Introduction to Data Science

This suggestion is prepared by Mustakim Billah Bedar (Theory) and Mahmuda Rahman (Math)

Subject: IDS

➢ Big data can be characterized by volume, velocity and variety- justify the statement with some relevant examples.

Big Data is any data that is expensive to manage and hard to extract value from

- Volume: The size of the data

Facebook, for example, stores photographs. That statement doesn't begin to boggle the mind until you start to realize that Facebook has more users than China has people. Each of those users has stored a whole lot of photographs. Facebook is storing roughly 250 billion images.

So, in the world of big data, when we start talking about volume, we're talking about insanely large amounts of data.

- Velocity: The latency of data processing relative to the growing demand for interactivity

Facebook users upload more than 900 million photos a day. A day. So that 250 billion number from last year will seem like a drop in the bucket in a few months. Velocity is the measure of how fast the data is coming in. Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and somehow, later, be able to retrieve it.

- Variety and Complexity: The diversity of sources, formats, quality, structures.

Take, for example, email messages. A legal discovery process might require sifting through thousands to millions of email messages in a collection. Not one of those messages is going to be exactly like another. Each one will consist of a sender's email address, a destination, plus a time stamp. Each message will have human-written text and possibly attachments.

Photos and videos and audio recordings and email messages and documents and books and presentations and tweets and ECG strips are all data, but they're generally unstructured, and incredibly varied. All that data diversity makes up the variety vector of big data.

➢ Define data science. Which subjects are covered by data science? Why data science can be said as a multidisciplinary subject?

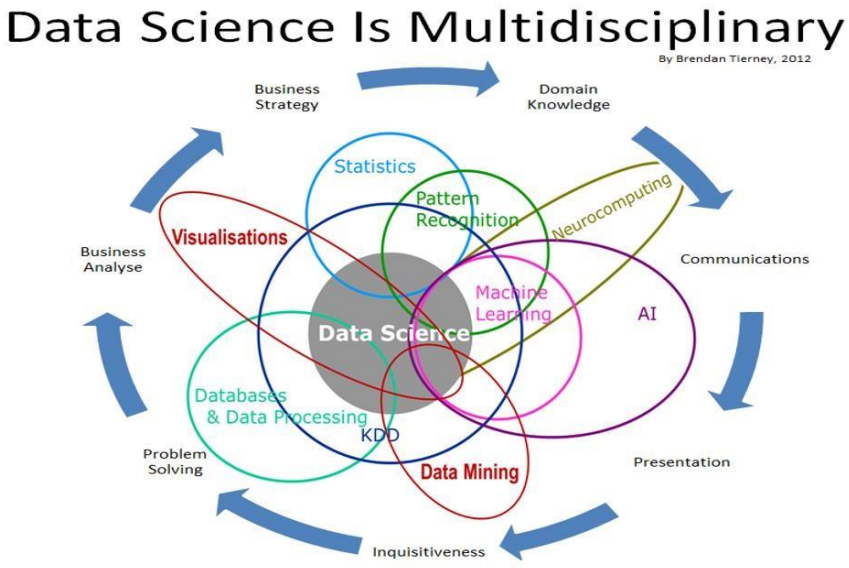
Data Science: An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data.

Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data.

Subjects covered by data science:

- Fundamental of Statistics and its types.
- Hypothesis Testing
- Fundamentals of Probability
- Linear Algebra
- Analysis of Variance & Covariance.
- Programming Python, R, SAS etc.
- Machine Learning & Deep Learning
- Visual Analytics
- Mathematical Modeling Techniques
- Simulation and Modeling Techniques
- Business Analytics and Data Mining

Data Science as a multidisciplinary subject:



- **What are the roles of a data scientist? How does a data scientist apply data science techniques to real world applications?**

Roles of Data Scientist:

- Work with stakeholders to determine how to use business data for valuable business solutions
- Search for ways to get new data sources and assess their accuracy
- Browse and analyze enterprise databases to simplify and improve product development, marketing techniques, and business processes
- Create custom data models and algorithms
- Use predictive models to improve customer experience, ad targeting, revenue generation, and more
- Develop the organization's test model quality and A/B testing framework
- Coordinate with various technical/functional teams to implement models and monitor results
- Develop processes, techniques, and tools to analyze and monitor model performance while ensuring data accuracy
- A natural inclination toward solving complex problems
- Knowledge/experience on/with statistical programming languages, including R, Python, SLQ, etc., to process data and gain insights from it
- Experience using and developing data architectures
- Knowledge of Machine Learning techniques, including decision tree learning, clustering, artificial neural networks, etc., and their pros and cons
- Knowledge and application experience in advanced statistical techniques and concepts, including, regression, distribution properties, statistical testing, etc.
- Experience/knowledge in statistics and data mining techniques, including, random forest, GLM/regression, social network analysis, text mining, etc.
- Experience with major web services, including S3, Spark, Redshift, etc.
- Experience/knowledge in distributed data and computing tools, including, MapReduce, MySQL, Hadoop, Spark, Hive, etc.

Apply data science in real world applications:

- Identifying and predicting disease
- Personalized healthcare recommendations
- Optimizing shipping routes in real-time
- Getting the most value out of soccer rosters
- Finding the next slew of world-class athletes
- Stamping out tax fraud
- Automating digital ad placement
- Algorithms that help you find love
- Predicting incarceration rates

➢ According to Gartner, why do 50% of the projects fail? How the failure can be avoided?

According to Gartner 50% of projects fail because:

Budget and Schedule:

Runaway budget costs are behind one-quarter of project failures for projects with budgets greater than $350,000.

Features and benefits:

they fail to deliver the features and benefits that are optimistically agreed on at their outset.

Big Project:

Small is beautiful — or at least small projects are easier to manage and execute.

Avoiding Failure:

1. Build with Organizational Buy in
2. Build with end in mind
3. Build with structural approach

Look for ways to limit the size, complexity and duration of individual projects, and ensure funding has been committed.

Stay on top of costs, especially for the largest projects. Ensure that there are the appropriate mechanisms in place to identify budget variances and/or overruns early. Regularly review how cost estimation is done to understand how accurate and effective your approaches are, and pursue improvement opportunities.

Keep the schedule realistic. Many large projects fail because business conditions keep changing after the project scope has been set, leaving a significant disconnect between the agreed-on scope and budget versus what the business will require and pay for by the time the project is delivered.

Invest in truly capturing and understanding the business expectations and functionality sought from the project, and ensure that there is initial, adequate allocated funding, as well as good processes in place for revisiting the expectations and required funding at multiple points during the project.

Increase the frequency of project status and review meetings, as well as ongoing confirmation of the project's alignment with business strategy — with an eye toward identifying and cancelling projects at the earliest possible stage that no longer meet company needs.

Mustakim Billah Bedar

➢ Why business understanding is critical to develop a successful data science project? Why business objectives require to change into technical objectives?

Business understanding is critical to develop a successful data science project.

In business understand phase we basically do is:

- Understands the business process
- Define and Frame the business problem
- Define the business objective
- Agree on success criteria
- Its Specific, measurable and time-bound
- List assumptions, constraints, and important factors.
- Identify secondary or competing objectives.
- Study existing solutions (if any).

Why business objectives require to change into technical objectives:

Business understanding require to change into technical objective to apply data science tools and techniques to solve a particular problem.

An aim is an overall goal, and objectives are the steps needed to achieve it. As a business grows, its aims and objectives change. One of the main reasons for this is that market conditions change. The term 'market conditions' refers to the size of the market, the business's competitors, and the proportions of large and small businesses in the market. If a business is in a growing market, over time its aims and objectives may change to focus on growth. An example of this would be a company focused on sustainable products, such as biodegradable packaging, where demand is growing. If a business is in a market where there is suddenly an increase in competition, its aims and objectives may have to change to focus on survival, which is when a business aims to keep its day-to-day operations running.

As technology continuously evolves, business aims and objectives also change. Common technological developments include:

- website developments
- manufacturing developments
- software developments
- mobile technology developments
- contactless, online, and mobile payment system developments

➢ What do you understand by “inactivity” in a telecom service? How can you apply data science in telecom service for profit accumulation?

Inactiveness of customer from using the service provided by the service provider is called the inactivity in a telecom service.

Inactivity in a telecom service defines:

- Incoming and outgoing calls
- Data usage
- Incoming text
- Promotional texts
- Voicemail usage • Call forwarding
- Etc.

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn.

Applying data science in telecom service:

- Fraud Detection
- Predictive Analysis
- Customer Segmentation
- Real Time Analytics
- Price Optimization
- Customer Sentiment Analysis
- Lifetime value prediction
- Customer Churn Prevention
- Customer Segmentation

Mustakim Billah Bedar

- **How data aggregation method helps finding some useful information from a data sheet? Give some real-life example.**

Data aggregation is the process where raw data is gathered and expressed in a summary form for statistical analysis.

For example, raw data can be aggregated over a given time period to provide statistics such as average, minimum, maximum, sum, and count. After the data is aggregated and written to a view or report, you can analyze the aggregated data to gain insights about particular resources or resource groups. There are two types of data aggregation:

Time aggregation: All data points for a single resource over a specified time period.

Spatial aggregation: All data points for a group of resources over a specified time period.

Data Aggregation method helps finding following information from a data sheet:

- Number of transactions (Frequency)
- Days since the last transaction (Recency)
- Days since the earliest transaction (Tenure)
- Avg. days between transaction
- # of transactions during weekends
- % of transactions during weekends
- # of transactions by day-part (breakfast, lunch, etc.)
- % of transactions by day-part
- Days since last transaction / Avg. days between transactions Example:

Companies often collect data on their online customers and website visitors. The aggregate data would include statistics on customer demographic and behavior metrics, such as average age or number of transactions. This aggregated data can be used by the marketing team to personalize messaging, offers, and more in the user's digital experience with the brand. It can also be used by the product team to learn which products are successful and which are not. And furthermore, the data can also be used by company executives and finance teams to help them choose how to allocate budget towards marketing or product development strategies.

- **Describe some data munging techniques. "Features are very important elements in data science project; however, we often require to reduce them" – justify the statement.**

Data munging techniques:

1. **Percent missing values**
2. **Amount of variation**
3. **Pairwise correlation**
4. **Multicolinearity**
5. **Principal Component Analysis (PCA)**
6. **Cluster analysis**
7. **Correlation (with the target)**
8. **Forward selection**
9. **Stepwise selection**
10. **LASSO**
11. **Tree-based selection**
12. **Backward elimination**

Feature reducing necessity:

It is required for dimensionality and overfitting problem.

True dimensionality <<< Observed dimensionality: The abundance of redundant and irrelevant features

Curse of dimensionality: With a fixed number of training samples, the predictive power reduces as the dimensionality increases. [Hughes phenomenon]. With d binary variables, the number of possible combinations is O (2d).

Law of Parsimony [Occam's Razor]: Other things being equal, simpler explanations are generally better than complex ones.

Art is the elimination of the unnecessary– Pablo Picasso

Mustakim Billah Bedar

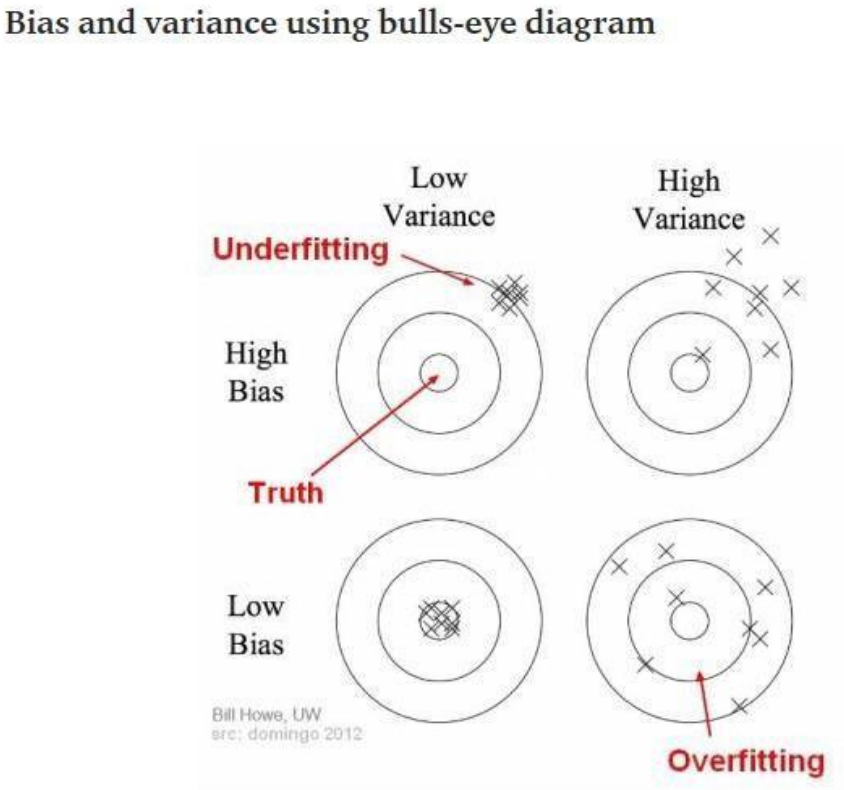
- **Write down the behavior of bias and variance in data. Why the tradeoff between bias and variance is required?**

Bias:

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance:

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.



Low Bias and Low Variance is Preferable.

Tradeoff between bias and variance:

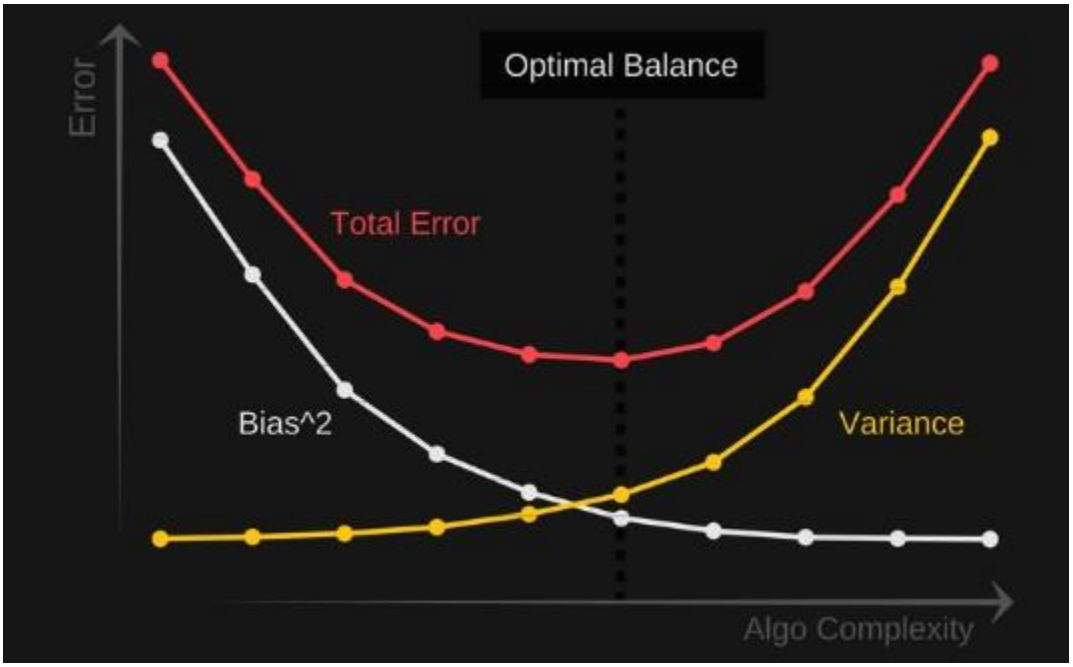
If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

Total Error

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

Total Error = Bias^2 + Variance + Irreducible Error



An optimal balance of bias and variance would never overfit or underfit the model. Therefore understanding bias and variance is critical for understanding the behavior of prediction models.

- **What do you mean by point estimate? Do you find any relation between point estimate and sampling distribution?**

Point Estimate:

In statistics, point estimation involves the use of sample data to calculate a single value (known as a point estimate since it identifies a point in some parameter space) which is to serve as a "best guess" or "best estimate" of an unknown population parameter (for example, the population mean). More formally, it is the application of a point estimator to the data to obtain a point estimate.

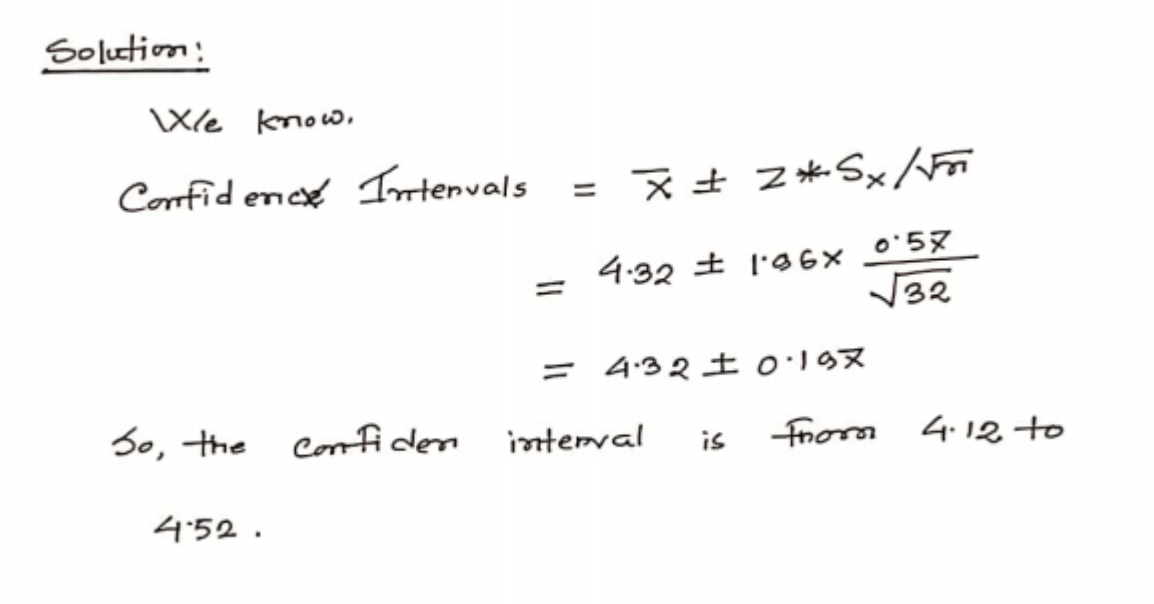
Relation between point estimate and sampling distribution:

The Probability distribution of statistic is called the sampling distribution A sampling distribution is a distribution of a statistic over all possible samples.

- Draw a sample from the population
- Calculate the point estimate
- Repeat the previous two steps many times
- Draw a frequency distribution of the point estimates
- That distribution is called a *sampling distribution*

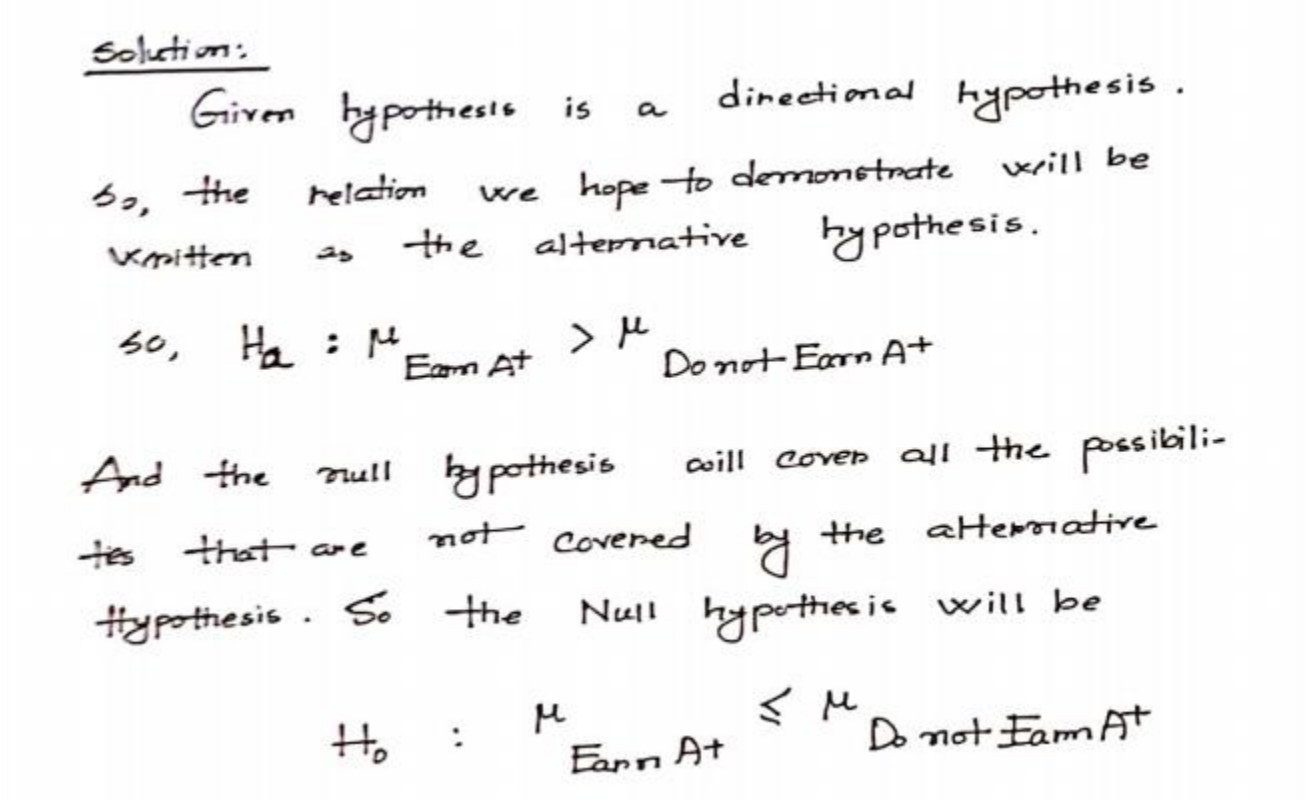
Mustakim Billah Bedar

- **The area under the normal curve between *z*-scores of -1.96 and +1.96 is .95; where mean, $\bar{X}$= 4.32, standard error of mean, $s_{\bar{X}}$=.57, and the number of samples, n = 32. Find out the confidence interval using central limit theorem.**



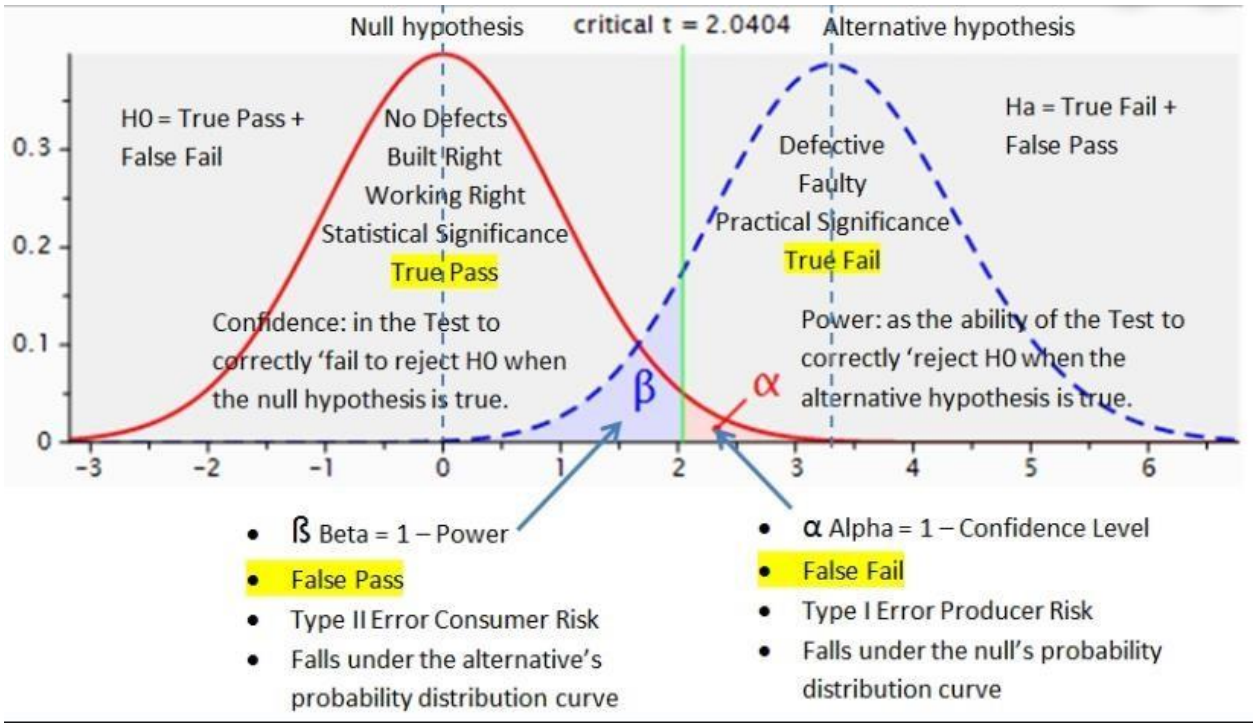
[This question is answered by Mahmuda Rahman]

- **Convert the following hypotheses into statistical hypotheses: "People who earn an A+ in Programming are more likely to get a high salary job than those who do not earn an A+".**



[This question is answered by Mahmuda Rahman]

- **Explain the below figure with respect to Inferential Reasoning.**



The given figure shows that it is not possible to minimize both type I and type II error.

Type I Error - is rejection of Null Hypothesis when it is true. In simpler words, Type I error occurs when we conclude that there is a statistical difference when there is actually no difference. This is also known as a false positive or producer's risk.

Type II Error - is failing to reject a Null Hypothesis when it is false or rejection of Alternate Hypothesis when it is true. In simpler words, Type II error occurs when we conclude that there is no difference when there is actually a statistical difference. This is also known as false negative or consumer's risk.

Basis for comparison	Type I error	Type II error
Definition	Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.	Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
Also termed	Type I error is equivalent to false positive.	Type II error is equivalent to a false negative.
Meaning	It is a false rejection of a true hypothesis.	It is the false acceptance of an incorrect hypothesis.
Symbol	Type I error is denoted by α.	Type II error is denoted by β.
Probability	The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus the power of the test.

Reduced	It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.
Cause	It is caused by luck or chance.	It is caused by a smaller sample size or a less powerful test.
What is it?	Type I error is similar to a false hit.	Type II error is similar to a miss.
Hypothesis	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.
When does it happen?	It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.

1.

a) Write down the Hunt's algorithm for constructing a decision tree. Using hunt's algorithm draw a decision tree with the following table.

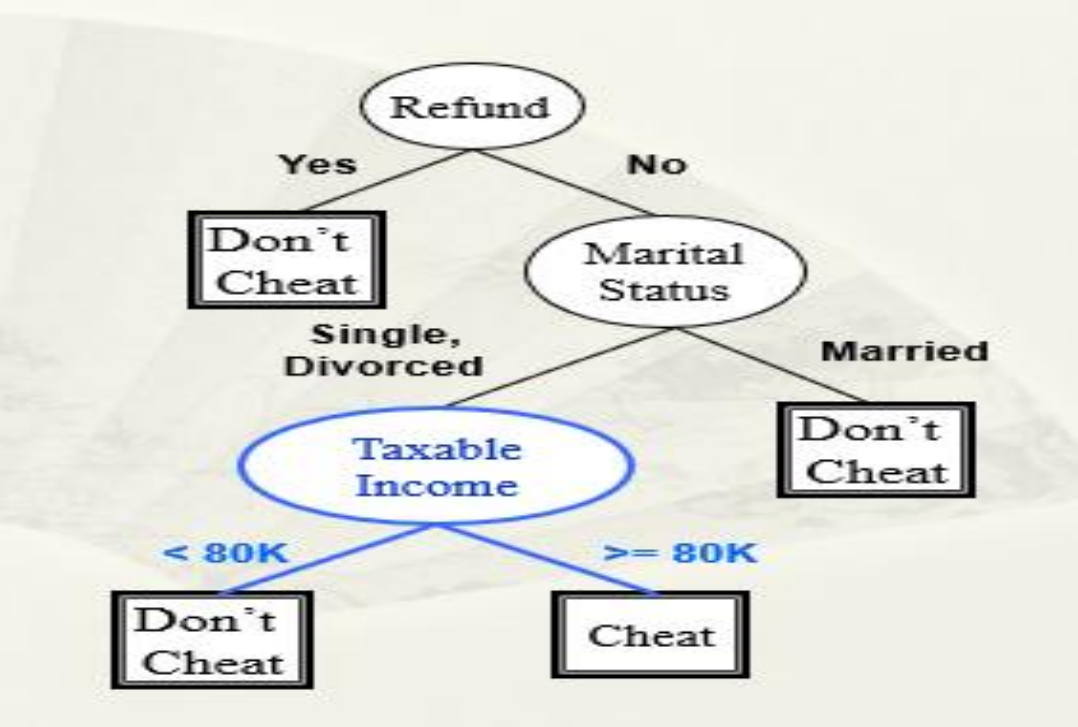
Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Hunts Algorithm:

- Let D_t be the set of training records that reach a node t

General Procedure:

- If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
- If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
- If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.



b) How can you determine that a piece of information is valuable or not? Support your opinion with the help of Entropy.

Ans:

A piece of information is valuable when it is Pure or Less valued Impurity.

We use entropy as a measure of impurity or disorder of data set D. (Or, a measure of information in a tree)

The Entropy formula is:

$$entropy(D) = -\sum_{j=1}^{|C|} \Pr(c_j) \log_2 \Pr(c_j)$$

$$\sum_{j=1}^{|C|} \Pr(c_j) = 1,$$

1. The data set D has 50% positive examples (Pr(*positive*) = 0.5) and 50% negative examples (Pr(*negative*) = 0.5).

$$entropy(D) = -0.5 \times \log_2 0.5 - 0.5 \times \log_2 0.5 = 1$$

2. The data set D has 20% positive examples (Pr(*positive*) = 0.2) and 80% negative examples (Pr(*negative*) = 0.8).

$$entropy(D) = -0.2 \times \log_2 0.2 - 0.8 \times \log_2 0.8 = 0.722$$

3. The data set D has 100% positive examples (Pr(*positive*) = 1) and no negative examples, (Pr(*negative*) = 0).

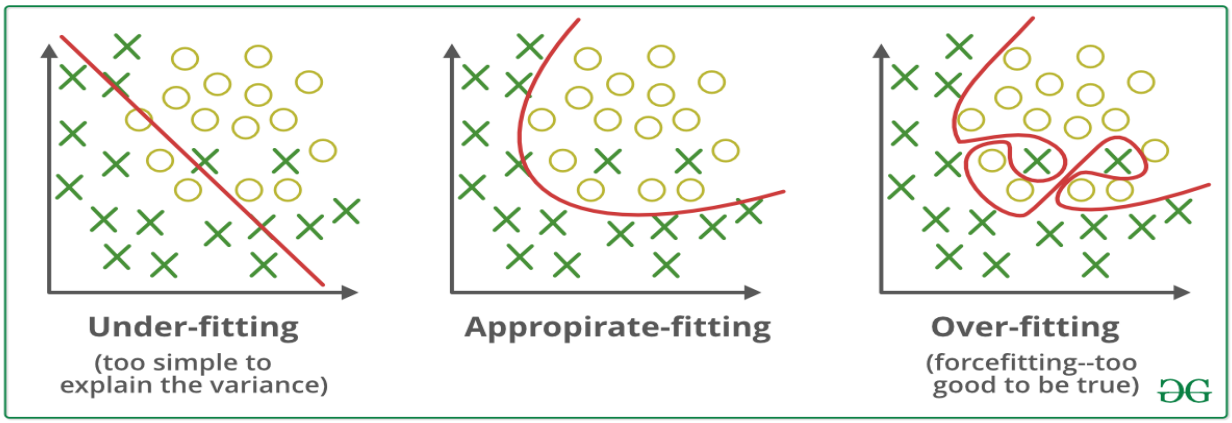
$$entropy(D) = -1 \times \log_2 1 - 0 \times \log_2 0 = 0$$

As the data become purer and purer, the entropy value becomes smaller and smaller. This is useful to us!

2.

a) Define the terms “overfitting” and “underfitting”. Discuss the various approaches of controlling an overfitting.

Overfitting	Underfitting
• Overfitting refers to a model that models the training data too well. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.	• Underfitting refers to a model that can neither model the training data nor generalize to new data. The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.



* Overfitting: A tree may overfit the training data
  * Good accuracy on training data but poor on test data
  * Symptoms: tree too deep and too many branches, some may reflect anomalies due to noise or outliers
* Two approaches to avoid overfitting
  * Pre-pruning: Halt tree construction early
    * Difficult to decide because we do not know what may happen subsequently if we keep growing the tree.
  * Post-pruning: Remove branches or sub-trees from a "fully grown" tree.
    * This method is commonly used. C4.5 uses a statistical method to estimates the errors at each node for pruning.
    * A validation set may be used for pruning as well.

b) What do you understand by "information gain"? Determine the sequence of attributes to construct a decision tree using following table based on information gain.

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Information gain

Entropy $D = \frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}$

$= 0.7 \log_2 0.7 + 0.3 \log 0.3$

$= 0.88$

yes = 3
NO = 7
total number = 10

refund

refund → evade

yes = 3 → yes → yes → 0
yes → NO → 3

~~NO = 7 → NO → yes → 3~~
~~→ yes → NO~~

refund → evade

NO = 7 → NO → yes → 3
NO → NO → 4

D_1 [yes] $= \frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}$

$= 0$

D_2 [NO] $= \frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}$

$= 0.429 \log_2 0.429 + 0.571 \log_2 0.571$

$= 0.985$

entropy $D_{(refund)} = \frac{3}{10} \times 0 + \frac{7}{10} \times 0.985$ (Yes, No)

$= 0 + 0.7 \times 0.985$

$= 0.6895$

Gain (D, refund) $= 0.88 - 0.6895$

$= 0.1905$

Marcital status:

	Yes	No
Singale → 4 →	2 →	2
married → 4 →	0 →	4
divorceed → 2 →	1 →	1

single = 4
married = 4
Divorced = 2

$D(single) \rightarrow \frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4}$

$= 0.5 \log_2 0.5 + 0.5 \log_2 0.5$

$= -1 = 1$

$D(married) = \frac{0}{4} \log_2 \frac{0}{4} + \frac{4}{4} \log_2 \frac{4}{4}$

$= 0$

$$D(\text{divorced}) = \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}$$

$$= -1 = 1$$

$$\text{Entropy}(\text{marital\_status}) = \frac{4}{10}\times 1 + \frac{4}{10}\times 0 + \frac{2}{10}\times 1$$

$$= 0.4 + 0 + 0.2$$

$$= 0.6$$

$$\text{Gain}(D.\text{marital}) = 0.88 - 0.6$$

$$= 0.28$$

~~TOP~~ Taxable

~~low Tax → 60k →~~

~~High tax → 220k →~~

taxable

	Yes	No
< 105	7	4
≥ 105	3	0

$$D[<105) \rightarrow \frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}$$

$$= 0.429 \log_2(0.429) + 0.571 \log_2(0.571)$$

$$= -0.985 = 0.985$$

$$D(\geq 105) = \frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}$$

$$= 0$$

$$\text{entropy } D_{(taxable)} = \frac{7}{10} \times 0.985 + \frac{3}{10} \times 0$$

$$= 0.6895 = 0.69$$

- Gain $\rightarrow$ $(D, taxable) = 0.88 - 0.69$

$$= 0.19$$

refund $= 0.1905$

marital $= 0.28$

taxable $= 0.19$

PMSCS-686: Introduction to Data Science

Assignment-2

Submitted By-

Mahmuda Rahman

CSE 20200127

Batch-22nd.

Question - 1: What is the purpose of regression analysis? Write down the properties of regression function.

Answer:

Purpose of regression Analysis:

Regression analysis serves three major purposes. Such as -

i) Description
ii) Control
iii) Prediction

The several purposes of regression analysis frequently overlap in practice.

Properties of Regression Function:

The properties are given in the following:

⇒ The line minimizes the sum of squared differences between observed values (the y values) and predicted values (the $\hat{y}$ values computed from the regression equation).

⇒ The regression line pass through the mean of X values ($\bar{x}$) and through the mean of Y values ($\bar{y}$).

⇒ The regression constant (b_0) is equal to the y interapt of the regression line.

⇒ The regression co-efficient (b_i) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line.

Question-2: The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table:

y	x
1250	41
1380	54
1425	63
1425	54
1450	48
1300	46
1400	62
1510	61
1575	64
1650	71

(a) What is the relationship between the advertising expenditure and weekly sales?

Solution:

y	x	xy	x^2
1250	41	51250	1681
1380	54	74520	2916
1425	63	89775	3969
1425	54	76950	2916
1450	48	69600	2304
1300	46	59800	2116
1400	62	86800	3844
1510	61	92110	3721
1575	64	100800	4096
1650	71	117150	5041
$\Sigma y =$	$\Sigma x = 564$	$\Sigma xy =$	$\Sigma xy =$
14365		818755	32604

Now, the least square estimates of the regression coefficients are,

$$b_1 = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{10(818755) - (564)(14365)}{10(32604) - (564)^2}$$

$$= 10.8$$

$n = 10$

Now,

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= (1436.50) - (10.8)(56.4)$$

$$= 828$$

∴ The estimated regression function is, $\hat{y} = 828 + 10.8x$

Sales = 828 + 10.8 Expenditure;

which means that if the weekly advertising expenditure is increased by $1 we would expect sales to increase by $10.8.

(b) If the advertising expenditure is $55, what will be the estimated weekly sales?

Solution:

From the solution of (a) we ~~got~~ the regression function that is.

Sales = 828 + 10.8 Expenditure.

For,
Expenditure, $55, Sales = 828 + 10.8 (55)

= 828 + 594

= 1422

Question-3: What do you understand by regression inference? State the conditions for a successful regression inference.

Answer:

Regression Inference:

Regression inference is the process which estimates the relationship between a dependent variable and one or more independent variables of an underlying population based on a sample or subset of the data.

Conditions of the successful regression inference:

Following conditions need to be met for successfull regression inference:

⇒ The sample is a simple random sample from the population.

⇒ Linearity of relationship between variables.

⇒ Independence of the residuals

⇒ Normality of the residuals

⇒ Equality of variance of the residuals.

Question-4. What is residual? Why residual plot analysis is important in statistical inference? Justify your answer by some examples.

Answer:

Residual:

A residual is the vertical distance between a data point and the regression line. Each data point has one residual They are positive if they are above the regression line and negative if they are below the regression line. If the regression line is actually passes through the point, the residual at the point is zero. Residuals are also called as "errors".

Importance of residual plot analysis in statistical inference:

Residual plot analysis is important in statistical inference because by checking the residual plots we can examine the conditions of the successful inference. Residual plots are:

→ Plot a histogram of the residuals: provides a check on the normality assumptions

→ plot of residuals against fitted values on independent variable : used to check the assumption of constant variance and the aptness of the model.

→ Plot of residuals against time: provides a check on the independence of the error terms assumption.

Consider the following examples:

example-1:

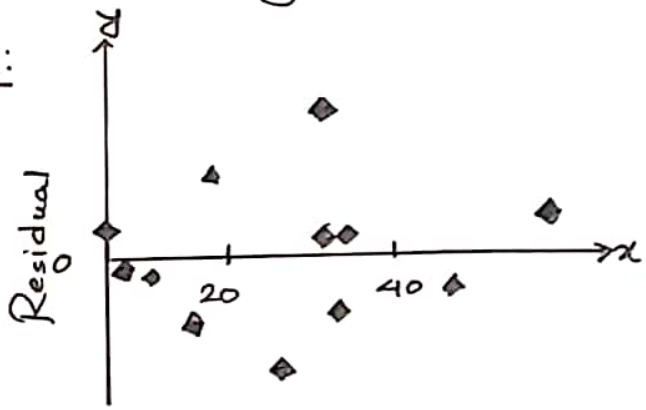


fig: 1

The residual plot of fig: 1 shows a scatter of the points with no individual observations or systematic changes as x increases.

Example-2:

The points of residual plot of fig-2 have a curve pattern, so a straight line fits poorly.

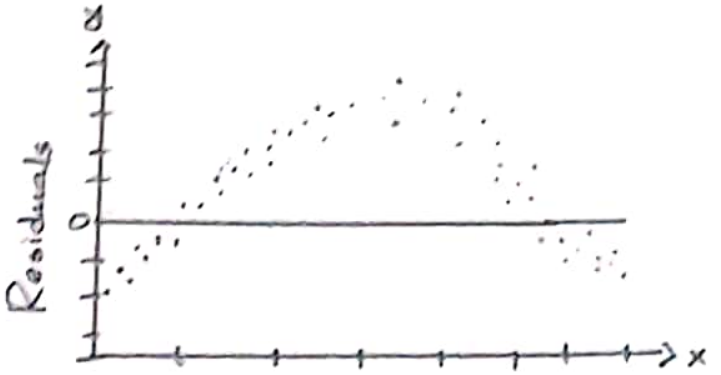
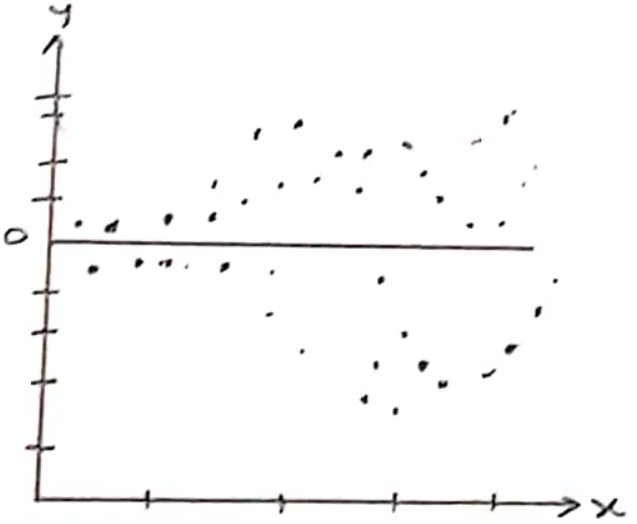


Fig: 2.

Example-2:

The points in the plot of fig-3 show more spreads for larger values of the explanatory variable x.



So, the prediction will be less accurate when x is large.

Question – 5: The MD of the company is interested in testing whether or not there is a linear association between advertising expenditure and weekly sales, using regression model with data Table-1 ($\alpha = 0.05$). Using standard t-test, find out the 95% Confidence interval.

Solution:

y	x	xy	x^2
1250	41	51250	1681
1380	54	74520	2916
1425	63	89775	3969
1425	54	76950	2916
1450	48	69600	2304
1300	46	59800	2116
1400	62	86800	3844
1510	61	92110	3721
1575	64	100800	4096
1650	71	117150	5041
$\Sigma y =$ 14365	$\Sigma x =$ 564	$\Sigma xy =$ 818755	$\Sigma x^2 =$ 32604

Here, $n = 10$

Now, the least square estimates of the regression coefficients are,

$$b_1 = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$= \frac{10(818755) - (564)(14365)}{10(32604) - (564)^2}$$

$$= 10.8$$

Now,

$$b_0 = \bar{y} - b_1\bar{x}$$
$$= (1436.5) - 10.8\,(564)$$
$$= 828.$$

The estimated regression function is

$$\hat{y} = 828 + 10.8x$$

Now,

y	x	$\hat{y}$	residuals (e)	Squares (e)
1250	41	1270.8	−20.8	432.64
1380	54	1411.2	−31.2	973.44
1425	63	1508.4	−83.4	6955.56
1425	54	1411.2	13.8	190.44
1450	48	1346.4	103.6	10732.96
1300	46	1324.8	−24.8	615.04
1400	62	1497.6	−97.6	9525.76
1510	61	1486.8	23.2	538.24
1575	64	1519.2	55.8	3113.64
1650	71	1594.8	55.2	3047.04

Total = 36124.76

We know,

$$\text{standard error, } s = \sqrt{\frac{1}{n-2}\sum e_i^2}$$
$$= \sqrt{\frac{1}{8} \times 36124.76}$$
$$= 67.2.$$

Now, Let,

Hypothesis: $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Decision Rule

Reject H_0 if $t > t_{0.025;8} \Rightarrow t > 2.306$

or, $t < -t_{0.025;8} \Rightarrow t < -2.306$

Test statistics:

$$t = \frac{b_1}{s(b_1)}$$

Now, $s(b_1) = \frac{s}{\sqrt{\Sigma (x-\bar{x})^2}} = \frac{6.72}{\sqrt{7044}} = 2.38$

and, $b_1 = 10.8$

$$\therefore t = \frac{10.8}{2.38} = 4.5.$$

Since $t = 4.5 > 2.306$, then we reject H_0.

There is a linear association between advertising expenditure and weekly sales.

Now, the 95% confidence interval is:

$$b_1 \pm t_{(\frac{\alpha}{2};n-2)} (s(b_1))$$

$$= 10.8 \pm 2.306 (2.38)$$

$$= 10.8 \pm 5.49$$

So the interval is $(5.31, 16.3)$.

Question No:-6: Question is too long. Please see the assignment-2.

Solution: let, y = wages, and x = LOS

Wages (y)	LOS (x)	xy	x2	Wages (y)	LOS (x)	xy	x2
389	94	36566	8836	486	60	29160	3600
395	48	18960	2304	393	7	2751	49
329	102	33558	10404	311	22	6842	484
295	20	5900	400	316	57	18012	3249
377	60	22620	3600	384	78	29952	6084
479	78	37362	6084	360	36	12960	1296
315	45	14175	2025	369	83	30627	6889
316	39	12325	1521	529	66	34914	4356
324	20	6480	400	270	47	12690	2209
307	65	19955	4225	332	97	32204	9409
403	76	30628	5776	547	228	124716	51984
378	48	18144	2304	347	27	9369	729
348	61	21,228	3721	328	48	15744	2304
488	30	14,640	900	327	7	2289	49
391	108	42228	11664	320	74	23680	5476
541	61	33001	3721	404	204	82416	41616
312	10	3120	100	443	24	10632	576
418	68	28424	4624	261	13	3393	169
417	54	22518	2916	417	30	12510	900
516	24	12384	576	450	95	42750	9025
443	222	98346	49284	443	104	46072	10816
353	58	20474	3364	566	34	19244	1156
349	41	14309	1722	461	184	84824	33856
499	153	76347	23409	436	156	68016	24336
322	16	5152	256	321	25	8025	625
408	43	17544	1849	221	43	9503	1849
393	96	37728	9216	547	36	19692	1296
277	98	27146	9604	362	60	21720	3600
649	150	97350	22500	415	102	42330	10404
272	124	33728	15376				

Now,

$n = 59$, $\Sigma x = 4159$, $\Sigma x^2 = 451031$

$\Sigma xy = 171.9376$, $\Sigma y = 23069$

Now, the least square estimate of the regression coefficients are,

$$b_1 = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$= \frac{59(171.9376) - (4159)(23069)}{59(451031) - (4159)^2}$$

$$= 0.59$$

So, $b_0 = \bar{y} - b_1 \bar{x}$

$= 391 - (0.59)(70.5)$

$= 349$

∴ The estimated regression function is

$$\hat{y} = 349 + 0.59x$$

$\hat{y}$	Residuals (e)	Squares (e^2)	$\hat{y}$	Residual (e)	Square (e^2)
4046	−15.46	239.0116	384.4	10.56	10322.56
377.32	17.68	312.5824	353.13	37.87	1589.6169
409.18	−80.18	6428.832	361.88	−50.98	2598.9604
360.8	−65.8	4329.64	382.63	−66.63	4439.5569
384.4	−7.4	54.76	395.02	−34.02	121.4404
395.02	83.98	7052.64	370.24	−18.24	104.8576
375.55	−60.55	3666.30	397.57	−89.57	839.2609
372.01	−56.01	3137.12	387.99	144.06	1087.9236
360.8	−36.8	1354.24	376.73	−106.83	11391.2929
387.35	−80.35	6456.12	406.23	−84.05	5510.0020
393.84	9.16	83.9056	463.52	69.48	4029.7104
377.32	0.68	0.4624	364.93	−17.93	321.4849
384.99	−36.99	1368.2601	377.32	−49.32	2432.4624
3667	121.3	14713.69	353.13	−26.13	682.7769
412.72	−21.72	471.7584	392.66	−72.66	5279.4756
384.99	156.01	24339.12	469.36	−65.36	4271.9296
354.9	−42.9	1840.41	368.16	79.84	6374.7225
3809.12	28.88	834.0544	356.67	−95.67	9152.7489
380.86	36.14	1306.0996	366.8	50.3	2530.09
363.16	152.84	23360.06	405.05	44.95	2020.5025
479.98	−36.98	1367.52	410.36	32.64	1065.3696
383.22	−30.22	913.24	369.06	196.94	38785.3636
373.19	−24.19	585.156	457.56	3.44	11.8336
439.27	59.73	3567.672	441.09	−5.09	25.9081
358.44	−36.44	1327.873	363.75	−42.75	1827.5625
374.37	33.63	1130.857	374.37	−153.37	23552.3569
405.64	−12.64	159.769	370.24	176.76	31244.0576
406.82	−129.82	16853.23	384.4	−22.4	501.76
497.5	211.5	44732.25	409.18	5.82	33.8724
422.16	−150.16	22548.02			

Total = 385468.5911

We know,

$$\text{standard Error. } s = \sqrt{\frac{1}{n-2}(\Sigma e^2)}$$

$$= \sqrt{\frac{1}{59-2} \cdot (385463.59)}$$

$$= 82.23.$$

Now, let us assume,

Hypothesis :

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Decision Rule,

Reject H_0 if $t > t_{0.025;57} \Rightarrow t > 2.016$

or, $t < -t_{0.025;57} \Rightarrow t > -2.016$

Now, t-statistics :

$$t = \frac{b_1}{s(b_1)}$$

$$s(b_1) = \frac{s}{\sqrt{\Sigma(x-\bar{x})^2}}$$

$$= \frac{82.23}{397.3}$$

$$= 0.21$$

and, $b_1 = 0.59$

$$\therefore t = \frac{b_1}{s(b_1)} = \frac{0.59}{0.21} = 2.81$$

Since, $t = 2.81 > 2.016$, then we reject H_0.

There is a linear association between wages and length of service.

Now the 95% Confidence interval is

$$b_1 \pm t_{(0.025;57)} (S(b_1))$$

$$= 0.59 \pm 2.016 (0.21)$$

$$= 0.59 \pm 0.42$$

So the interval is $(0.17, 1.01)$.

Question-7: Why does linear regression model fail to infer about an event in the case of dichotomous or binary output?

Answer:

The linear regression model fail to infer about an event in the case of dichotomous or binary output because, the linear regression model assums that the dependent variable (usually 'y') is normally distributed. But in the case of dichotomous or binary outcome, we have the binary dependent variable which heavily violates the assumptions of the linear regression model. Thus, it doesn't make sense to use linear regression when the dependent variable is binary.

Question-8: Define the term probability, odds and odds ratio. Is there any relationship among them?

Answer:

Probability:

The probability of an event is a measure of the likelihood that the event will occur.

Odds:

The odds are defined as the probability that the event will occur divided by the probability that the event will not occur.

Odds Ratio:

Ratio of the two odds estimates.

Relationship between probability and odds:

Probability and odds are closely related. Odds can be defined as a probability ratio. as follows

$$\text{Odds (event)} = \frac{\text{Probability (event)}}{1 - \text{probability (event)}}$$

Similarly, probability can be expressed by following:

$$\text{probability (event)} = \frac{\text{odds (event)}}{1 + \text{odds (event)}}$$

Question-9: Assuming all independent events of, derive the following expression likelihood function where the symbols bear the usual meaning.

Answer:

Assuming all independent events of y, the expression of likelihood function is derived in the following where the symbols bear the usual meaning.

define, $p = P_r(y=1)$

Then for dichotomous outcome

$$P_r(y=0) = 1 - P_r(y=1)$$
$$= 1-p.$$

Then $$P_r(y) = p^y (1-p)^{1-y}$$

For,
$y=1$, $P_r(1) = p^1 (1-p)^0 = p$

Again,
for,
$y=0$, $P_r(0) = p^0 (1-p)^{1-0} = 1-p.$

So, given that;
$$P_r(y) = p^y (1-p)^{1-y}$$

$$L = \prod_{i=1}^{N} Pr(y_i) = \prod_{i=1}^{N} P_i^{y_i} (1-P_i)^{1-y_i}$$

$$= \prod_{i=1}^{N} P_i^{y_i} \left(\frac{1}{1-P_i}\right)^{y_i} (1-P_i)$$

$$= \prod_{i=1}^{N} \left(\frac{P_i}{1-P_i}\right)^{y_i} (1-P_i)$$

Next, taking the logarithm of both sides,

$$\ln L = \sum_i y_i \ln\left(\frac{P_i}{1-P_i}\right) + \sum \ln(1-P_i)$$

Remember,

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K$$

$$= \beta x_i$$

Substituting in using logistic regression model

$$\ln L = \sum y_i \beta x_i - \sum \ln(1 + \exp(\beta x_i))$$

Question-10: Please see the Question in assignment -2.

Solution:

(a) **Effect of age on Pregnancy:**

$$\ln\left(\frac{Pr(\text{Pregnancy})}{1-Pr(\text{Pregnancy})}\right) = 2.67 - 0.13 * \text{age}$$

The $\hat{OR}$ age $= \exp(-0.13)$

$= 0.88$

this implies that for every 1 yr. increases in age, the odds of the pregnancy decreases by 12%.

(b) **For 27 years old having Success:**

$$\hat{Pr}(\text{Pregnancy}) = \frac{\exp(2.67 - 0.13 \times 27)}{1 + \exp(2.67 - 0.13 \times 27)}$$

$= 0.30$

$= 30\%$.

this implies from this model, a 27 yr. old has about a 30% chance of pregnancy success.

Lecture - 3

→ Data aggregation with example

Data Inhancemunt এর জন্য data aggregation importand

→ Data munging → ~~Feature~~ feature reduction → why why

→ Bias - variance → কোনটা বেশি slide → 43 4tr figure এর মধ্যে

→ slide - 50 → model persistance, model transience

বলতে কি বোঝায়

Point estimate কি? what makes point estimate good.

Standerd error of the mean বলতে কি বোঝায়

Centrical limit theorem এর কি standerd error & central limit

Confidence intervals & math
দেখি slide – ৮৩, ৮৪

Hypothesis দেখি Hypothesis এর প্রকার

Hypothesis এর property দেখি

word Hypothesis convert to statistical Hypotheses -এ ঠিক করতে হবে কোন প্রকার ব্যবহার।

α error / Type 1 error
β error / Type 2 error } দেখি

Simple linear ~~Regress~~ Regression

Regression function →

general regression function.

variable transformations → কি হবে দেখ

Regression line থেকে math আসবে

Point estimation → math →

Residual → orginal and prediction এর মধ্যে পার্থক্য কে residual বলে error বলে।

slide → ২৮

গুরুত্বপূর্ণ

what is the advantage of residual analysis → figure সহ describe করতে হবে।

→ slide – 43 → 1st formula

" → 4৮ → math → ৭ক

" → 46 → " →

" → ৮0

12 → ৮0 পর্যন্ত ছবি

math →

Residual

Logistic regression.

machine learning কি?

→ classification & Association

এর কাজ আছে।

Supervised learning and unsupervised learning কী পার্থক্য আছে?

nearest neighbor কী?

↳ 1, 2, 3 নং মডেল কী এর error কমে না কমে

distance base classification কাজ করে

Scaling কী প্রয়োজন

lazy learner की दो उदाहरण दो ।

Nearest neighbor search

Support vector machine.

what is the rule of this margin

naive Bayes → इसका problem →

Decision tree

Hunts algorithm →

Hunts algorithm apply करके tree

is the decision tree unique?

Information gain → entropy →

↓

decision tree.

Lecture 01:

MCQ: Source of data, Amount of data, Why Data Science, Types of data

Written: Data Dimensions, Define Data Science, Roles of Data Scientist

Lecture 02:

MCQ: Data Science Process with sub steps, Why information leakage, Model Persistence vs Model Transience

Written: Why Project fail + How to avoid, Define Data Aggregation + Methods, Data Munging + Why feature reduction, Bias vs Variance with figure

Lecture 03:

MCQ: Type 1 error vs Type 2 error

Written: Point estimate and Sampling distribution, Define Confidence Interval + math, Hypothesis Properties + Word into Statistical

Lecture 04:

MCQ: Condition of Statistical relation, Use of Regression analysis, Regression model, Variable Transformation,

Written: Residual + Why important + Advantages, Math (12-50 slide),

Lecture 06:

MCQ: Classification, Supervised vs Unsupervised, Distance base Classification, Euclidean measure problem, Lazy learners, Complexity of binary tree

Written: Nearest Neighbor (1/2/3 nearest), Define Margin + Error find out and how to avoid, Naïve Byes Math

Lecture 07:

MCQ: Simpler Tree, Entropy, Information Gain

Written: Follow Assignments

1. What is Data Science?

Data Science is a combination of algorithms, tools, and machine learning technique which helps you to find common hidden patterns from the given raw data.

2. What is logistic regression in Data Science?

Logistic Regression is also called as the logit model. It is a method to forecast the binary outcome from a linear combination of predictor variables.

3. Name three types of biases that can occur during sampling

In the sampling process, there are three types of biases, which are:

- Selection bias
- Under coverage bias
- Survivorship bias

4. Discuss Decision Tree algorithm

A decision tree is a popular supervised machine learning algorithm. It is mainly used for Regression and Classification. It allows breaks down a dataset into smaller subsets. The decision tree can able to handle both categorical and numerical data.

5. What is Prior probability and likelihood?

Prior probability is the proportion of the dependent variable in the data set while the likelihood is the probability of classifying a given observant in the presence of some other variable.

6. Explain Recommender Systems?

It is a subclass of information filtering techniques. It helps you to predict the preferences or ratings which users likely to give to a product.

7. Name three disadvantages of using a linear model

Three disadvantages of the linear model are:

- The assumption of linearity of the errors.
- You can't use this model for binary or count outcomes
- There are plenty of overfitting problems that it can't solve

8. Why do you need to perform resampling?

Resampling is done in below-given cases:

- Estimating the accuracy of sample statistics by drawing randomly with replacement from a set of the data point or using as subsets of accessible data
- Substituting labels on data points when performing necessary tests
- Validating models by using random subsets

9. List out the libraries in Python used for Data Analysis and Scientific Computations.

- SciPy
- Pandas
- Matplotlib
- NumPy
- SciKit
- Seaborn

10. What is Power Analysis?

The power analysis is an integral part of the experimental design. It helps you to determine the sample size requires to find out the effect of a given size from a cause with a specific level of assurance. It also allows you to deploy a particular probability in a sample size constraint.

11. Explain Collaborative filtering

Collaborative filtering used to search for correct patterns by collaborating viewpoints, multiple data sources, and various agents.

12. What is bias?

Bias is an error introduced in your model because of the oversimplification of a machine learning algorithm." It can lead to underfitting.

13. Discuss 'Naive' in a Naive Bayes algorithm?

The Naive Bayes Algorithm model is based on the Bayes Theorem. It describes the probability of an event. It is based on prior knowledge of conditions which might be related to that specific event.

14. What is a Linear Regression?

Linear regression is a statistical programming method where the score of a variable 'A' is predicted from the score of a second variable 'B'. B is referred to as the predictor variable and A as the criterion variable.

15. State the difference between the expected value and mean value

They are not many differences, but both of these terms are used in different contexts. Mean value is generally referred to when you are discussing a probability distribution whereas expected value is referred to in the context of a random variable.

16. What the aim of conducting A/B Testing?

AB testing used to conduct random experiments with two variables, A and B. The goal of this testing method is to find out changes to a web page to maximize or increase the outcome of a strategy.

17. What is Ensemble Learning?

The ensemble is a method of combining a diverse set of learners together to improvise on the stability and predictive power of the model. Two types of Ensemble learning methods are:

Bagging

Bagging method helps you to implement similar learners on small sample populations. It helps you to make nearer predictions.

Boosting

Boosting is an iterative method which allows you to adjust the weight of an observation depends upon the last classification. Boosting decreases the bias error and helps you to build strong predictive models.

18. Explain Eigenvalue and Eigenvector

Eigenvectors are for understanding linear transformations. Data scientist need to calculate the eigenvectors for a covariance matrix or correlation. Eigenvalues are the directions along using specific linear transformation acts by compressing, flipping, or stretching.

19. Define the term cross-validation

Cross-validation is a validation technique for evaluating how the outcomes of statistical analysis will generalize for an Independent dataset. This method is used in backgrounds where the objective is forecast, and one needs to estimate how accurately a model will accomplish.

20. Explain the steps for a Data analytics project

The following are important steps involved in an analytics project:

- Understand the Business problem
- Explore the data and study it carefully.
- Prepare the data for modeling by finding missing values and transforming variables.
- Start running the model and analyze the Big data result.
- Validate the model with new data set.
- Implement the model and track the result to analyze the performance of the model for a specific period.

21. Discuss Artificial Neural Networks

Artificial Neural networks (ANN) are a special set of algorithms that have revolutionized machine learning. It helps you to adapt according to changing input. So the network generates the best possible result without redesigning the output criteria.

22. What is Back Propagation?

Back-propagation is the essence of neural net training. It is the method of tuning the weights of a neural net depend upon the error rate obtained in the previous epoch. Proper tuning of the helps you to reduce error rates and to make the model reliable by increasing its generalization.

23. What is a Random Forest?

Random forest is a machine learning method which helps you to perform all types of regression and classification tasks. It is also used for treating missing values and outlier values.

24. What is the importance of having a selection bias?

Selection Bias occurs when there is no specific randomization achieved while picking individuals or groups or data to be analyzed. It suggests that the given sample does not exactly represent the population which was intended to be analyzed.

25. What is the K-means clustering method?

K-means clustering is an important unsupervised learning method. It is the technique of classifying data using a certain set of clusters which is called K clusters. It is deployed for grouping to find out the similarity in the data.

26. Explain the difference between Data Science and Data Analytics

Data Scientists need to slice data to extract valuable insights that a data analyst can apply to real-world business scenarios. The main difference between the two is that the data scientists have more technical knowledge then business analyst. Moreover, they don't need an understanding of the business required for data visualization.

27. Explain p-value?

When you conduct a hypothesis test in statistics, a p-value allows you to determine the strength of your results. It is a numerical number between 0 and 1. Based on the value it will help you to denote the strength of the specific result.

28. Define the term deep learning

Deep Learning is a subtype of machine learning. It is concerned with algorithms inspired by the structure called artificial neural networks (ANN).

29. Explain the method to collect and analyze data to use social media to predict the weather condition.

You can collect social media data using Facebook, twitter, Instagram's API's. For example, for the tweeter, we can construct a feature from each tweet like tweeted date, retweets, list of follower, etc. Then you can use a multivariate time series model to predict the weather condition.

30. When do you need to update the algorithm in Data science?

You need to update an algorithm in the following situation:

- You want your data model to evolve as data streams using infrastructure
- The underlying data source is changing

  If it is non-stationarity

31. What is Normal Distribution

A normal distribution is a set of a continuous variable spread across a normal curve or in the shape of a bell curve. You can consider it as a continuous probability distribution which is useful in statistics. It is useful to analyze the variables and their relationships when we are using the normal distribution curve.

32. Which language is best for text analytics? R or Python?

Python will more suitable for text analytics as it consists of a rich library known as pandas. It allows you to use high-level data analysis tools and data structures, while R doesn't offer this feature.

33. Explain the benefits of using statistics by Data Scientists

Statistics help Data scientist to get a better idea of customer's expectation. Using the statistic method Data Scientists can get knowledge regarding consumer interest, behavior, engagement, retention, etc. It also helps you to build powerful data models to validate certain inferences and predictions.

34. Name various types of Deep Learning Frameworks

- Pytorch
- Microsoft Cognitive Toolkit
- TensorFlow
- Caffe
- Chainer
- Keras

35.Explain Auto-Encoder

Autoencoders are learning networks. It helps you to transform inputs into outputs with fewer numbers of errors. This means that you will get output to be as close to input as possible.

36. Define Boltzmann Machine

Boltzmann machines is a simple learning algorithm. It helps you to discover those features that represent complex regularities in the training data. This algorithm allows you to optimize the weights and the quantity for the given problem.

37. Explain why Data Cleansing is essential and which method you use to maintain clean data

Dirty data often leads to the incorrect inside, which can damage the prospect of any organization. For example, if you want to run a targeted marketing campaign. However, our data incorrectly tell you that a specific product will be in-demand with your target audience; the campaign will fail.

38. What is skewed Distribution & uniform distribution?

Skewed distribution occurs when if data is distributed on any one side of the plot whereas uniform distribution is identified when the data is spread is equal in the range.

39. When underfitting occurs in a static model?

Underfitting occurs when a statistical model or machine learning algorithm not able to capture the underlying trend of the data.

40. What is reinforcement learning?

Reinforcement Learning is a learning mechanism about how to map situations to actions. The end result should help you to increase the binary reward signal. In this method, a learner is not told which action to take but instead must discover which action offers a maximum reward. As this method based on the reward/penalty mechanism.

41. Name commonly used algorithms.

Four most commonly used algorithm by Data scientist are:

- Linear regression
- Logistic regression
- Random Forest
- KNN

42. What is precision?

Precision is the most commonly used error metric is n classification mechanism. Its range is from 0 to 1, where 1 represents 100%

43. What is a univariate analysis?

An analysis which is applied to none attribute at a time is known as univariate analysis. Boxplot is widely used, univariate model.

44. How do you overcome challenges to your findings?

In order, to overcome challenges of my finding one need to encourage discussion, Demonstrate leadership and respecting different options.

45. Explain cluster sampling technique in Data science

A cluster sampling method is used when it is challenging to study the target population spread across, and simple random sampling can't be applied.

46. State the difference between a Validation Set and a Test Set

A Validation set mostly considered as a part of the training set as it is used for parameter selection which helps you to avoid overfitting of the model being built.

While a Test Set is used for testing or evaluating the performance of a trained machine learning model.

47. Explain the term Binomial Probability Formula?

"The binomial distribution contains the probabilities of every possible success on N trials for independent events that have a probability of π of occurring."

48. What is a recall?

A recall is a ratio of the true positive rate against the actual positive rate. It ranges from 0 to 1.

49. Discuss normal distribution

Normal distribution equally distributed as such the mean, median and mode are equal.

50. While working on a data set, how can you select important variables? Explain

Following methods of variable selection you can use:

- Remove the correlated variables before selecting important variables
- Use linear regression and select variables which depend on that p values.
- Use Backward, Forward Selection, and Stepwise Selection
- Use Xgboost, Random Forest, and plot variable importance chart.
- Measure information gain for the given set of features and select top n features accordingly.

51. Is it possible to capture the correlation between continuous and categorical variable?

Yes, we can use analysis of covariance technique to capture the association between continuous and categorical variables.