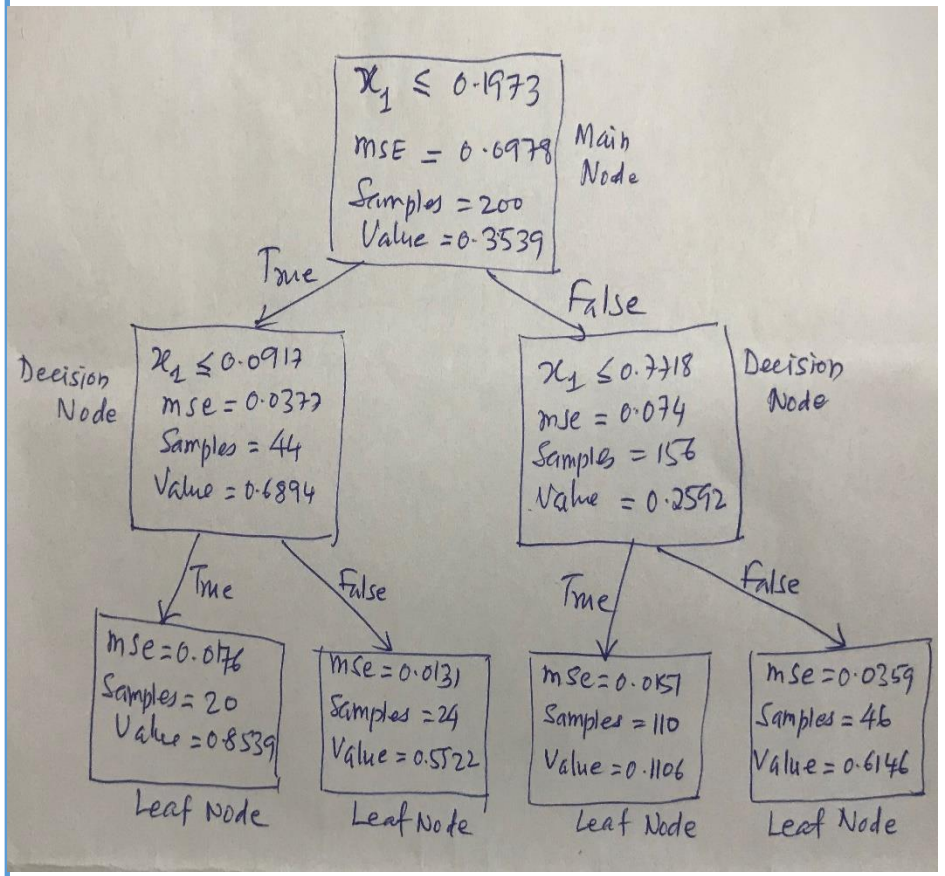
Applied Machine Learning

Unit-2: Regression Techniques

10

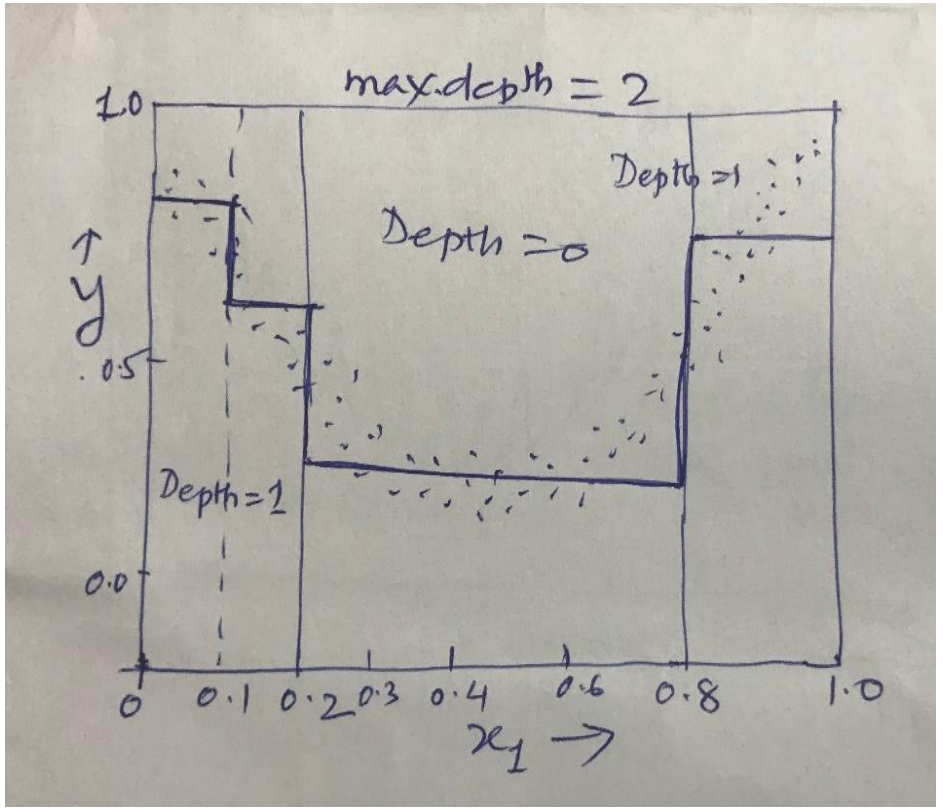
Unit 2: Decision Tree Regression

- In case of Regression, the Decision Trees are generated from the Y and X training data which are numerical and continuous variables.
- A typical Decision Tree based on all the numerical data is shown in the figure.
- Note, all the rules applied to any decision tree containing nominal or multinomial, logical variables are also applicable to this type of decision tree.
- In the Decision Tree based model, the average values of data in each Leaf Node are treated as dependent variable and the data in the other nodes as Independent variables.
- Each respective path in the Decision Tree from root node to leaf node represent a regression model (but not in usual sense).



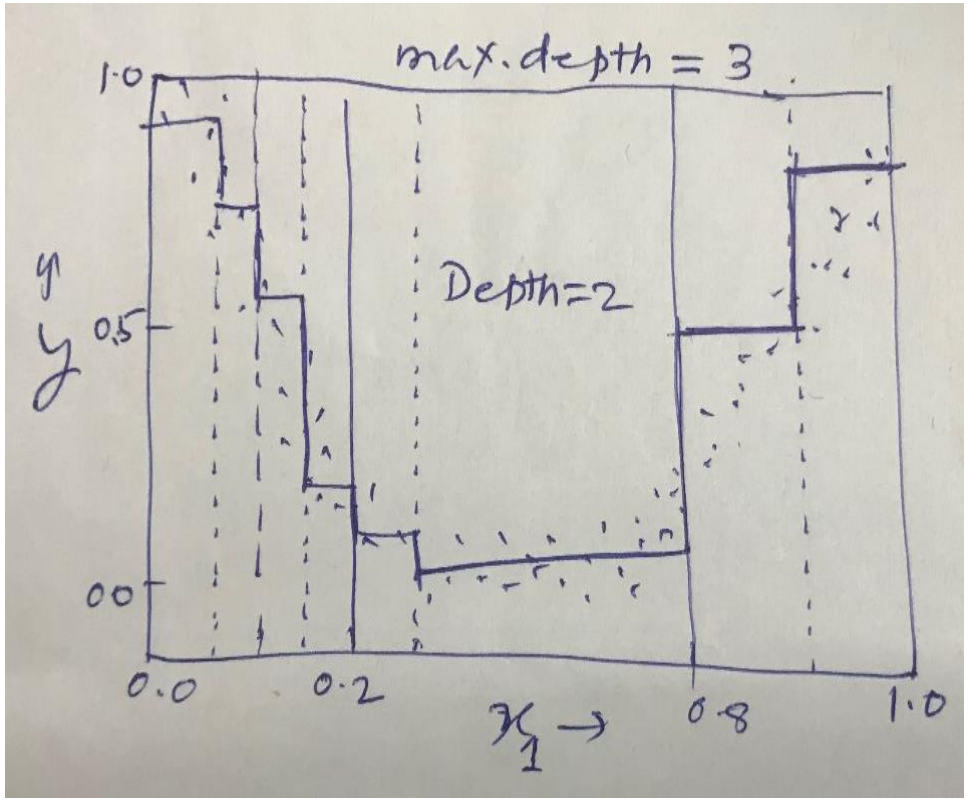
Unit 2: Decision Tree Regression

- A Decision Tree used for regression works very much similar to the Classification Tree.
- The main difference is that instead of predicting class in the each node, it predicts values in the leaf nodes.
- The predicted value is simply the average target value of training instances associated with the respective leaf node.
- This prediction results in SSE or MSE indicated in the respective leaf node.



Unit 2: Decision Tree Regression

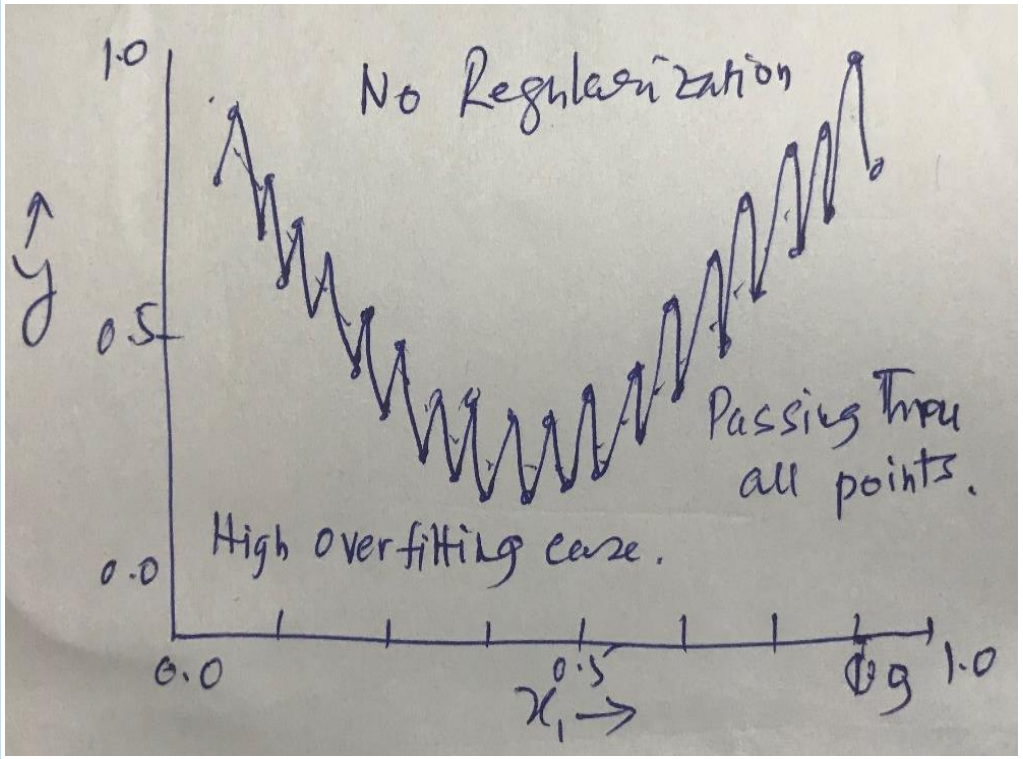
- The model's predictions are shown in the figure in case of tree with max. depth of 2.
- Note the vertical dividing lines correspond to criteria values in Decision nodes.
- And horizontal lines correspond to the predicted averages in the leaf nodes.
- In case of max. depth = 3, the predictions are as shown in the second figure. It has more vertical and horizontal lines.



Unit 2: Decision Tree Regression

Regularization of Decision Tree Regression: Case of Overfitting

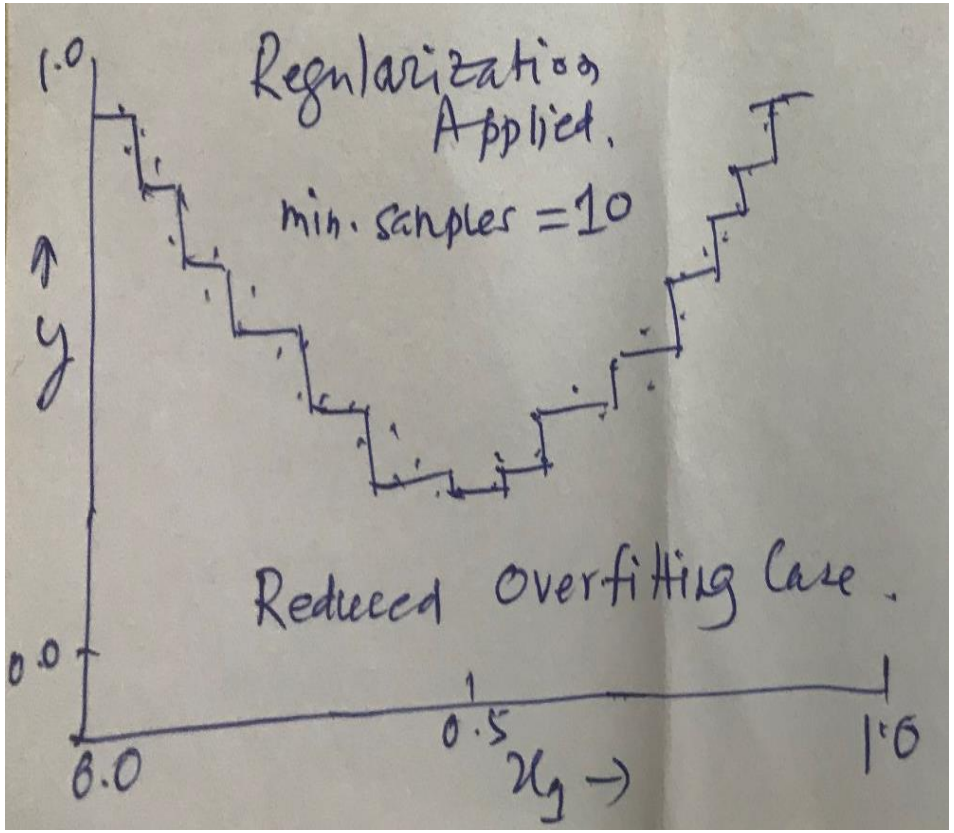
- If no regularization is done, then the regression uses the default parameters.
- Then the fitting is completely perfect.
- In this case the regression curve passes through all points, as shown in the top figure.
- This is case of very high overfitting. Such a model does not generalize very well on the testing data.



Unit 2: Decision Tree Regression

Regularization of Decision Tree Regression:

- To regularize this overfitting, the hyper parameter of minimum samples in a leaf (named 'min-samples_in_a_leaf' in CART algorithm with default value of 1) needs to be tuned.
- By increasing the value of hyper parameter, the problem of overfitting can be reduced.
- If the hyper parameter is set to 10, for example, the overfitting reduces as shown in the lower figure.
- Another way to reduce overfitting is to reduce the max. depth hyper parameter of the tree.



Unit 2: Decision Tree Regression

The CART (Classification And Regression Technique) Algorithm:

- The CART algorithm (Scikit-Learn in python library) tries to split the training data set in such a way that minimises the cost function associated with each of the leaf node.
- CART cost function for regression is given by:

$J(k,t_k) = \frac{n_{of}}{n}(SSE)_{of} + \frac{n_{uf}}{n}(SSE)_{uf}$

where $(SSE)_{node} = \sum_{i=1}^{n} (\hat{y}_{node} - yi)2$ and

$\hat{y}_{node} = (1/n_{node}) \sum_{i=1}^{n} (y_i)$,

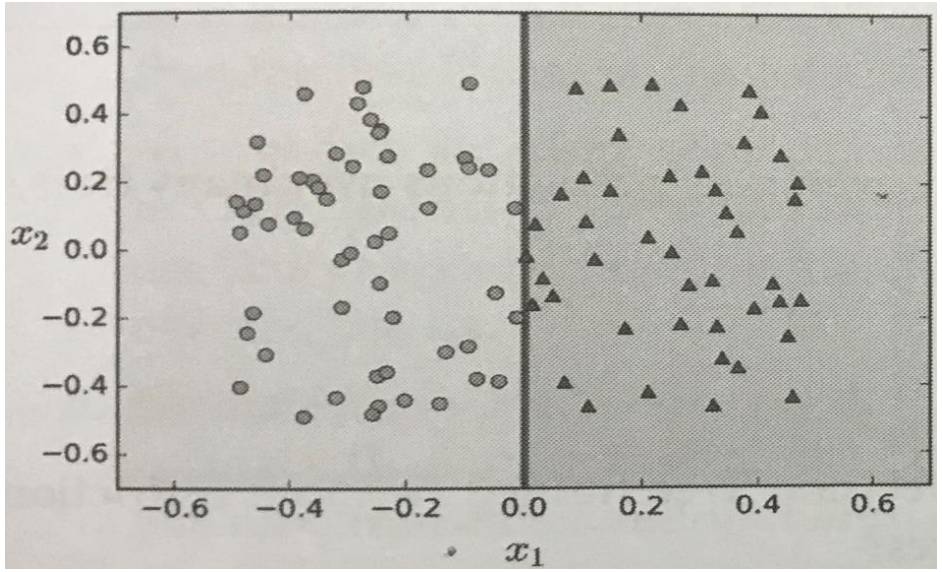
of = overfitting and uf = underfitting.

- The decision trees are usually very prone to overfit the training data.
- This is an unacceptable situation.

Unit 2: Decision Tree Regression

Limitations of the Decision Tree Regression:

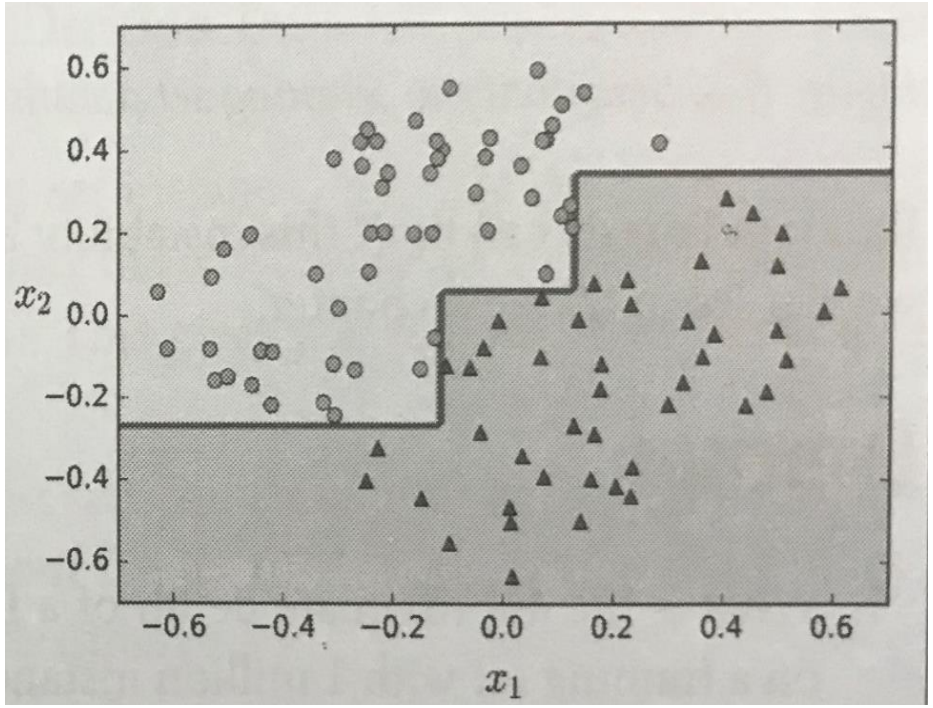
- Decision Trees have orthogonal decision boundaries (i.e. the boundaries perpendicular to each other).
- This makes it sensitive to rotation applied to training data set, if any.
- As shown in the figure, rotation of data may increase in the decision boundaries.
- One way to limit this problem is to use Principal Component Analysis (PCA transformation)



Unit 2: Decision Tree Regression

Limitations of the Decision Tree Regression:

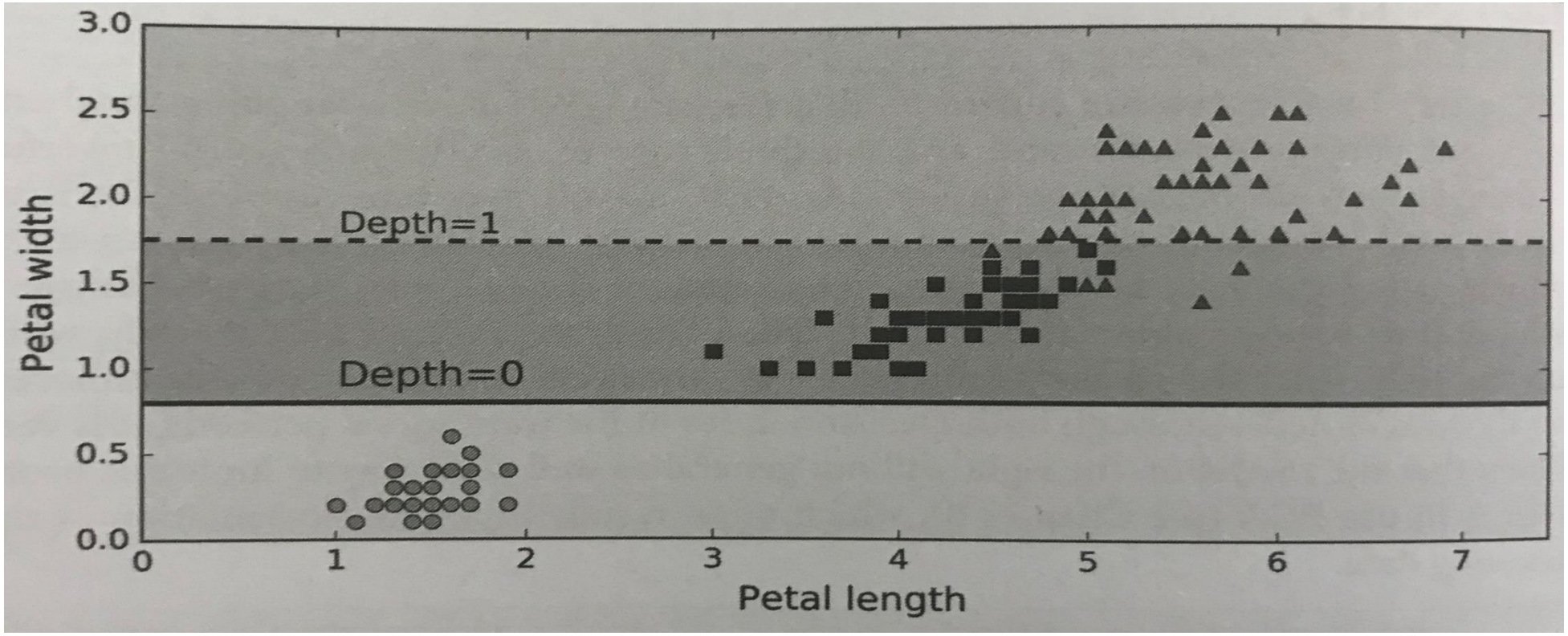
- The PCA may result into better orientation of the data spread.
- The Decision Trees are very sensitive to small variations in the training data values.
- For example, the removal of some data points (especially boundary points) may change the model significantly.
- Moreover, if the underlying modelling algorithm is Stochastic, it may lead to very different models even if it is based on the same data set.



Unit 2: Decision Tree Regression

Limitations of the Decision Tree Regression:

- Moreover, if the underlying modelling algorithm is Stochastic, it may lead to very different models even if based on the same data set.



Applied Machine Learning

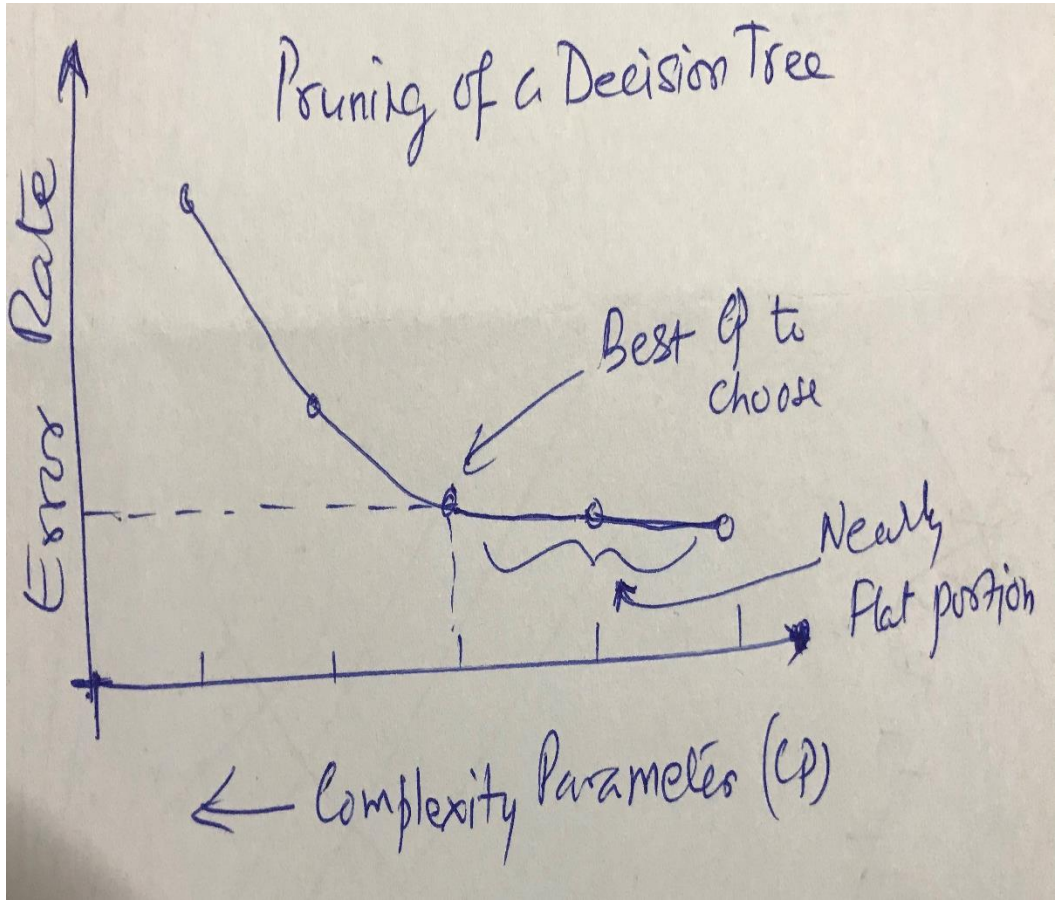
Unit-2: Regression Techniques

11

Unit 2: Pruning the Decision Tree

Pruning the Decision Tree

- Pruning is done based on the Complexity Parameter (CP).
- CP is used to determine to determine the number of branches in the Decision Tree.
- The aim is to set CP value such that the Classification Error is minimum or the least.
- The error rate is minimum when CP = 0. This means the tree does not require any pruning.
- Pruning should be done when there is possibility of reducing error.
- However, at CP = 0, the model may be overfitting the training data. complexity



Unit 2: Pruning the Decision Tree

- Allowing the Tree to grow as many branches as possible is not always a practical approach
- Initially, the CP value is set to zero, allowing the Tree to grow as many branches as possible.
- Decision Tree models may give better results with training data, but may not give equally good results with the testing data (the situation of Overfitting).
- On the other hand, the Decision Tree model may NOT perform well on both Training and Testing data sets (A case of underfitting).
- In such a situation, the pruning (cutting down branches of the Decision tree) avoids the problem of overfitting.
- Pruning tree involves Reducing the number of Decision Nodes, and also reducing the depth of the decision tree.

Unit 2: Cross-Validation of Decision Tree

- Steps in Multiple Fold Cross-Validation of Decision Tree
  - Step1: The training data is split into P number of equal data sets. P is usually is around 10).
  - Step2: The $(P-1)^{th}$ split data is used as a training data set and the other data set from the P^{th} split is used as a test data set.
  - Step3: Next the $(P-1)^{th}$ set is used as test data and the remaining data as training data.
  - Step4: Above steps are repeated P number of times. Then, the average of the P fold validation accuracies are determined.
- This is known as P-fold cross validation.
- Every single data sample/example will be part of training data (P-1) times and will be part of test data at least once.
- In this way, the Decision Tree model is rigorously trained multiple times to make model strong/robust for prediction.

Unit 2: Evaluation of Regression Models

- **The correlation coefficient R is a measure of relation between the dependent variable and independent variables.**
- **The independent variables are known as explanatory variables.**
- **If R = 0, there exists no relations. The regression line or plane/hyperplane is flat.**
- **The R^2 value, which is always positive, is a measure of how much is variance is explained by the model.**
- **If R^2 is 0.80, then 80 % of the variance in the data is explained or accounted by the model.**
- **The quantity $(1 – R^2)$ explains the residual of the variance which remains unexplained. This quantity is known as Coefficient of Determination.**

Unit 2: Measures of Regression

- **How good is the Least Square Fitting is determined by the minimum/optimum value of the SSE achieved.**
- **The least Square Fitting is measured by correlation coefficient R which is related to minimum SSE achieved.**
- **In case of Linear Regression equation with single variable x, the R is given by:**
- $$R = \frac{\sum_{i=1}^{n}(xi - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(xi - \bar{x})^2 \sum_{i=1}^{n}(yi - \bar{y})^2]}}$$

  Where x and y are the means of X and Y, respectively.

  The R ranges from -1 to +1, through 0.
- **In case of multiple linear regression, the R^2 is defined as:**

$$R^2 = \frac{(SSE)_{Explained}}{(SSE)_{Total}}$$

- **The $(SSE)_{Total}$ is related to both $(SSE)_{Explained}$ and $(SSE)_{Residual}$:**

  $(SSE)_{Total} \simeq (SSE)_{Explained} + (SSE)_{Residual}$

Unit 2: Coeff. Of Determination and SEE

- As said above $(1 - R^2)$ is measure of unexplained variance.
- The R^2 is related to explained and residual variances.
- It is given by:

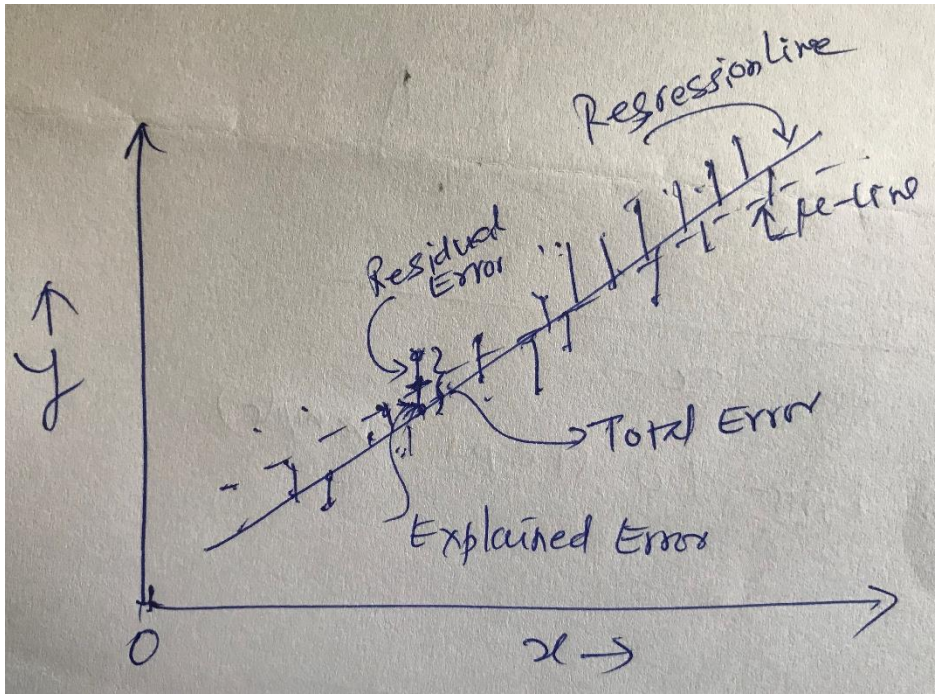
$$R^2 = \frac{\text{Explained } Variance}{\text{Total } Variance} = \frac{(SSE)_{Exp}}{(SSE)_{Total}} = \frac{(SSE)_{Exp}}{(SSE)_{Exp}+(SSE)Residual}$$

$$= \frac{(SSE)_{Exp}}{(SSE)_{Exp}+(SSE)_{Residual}} \simeq \left[1 - \frac{(SSE)_{Res}}{(SSE)_{Exp}}\right]$$

- As $R \rightarrow 1$, the ratio $\frac{(SSE)_{Res}}{(SSE)_{Exp}} \rightarrow 0$.
- Coeff. of Determination $= (1 - R^2) = \frac{(SSE)_{Res}}{(SSE)_{Exp}}$
- The standard error of estimation of mean (of y) SEE is given by:

$$\text{SEE of Mean} = \text{SSE x (Coeff. of Determination)}$$

$$= \text{SSE}\sqrt{(1-R^2)} = \text{SSE}\sqrt{\left\{\frac{(SSE)_{Res}}{(SSE)_{Exp}}\right\}}$$



Unit 2: Loss Function

- **The LOSS FUNCTION is the distance between the most recent/current output of an algorithm and the expected output.**
- **It is a method to evaluate the performance of an algorithm (i.e. how good the algorithm models the data).**
- **The Loss Function can be categorised in to two groups.**
- **One group for Regression models and second group for Classification models.**
- **In case of Regression modelling, the dependent variable is continuous.**
- **In case of classification, the dependent variable is discrete/nominal.**
- **The commonly used Loss Functions are (i) Cross-entropy, (ii) Log Loss, (iii) Exponential loss, (iv) Hinge Loss, (v) Kullback Leibler Divergence Loss, (vi) Mean Square Error (MSE, l_2-fold), (vii) Mean Absolute Error (MAE, l_1-fold), and (viii) Huber Loss.**

Unit 2: MSE Loss Function

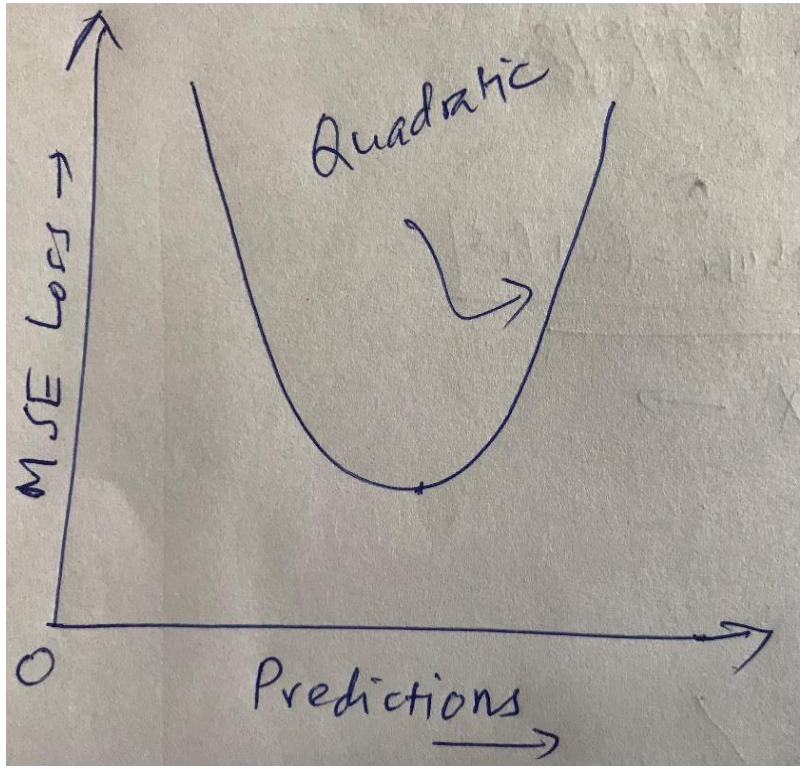
Evaluation of Regression Models:

- **SSE loss, MSE Loss and RMSE Loss (l_2-regularisation):**

  SSE = $\sum_{i=1}^{n} (\hat{y} - y_i)^2$ **,** **MSE =** $\frac{1}{n} \sum_{i=1}^{n} (\hat{y} - y_i)^2$ **, and**

  RMSE = $\sqrt{(MSE)}$

- **Here** $\hat{y}$ **is regression model estimated value and** y_i **is the measured value for given** x_{ji}**, and i = 1 to n and j = 1 to m.**
- **All the SSE, MSE and RMSE are sensitive to outliers, because the difference are squared which gives importance to the outliers.**
- **Behaviour of the MSE loss function is a quadratic function often useful for Gradient Descent algorithm. The Gradient will be smaller close to the minimum.**

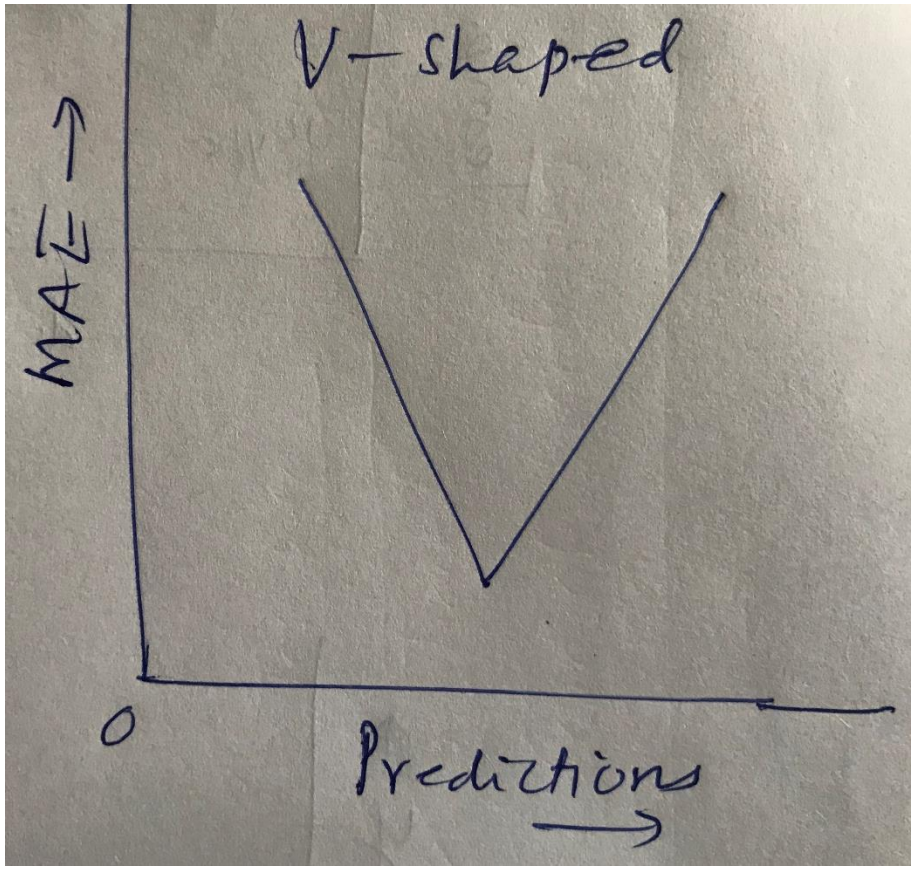


Unit 2: MAE Loss Function

Mean Absolute Error (MAE) is defined as:

- **MAE =** $\frac{1}{n}\sum_{i=1}^{n} |\hat{y} - y_i|$

  (l_1 i.e. Manhattan Distance/norm based)
- **The MAE is not sensitive to the outliers.**
- **The MAE loss function has V-shaped curve unlike the MSE function.**
- **The MAE function is more robust to outliers.**
- **It is like Median, i.e. outliers cannot really impact MAE**
- **The Gradient Descent is same at each point on the V-Curve of MAE.**



Unit 2: Huber Loss Function

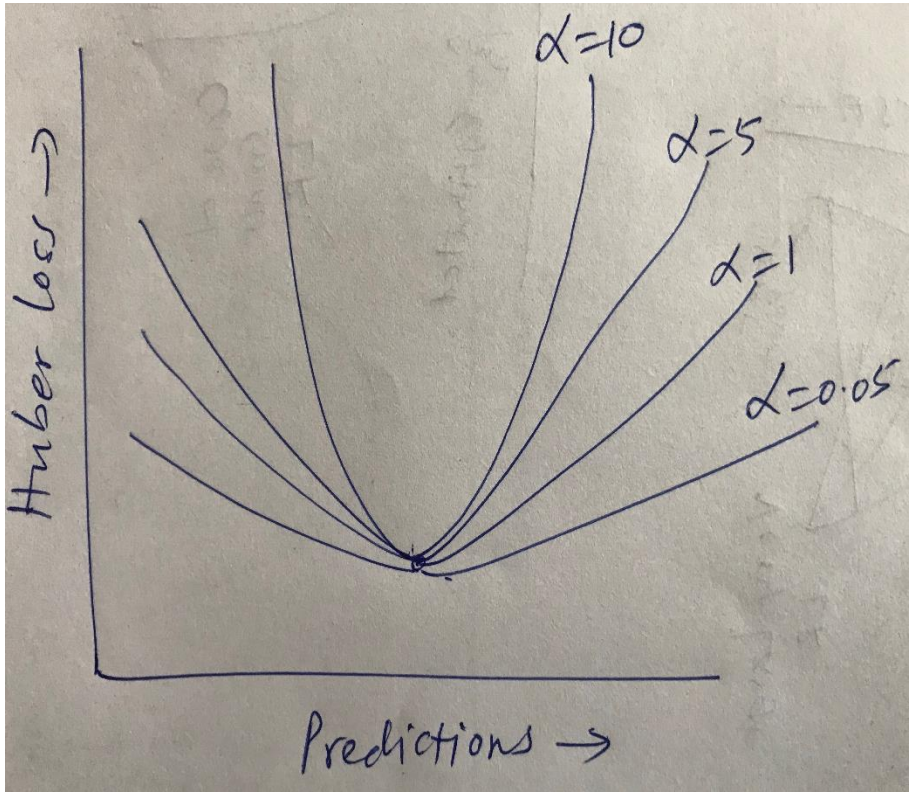
Huber Loss:

- The Mean Square Logarithmic Error is known Huber Loss is a combination of MAE and MSE (l_1, l_2 respectively).
- The response variable is transformed as: y ← log(y) to reduce the range or variability of y values.
- Huber Loss function L_α is given by:

$L_\alpha = \frac{1}{2}\sum_{i=1}^{n}(\hat{y} - y_i)^2$ for $y - f(x) \le \delta$ and

$= \alpha \sum_{i=1}^{n} |\hat{y} - y_i| - \frac{1}{2}\alpha^2$

- It depends on the additional parameter called α that influences the shape of the loss function.
- The hyperparameter α needs to be fine tuned by the algorithm.
- When the values are large (far from the minimum), this function has behaviour like that of MAE and when close to the minimum, this function behaves like MSE.
- The α- parameter is sensitive to outliers.
- It is supposed to reduce the influence of outliers.



Applied Machine Learning

Unit-2: Regression Techniques

End