**Statistics is the study of how to collect, organize, analyze and interpret numerical information and data.**

**\*Why Study Statistics?**

**Statistics are part of your daily life and are all around you. Statistics don’t lie but Statisticians will in any of the following situations:**

**• Data Gathering**

**• Data Understanding**

**• Data Analysis/Interpretation**

**• Data Presentation**

* **Complete Total data with respect to the problem statement is called as population.**
* **Sample is representation of population. Sample should hold the all properties of population. Sample is subset of population.**

**Census:**

**Gathering Data from whole population of interest is called Census.**

**Survey:**

**Gathering data from sample in order to make conclusions about population is called Survey.**

**The common data assumptions are: random samples, independence, normality, equal variance, stability, and that your measurement system is accurate and precise.**

**Assumptions of data:**

* **Data have a normal distribution (Normality)**
* **A sample is random when each data point in your population has an equal chance of being included in the sample (Random samples).**
* **The value of one observation does not influence or affect the value of other observations (Statistical Independence).**
* **Different samples can come from populations with different means, they have the same variance (Equal Variance).**
* **A stable process is one in which the inputs and conditions are consistent over time. This means the sources of variation are consistent over time, and the process does not exhibit unpredictable variation. In contrast, if a process is unstable and changing over time, the sources of variation are inconsistent and unpredictable. As a result of the instability, you cannot be confident in your statistical test results. (Stability)**

**Why Machine Learning?**

**There are problems whose solutions are not formulated by using standard rule based algorithms.**

**\*Parameter: A descriptive measure of the population.**

**For example, population mean, population variance, population standard deviation, etc.**

**Greek – Population Parameter**

**Mean – μ**

**Variance – σ2**

**Standard Deviation - σ**

**\*Statistic: A descriptive measure of the sample.**

**For example, sample mean, sample variance, sample standard deviation, etc.**

**Roman – Sample Statistic**

**Mean – x**

**Variance – s2**

**Standard Deviation – s**

**\*Descriptive Statistics – Data gathered about a group to reach conclusions about the same group.**

**\*Inferential Statistics – Data gathered from a sample and the statistics generated to reach conclusions**

**about the population from which the sample is taken. Also known as Inductive Statistics.**

**Ordinal Variables:**

**While there is an order, difference between consecutive levels are not always equal.**

**Measures of central tendency:**

**\*Represent entire sample set by single value.**

**1.Mean (Affected by outliers)**

**2.Mode (Most probable value/The most frequently occurring data point)**

**3.Median (Robust with respect to outliers in the data)**

**\*Mean and Median need not be in the dataset but Mode has to be in it.**

**\*Mode is also the only average that works with categorical data.**

**\*Range = Max – Min**

**\*Lower quartile (25th percentile, Q1) = (𝑛+1) /4 th**

**\*Middle quartile = Median = 2∗(𝑛+1)/ 4 th**

**\*Upper quartile (75th percentile, Q3) = 3∗(𝑛+1)/ 4 th**

**\*Interquartile range, IQR = Q3-Q1 (central 50% of data)**

**\*Range and IQR give the spread but still do not describe variability (consistency).**

**\*Average Squared Distance from the mean is nothing but a variability.**

**Variance = Σ(𝑥−𝜇)^2 /𝑛**

**Units are squared, which is not intuitive.**

**Standard Deviation, 𝜎 = (𝑉𝑎𝑟𝑖𝑎𝑛𝑐𝑒)^1/2**

**\*Standard score, Z will describe the no of stdevs from the mean.**

**Z = x-u/ 𝜎**

**\*Probability is basis for inferential statistics, probability is predicting the likelihood of a feature event.**

**\*Statistics analyze the past events.**

**\*Probability measure is a function that assigns real value to every outcome of experiment.**

**1.Classical Method – A priori or Theoretical Probability can be determined prior to**

**conducting any experiment.**

**𝑃 (𝐸) =# 𝑜𝑓 𝑜𝑢𝑡𝑐𝑜𝑚𝑒𝑠 𝑖𝑛 𝑤ℎ𝑖𝑐ℎ 𝑡ℎ𝑒 𝑒𝑣𝑒𝑛𝑡 𝑜𝑐𝑐𝑢𝑟𝑠 /𝑡𝑜𝑡𝑎𝑙 𝑝𝑜𝑠𝑠𝑖𝑏𝑙𝑒 # 𝑜𝑓 𝑜𝑢𝑡𝑐𝑜𝑚𝑒𝑠**

**Example: Tossing of a fair die**

**2.Empirical Method – A posteriori or Frequentist Probability can be determined post**

**conducting a thought experiment.**

**𝑃 (𝐸) = # 𝑜𝑓 𝑡𝑖𝑚𝑒𝑠 𝑎𝑛 𝑒𝑣𝑒𝑛𝑡 𝑜𝑐𝑐𝑢𝑟𝑟𝑒𝑑 / 𝑡𝑜𝑡𝑎𝑙 # 𝑜𝑓 𝑜𝑝𝑝𝑜𝑟𝑡𝑢𝑛𝑖𝑡𝑖𝑒𝑠 𝑓𝑜𝑟 𝑡ℎ𝑒 𝑒𝑣𝑒𝑛𝑡 𝑡𝑜**

**ℎ𝑎𝑣𝑒 𝑜𝑐𝑐𝑢𝑟𝑟𝑒𝑑**

**\*Set of all possible outcomes is called as sample space, subset of sample space is called as an event.**

**\*Two events are independent if occurrence of one has no influence on occurrence of the other.**

**\*Two events are mutually exclusive if occurrence of one implies other event does not occur.**

**\*If two events A and B are not independent, then the information available about event A can**

**influence the predictability of B is called as conditional probability.**

**\*Joint Probability: Probability describing a combination of attributes.**

**\*Marginal Probability: Probability describing a single attribute**

**𝑃𝑜𝑠𝑡𝑒𝑟𝑖𝑜𝑟 𝑃𝑟𝑜𝑏𝑎𝑏𝑖𝑙𝑖𝑡𝑦 =𝑃𝑟𝑖𝑜𝑟 𝑃𝑟𝑜𝑏𝑎𝑏𝑖𝑙𝑖𝑡𝑦 ∗ 𝐿𝑖𝑘𝑒𝑙𝑖ℎ𝑜𝑜𝑑/ 𝐸𝑣𝑖𝑑𝑒𝑛𝑐𝑒**

**\*** **Bayes’ Theorem allows you to find reverse probabilities, and to allow revision of original probabilities with new information.**

**\*Random variable is a function map from sample space to real line such that there is unique real number is corresponding to every outcome of sample space.**

\*A **variable that can take multiple values with different probabilities.**

**\*The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.**

**\*For Discrete RV the probability mass function assign probability to every outcome in sample space;**

**\*For Continuous RV the probability density function assign probability to every interval on a line.**

**\*A series of contiguous rectangles that represent the frequency of data in given class intervals is called as histogram.**

**How many class intervals?**

**Rule of thumb: 5-15 (not too many and not too few)**

**Freedman-Diaconis rule:**

**𝑁𝑜.𝑜𝑓 𝑏𝑖𝑛𝑠 = (𝑚𝑎𝑥 − 𝑚𝑖𝑛)/(2 ∗ 𝐼𝑄𝑅 ∗ 𝑛^−1/ 3)**

**,**

**𝑤ℎ𝑒𝑟𝑒 𝑡ℎ𝑒 𝑑𝑒𝑛𝑜𝑚𝑖𝑛𝑎𝑡𝑜𝑟 𝑖𝑠 𝑡ℎ𝑒 𝑏𝑖𝑛 − 𝑤𝑖𝑑𝑡ℎ**

**\*Variance does not change when a constant is added.**

**Q) What is the need of Sampling?**

**A) From finite sample we are deriving conclusions pdf and its parameters of population.**

**\*Variance will describes the spread of the data**

**\*Variance is average squared deviation from the population mean.**

**Units are squared, which is not intuitive. Standard Deviation, 𝜎 = (𝑉𝑎𝑟𝑖𝑎𝑛𝑐𝑒)^1/2**

**Standard score, 𝑧 =(𝑥−𝜇)/ 𝜎( # of stdevs from the mean)**

**Var(X+a) = Var(X) (Variance does not change when a constant is added)**

**• Var(X+Y) = Var(X) + Var(Y) for Independent Observations**

**• Var(X-Y) = Var(X) + Var(Y)**

**\*Understanding the shape of a PDF – Skewness**

**\*A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means**

**median is less than mean.**

**𝑠𝑘𝑒𝑤 (𝑋) = 𝐸[(X − 𝜇 /𝜎)^3]**

**\*Understanding the shape of a PDF – Kurtosis**

**A measure of the ‘tailed’ness of the data distribution as compared to a normal distribution. Negative kurtosis means a distribution with light tails (fewer extreme deviations from mean (or outliers) than in normal distribution). Positive kurtosis means a distribution with heavy tails (more outliers than in normal distribution).**

**𝑒𝑥𝑐𝑒𝑠𝑠 𝑘𝑢𝑟𝑡 (𝑋) = 𝐸[(X − 𝜇 /𝜎)^4]-3**

**Measure Formula Description**

**Mean (𝜇): 𝐸(𝑋): Measures the Centre of the distribution of X**

**Variance (𝜎2): 𝐸[( 𝑋 − 𝜇 )^2]:Measures the spread of the distribution of X about the mean**

**Skewness :𝐸[(X − 𝜇 /𝜎)^3]:Measures asymmetry of the distribution of X**

**Kurtosis (excess) :𝐸[(X − 𝜇 /𝜎)^4]-3:Measures ‘tailed’ness of the distribution of X and useful in outlier identification.**

**\*Geometric: For estimating number of attempts before first success**

**\*Binomial: For estimating number of successes in n attempts**

**\*Poisson: For estimating n number of events in a given time period when on average we see m events**

**\*Exponential: Time between events**

**Central Limit Theorem:**

**\*Mean of all sample means of size n is the mean of the population.**

**𝐸 (𝑋) = 𝜇**

**\*Standard deviation of 𝑋 tells how far away from the population mean the sample mean is likely to be and is called the Standard Error of the Mean,**

**𝑉𝑎𝑟 𝑋 = 𝜎^2/ 𝑛**

**𝑆𝑡𝑎𝑛𝑑𝑎𝑟𝑑 𝐸𝑟𝑟𝑜𝑟 𝑜𝑓 𝑡ℎ𝑒 𝑀𝑒𝑎𝑛 = 𝜎/(𝑛)^1/2**

**\* Margin of error is the maximum expected difference between the true population parameter and**

**a sample estimate of that parameter.**

**𝑀𝑎𝑟𝑔𝑖𝑛 𝑜𝑓 𝐸𝑟𝑟𝑜𝑟 = 𝑧 ∗ 𝑆𝐸**

**Degrees of freedom:**

* **In statistics, the degrees of freedom (DF) indicate the number of independent values that can vary in an analysis without breaking any constraints.**

**(Or)**

* **Degrees of freedom are the number of independent values that a statistical analysis can estimate.**

**(Or)**

* **It as the number of values that are free to vary as you estimate parameters.**

**(Or)**

* **The degrees of freedom equal to your sample size minus the number of parameters you need to calculate during an analysis.**
* **what is the minimum number of observations required to estimate the regression?**

<https://www.youtube.com/watch?v=4otEcA3gjLk>

**Hypothesis Testing:**

**\*Hypothesis is a statement about the parameter distribution.**

**\*Hypothesis is normally converted to test of the mean or variance parameter of population (or differences in means or variations in populations)**

**\*** **Hypothesis tests give a way of using samples to test whether or not statistical claims are likely to be**

**true or not.**

**\*The hypothesis test only states whether the evidence is enough to reject the null hypothesis or not at the chosen significance level.**

**\*We always start with the assumption that Null Hypothesis is true.**

* **Null hypothesis: It is a tentative assumption about the population parameter.**
* **Cutoff value indicates when the Ho is rejected or not rejected**
* **The type of test to be conducted depends on alternative hypothesis**
* **Rejection region is the range of values that leads to reject the Ho**
* **A P value is a probability that provides measure of the evidence against the Ho provided by the sample.**

**What is p-value?**

* **Now, let’s say we pick any random value from this distribution. The probability that we will pick values close to the mean is highest as it has the highest peak (due to high occurrence values in that region).**
* **We can clearly see that if we move away from the peak, the occurrence of the values decreases rapidly and so does the corresponding probability, towards a very small value close to zero.**
* **P-value tells us about the ‘total probability’ of getting any value, when the values are picked randomly from the population distribution.**
* **P-value does not hold any value by itself. A large p-value implies that sample scores are more aligned or similar to the population score.**
* **P-value is an important metric in the process of feature selection. In feature selection, we try to find out the best subset of the independent variables to build the model.**
* **Null Hypothesis: The independent variable has no significant effect over the target variable**
* **Alternate Hypothesis: The independent variables have a significant effect on the target variable**

**Procedure to test hypothesis:**

**1.Identify parameter of interest (mean, variance, proportion) which you wish to test**

**2.Construct null and alternative hypothesis.**

**3.Compute test statistic which is function of the sample set of observations.**

**4.Derive distribution of the test statistic under the null hypothesis assumption.**

**5.Choose a test criterion (threshold) against which the test statistic is compared to reject or not reject the null hypothesis.**

**\*\*The performance of a hypothesis test depends on**

**1.Extent of variability in data**

**2.Number of observations (Sample size)**

**3.Test statistic (function of observations)**

**4.Test criterion(threshold)**

**\*Two types of tests in hypothesis:**

**1.one sided**

**2.two sided**

**\*Two types of errors in hypothesis testing**

**1.Type-I Error (Alpha)**

**2.Type -II Error (Beta)**

**\*Significance level is controlled by choosing test criterion from the distribution of the test statistic under the null hypothesis.**

**\*Significance Level(α) is a measure of how unlikely you want the results of the sample to be before you reject the null hypothesis, H0.**

**Statistical test power = 1-type II error**

**\*Trade-off: If we decrease the type I error probability, then type II probability will increase.**

**4 test statistics:**

**1.Closely related to Sampling Distribution of Means**

**Z And T**

**2.Closely related to Sampling Distribution of Variances • Derived from Normal Distribution**

**𝜒2 (Chi-squared) and F**

**\* If the sample size is small (<30), the variance of the population is not adequately captured by the**

**variance of the sample. Instead of z-distribution, t-distribution is used. It is also the appropriate**

**distribution to be used when population variance is not known, irrespective of sample size.**

**𝑡 𝑠𝑡𝑎𝑡𝑖𝑠𝑡𝑖𝑐 𝑜𝑟 𝑡 𝑠𝑐𝑜𝑟𝑒 ,𝑡 =( 𝑥 − 𝜇) /(𝑠/ (𝑛)^1/2)**

**Z-Test: Test coefficients of regression model**

**T-Test: Test coefficients of regression model**

**Chi-Square Test: Test Quality of regression model**

**F-Test:Choose between regression models having different number of parameters.**

**F distribution:**

* + **𝜒2 was useful in testing hypotheses about a single population variance.**
  + **Sometimes we want to test hypotheses about difference in variances of two populations.**
  + **Ratio of 2 variance estimates: 𝐹 =𝑠1^ 2/ 𝑠2 ^2 =𝑒𝑠𝑡.𝜎1 ^2 /𝑒𝑠𝑡.𝜎2 ^2**

**ANOVA:**

**The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.**

**Assumptions of T 𝑠𝑡𝑎𝑡𝑖𝑠𝑡𝑖𝑐:**

* **The data should follow a continuous or ordinal scale**
* **The observations in the data should be randomly selected**
* **The data should be normally distributed.**
* **Large sample size should be taken for the data to approach a normal distribution (although t-test is essential for small samples as their distributions are non-normal)**
* **Variances among the groups should be equal (for independent two-sample t-test)**
* **There are three types of t-tests we can perform based on the data at hand:**

**1.One sample t-test**

**2.Independent two-sample t-test**

**3.Paired sample t-test**

**Chi-Square Test:**

* **A Chi-Square test is a test of statistical significance for categorical variables.**
* **Chi-square test in hypothesis testing is used to test the hypothesis about the distribution of observations/frequencies in different categories.**
* **Uses of 𝝌𝟐 distribution:**

**To test goodness of fit.**

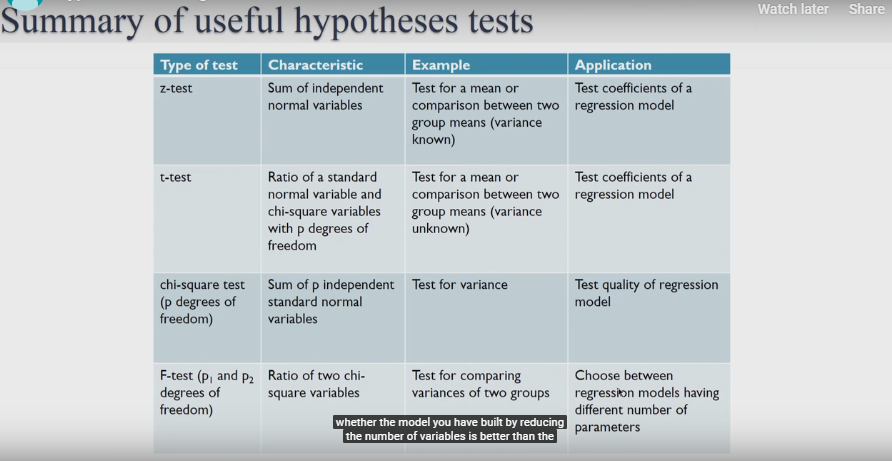
**To test independence of two variables.**

**To test hypothesis about variance of a population.**

**𝑋^2 = sum((𝑂−𝐸)^2 /𝐸)**

**𝜒2 =(𝑛 − 1)𝑠^2 /𝜎^2**

* **Assumptions:**
  + - **The χ2 assumes that the data for the study is obtained through random selection, i.e. they are randomly picked from the population.**
    - **The categories are mutually exclusive i.e. each subject fits in only one category.**
    - **The data should be in the form of frequencies or counts of a particular category and not in percentages.**
    - **The data should not consist of paired samples or groups or we can say the observations should be independent of each other.**
    - **When more than 20% of the expected frequencies have a value of less than 5 then Chi-square cannot be used. To tackle this problem: Either one should combine the categories only if it is relevant or obtain more data**



**Non Parametric Tests:**

* **Many real time examples tell us that average is not a good indicator of the center of the data. It can be extremely influenced by Outliers. In such cases, looking at median is a better choice. It is a better indicator of the center of the data because half of the data lies below the median and the other half lies above it.**
* **Statistical tests are used for making decisions.**
* **To perform analysis using median, we need to use non-parametric tests.**
* **Non-parametric tests are distribution independent tests whereas parametric tests assume that the data is normally distributed.**
* **Parametric tests are used when the information about the population parameters is completely known whereas non-parametric tests are used when there is no or few information available about the population parameters.**
* **To summarize, non-parametric tests can be applied to situations when:**

**1.The data does not follow any probability distribution**

**2.The data constitutes of ordinal values or ranks**

**3.There are outliers in the data**

**4.The data has a limit of detection**

* **The pros of using non-parametric tests over parametric tests are**

**1. Non-parametric tests deliver accurate results even when the sample size is small.**

**2. Non-parametric tests are more powerful than parametric tests when the assumptions of normality have been violated.**

**3. They are suitable for all data types, such as nominal, ordinal, interval or the data which has outliers.**

* **Cons**

**1. If there exists any parametric test for a data then using non-parametric test could be a terrible blunder.**

**2. The critical value tables for non-parametric tests are not included in many computer software packages so these tests require more manual calculations.**

<https://www.analyticsvidhya.com/blog/2017/11/a-guide-to-conduct-analysis-using-non-parametric-tests/>

**\*Euclidean Distance: Euclidean distance is calculated as the square root of the sum of the squared**

**differences between a new point (x) and an existing point (y).**

**\*Manhattan Distance: This is the distance between real vectors using the sum of their absolute**

**difference.**

**\*Hamming Distance: It is used for categorical variables. If the value (x) and the value (y) are same, the distance D will be equal to 0 Otherwise D=1.**

**\*A Variable is any number, characteristic or quantity that can be measured or counted.(Values can vary and it usually does within the population) .**

**Parametric and Nonparametric Machine Learning Algorithms:**

* **Machine learning can be summarized as learning a function (f) that maps input variables (X) to output variables (Y).**

**Y = f(x)**

* **An algorithm learns this target mapping function from training data.**
* **The form of the function is unknown, so our job as machine learning practitioners is to evaluate different machine learning algorithms and see which is better at approximating the underlying function.**

**Parametric Machine Learning Algorithms:**

* **Different algorithms make different assumptions or biases about the form of the function**
* **Assumptions can greatly simplify the learning process, but can also limit what can be learned. Algorithms that simplify the function to a known form are called parametric machine learning algorithms.**
* **A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.**
* **No matter how much data you throw at a parametric model, it won’t change its mind about how many parameters it needs.**
* **The algorithms involve two steps:**

**1.Select a form for the function.**

**2.Learn the coefficients for the function from the training data.**

* **The assumed functional form is a linear combination of the input variables and as such parametric machine learning algorithms are often also called “linear machine learning algorithms “.**
* **Examples of parametric machine learning algorithms include:**

**Linear Regression**

**Logistic Regression**

**Linear Discriminant Analysis**

**Perceptron**

**Naive Bayes**

**Simple Neural Networks**

* **Benefits of Parametric Machine Learning Algorithms:**

**Simpler: These methods are easier to understand and interpret results.**

**Speed: Parametric models are very fast to learn from data.**

**Less Data: They do not require as much training data and can work well even if the fit to the data is not perfect.**

* **Limitations of Parametric Machine Learning Algorithms:**

**Constrained: By choosing a functional form these methods are highly constrained to the specified form.**

**Limited Complexity: The methods are more suited to simpler problems.**

**Poor Fit: In practice the methods are unlikely to match the underlying mapping function.**

**Nonparametric Machine Learning Algorithms:**

* **Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data.**
* **Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don’t want to worry too much about choosing just the right features.**
* **Examples of popular nonparametric machine learning algorithms are:**

**k-Nearest Neighbors**

**Decision Trees like CART and C4.5**

**Support Vector Machines**

* **Benefits of Nonparametric Machine Learning Algorithms:**

**Flexibility: Capable of fitting a large number of functional forms.**

**Power: No assumptions (or weak assumptions) about the underlying function.**

**Performance: Can result in higher performance models for prediction.**

* **Limitations of Nonparametric Machine Learning Algorithms:**

**More data: Require a lot more training data to estimate the mapping function.**

**Slower: A lot slower to train as they often have far more parameters to train.**

**Overfitting: More of a risk to overfit the training data and it is harder to explain why specific predictions are made.**

* **A**parametric**algorithm has a**fixed number of parameters**.  A parametric algorithm is**computationally faster**, but makes**stronger assumptions**about the data; the algorithm may work well if the assumptions turn out to be correct, but it may perform badly if the assumptions are wrong.  A common example of a parametric algorithm is**linear regression**.**
* **In contrast, a**non-parametric**algorithm uses a**flexible number of parameters**, and the number of parameters often**grows as it learns from more data**.  A non-parametric algorithm is**computationally slower**, but makes**fewer assumptions**about the data.  A common example of a non-parametric algorithm is**K-nearest neighbor**.**
* **To summarize, the**trade-offs**between parametric and non-parametric algorithms are in**computational cost**and**accuracy**.**

**Algorithms:**

**Linear regression:**

* **Method of fitting the line of best fit is called least squares regression.**
* **Regression is basically a statistical approach to find the relationship between variables.**
* **Correlation coefficient, r, is a number between -1 and 1 and tells us how well a regression line fits the data.** **It gives the strength and direction of the relationship between two variables.**

**𝑟 =𝑏S𝑥 /S𝑦**

**where b is the slope of the line of best fit, 𝑠𝑥 is the standard deviation of the x values in the sample, and 𝑠𝑦 is the standard deviation of the y values in the sample.**

**Covariance:**

* **The value of covariance itself doesn’t say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).**

**– Affected by units (measuring height in ft vs mm)**

**– Not intuitive comparing covariance values between 2 sets of variables (how does height-weight covariance compare with oil price ($)-potato price (Rupee) covariance)**

**– Unintuitive units**

* **If both x and y are large distance away from their respective means, the resulting covariance will be even larger.**
* **The value will be positive if both are below the mean or both are above.**
* **If one is above and the other below, the covariance will be negative.**
* **If even one of them is very close to the mean, the covariance will be small.**
* **To know the strength of how the variables, move together, covariance is standardized to the dimensionless quantity, correlation.**

**Linear Regression: A regression method is called linear if the prediction f is a linear function of the unknown parameters w.**

**Y = WX+C**

* **The regression coefficient w represents the expected change in Y for a one unit change in X**.

**SST = SSR + SSE**

**\*R – Squared measures the how much variation the model explain or what proportion of the variation in the outcome Y can be explained by the covariates/predictors or the proportion of variation in Y being explained by the variation in X .**

**Degrees of freedom:**

**\*what is the minimum number of observations required to estimate this regression.**

**\*As degree of freedom decreases, (i.e more variables added to a given model )R-squared will only increases.**

**\* R2 assumes that every single variable explains the variation in the dependent variable. The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.**

**Important assumptions in regression analysis:**

* **There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X¹ is constant, regardless of the value of X¹. An additive relationship suggests that the effect of X¹ on Y is independent of other variables.**
* **There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.**
* **The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.**
* **The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to heteroscedasticity.**
* **The error terms must be normally distributed.**

**Assumptions:**

**1.Linearity with parameters. (If the function form is incorrect ,both coefficients and standard errors unreliable**

**2.constatnt error variance.(Homoscedasticity )**

**3.error terms are independent.**

**4.Normal Errors**

**5.No Multi-collinearity (X variables are themselves not related)**

**6.Exogenity (No omitted variable bias )**

**What if these assumptions get violated ?**

**Let’s dive into specific assumptions and learn about their outcomes (if violated):**

**1. Linear and Additive: If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.**

**How to check: Look for residual vs fitted value plots (explained below). Also, you can include polynomial terms (X, X², X³) in your model to capture the non-linear effect.**

**2. Autocorrelation: The presence of correlation in error terms drastically reduces model’s accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.** **Also, lower standard errors would cause the associated p-values to be lower than actual.**

**This will make us incorrectly conclude a parameter to be statistically significant.**

**How to check: Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, 0 < DW < 2 implies positive autocorrelation while 2 < DW < 4 indicates negative autocorrelation. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.**

**3**. **Multicollinearity**:**This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.**

**How to check:**

**1.Correlation factor**

**2.VIF(variance inflation factors):**

**a)Create auxiliary regression for each X variable**

**b)Find the VIF’s using the R-squared from each regression**

**VIF = 1/1-R-squared**

**4.Heteroskedasticity:**

**The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model’s performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.**

**How to check: You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.**

**5. Normal Distribution of error terms: If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.**

**How to check: You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.**

**Heteroscedasticity: It will occurs when variance(Y|X) is not constant**

**Standard errors in regression output are very large, we can make model should be significant when the standard errors are small .( The standard error is an estimate of the *standard deviation* of the coefficient, the amount it varies across cases)**

**Detection Methods:**

**1.Residual plot(Residual Vs X)**

**2.Goldfeld-Quandnt test:**

**a)order the data by the offending variable**

**b)split the data into two segments**

**c)run separate regression on each segment**

**d)compute variance estimates using f-statistic**

**3. Breusch-pagan test(whites test):**

**a)Estimate a single linear regression**

**b)save residuals,ei**

**c)Estimate a new regression using standard errors as Y**

**d)Find R-Squared**

**e) LM(LaGrange multiplier ) = Number of samples \* R-Squared ~ chi square**

**Correlation Vs Regression:**

* **Correlation is a statistical measure which determines association of two variables (i.e. dependent and independent variables). Regression on other hand side describes how an independent variable is numerically related to dependent variable.**
* **Correlation is used to represent the linear relationship between two variables. On the contrary, regression is used to fit the best line and estimate one variable on the basis of another variable.**
* **correlation, there is no difference between dependent and independent variables i.e. correlation between x and y is similar to y and x. Conversely, the regression of y on x is different from x on y.**
* **Correlation indicates the strength of association between variables. As opposed to, regression reflects the impact of the unit change in the independent variable on the dependent variable.**
* **Correlation aims at finding a numerical value that expresses the relationship between variables. Unlike regression whose goal is to predict values of the random variable on the basis of the values of fixed variable.**

**Logistic Regression:**

* **Logistic regression is similar to linear regression but it will predict whether something is true or not instead of predicting something continuous also instead of fitting line to the data it will fits an s shaped logistic function.**
* **The response y is binary: 1 if the coin is Head, 0 if the coin is Tail. This is represented by a Bernoulli variable where the probabilities are bounded on both ends (they must be between 0 and 1).**
* **Linear regression is only dealing with continuous variables instead of Bernoulli variables.**
* **The problem of Linear Regression is that these predictions are not sensible for classification since the true probability must fall between 0 and 1, but it can be larger than 1 or smaller than 0.**
* **Noted that classification is not normally distributed which is violated assumption Normality.**
* **Moreover, both mean and variance depend on the underlying probability. Any factor that affects the probability will change not just the mean but also the variance of the observations, which means the variance is no longer constantly violating the assumption 2: Homoscedasticity.**
* **Assumptions:**
  + **The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.**
  + **There is no influential values (extreme values or outliers) in the continuous predictors**
  + **There is no high intercorrelations (i.e. multicollinearity) among the predictors.**
  + **There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome**
* **Probability is the ratio of something happening to everything that could happen.**
* **Odds are the ratio of something happening to something not happening.**
* **The log of the odd ratio of the probabilities is called the logit function and forms basis for logistic regression.**
* **Logistic regression goal is to draw the “best fitting” squiggle for the data.**
* **In logistic regression we are fitting line with maximum Likelihood.**
* **Logs of odds makes everything symmetrical easier to interpret and easier for fancy statistics.**
* **The odds are not constrained to lie between 0 and 1 but can take any value from zero to infinity.**
* **Logistic Regression is a type of Generalized Linear Models.**
* **In Probability, we predict data based on known parameters.**
* **In Likelihood, we predict parameters based on known data.**
* **Odds ratio will determine the relationship between predictors and output variable. Larger value indicates good predictor and smaller value indicates not a good predictor.**
* **The problem in the logistic regression is that the transformation pushes the raw data to positive and negative infinity.**
* **There are three ways to determine if odds or log odds ratio statistically significant.**

**1.Fishers Exact Test**

**2.Chi-Square Test**

**3.The wald Test**

* **R2 in logistic regression = LL (overall probability)-LL(fit)/LL (overall probability)**
* **Chi squared value = 2(LL(fit) - LL (overall probability))**
* **Deviance or Residual Deviance is similar to SSE in the sense it measures how much remains unexplained by the model built with predictors included.**

**𝑫 = −𝟐𝑳𝑳, where LL is the log-likelihood.**

* **Null Deviance shows how well the model predicts the response with only the intercept as a parameter. The intercept is the logarithm of the ratio of number of cases with y=1 to the number of cases with y=0. This is similar to SST, which gives total variation when all coefficients are zero (null hypothesis).**
* **The goal of maximum likelihood is to find the optimal way to fit a distribution of data.**
* **Measure of a poorly fitted line that is analogous to SS(mean)**
* **AIC is a measure of the statistical model for a given set of data.AIC deals with trade-off between good of fit of the model and complexity of the model.**

**\*** **To start with logistic regression, I’ll first write the simple linear regression equation with dependent variable enclosed in a link function:**

**g(y) = βo + β(Age)**

**\*** **In logistic regression, we are only concerned about the probability of outcome dependent variable ( success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success(p) and Probability of Failure(1-p). p should meet following criteria:**

1. **It must always be positive (since p >= 0)**
2. **It must always be less than equals to 1 (since p <= 1)**

**\*Since probability must always be positive, we’ll put the linear equation in exponential form.**

**p = exp(βo + β(Age)) = e^(βo + β(Age))**

**\*** **To make the probability less than 1, we must divide p by a number greater than p. This can simply be done by:**

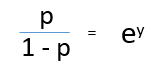
**p  =  exp(βo + β(Age)) / exp(βo + β(Age)) + 1   =   e^(βo + β(Age)) / e^(βo + β(Age)) + 1**

**\*** **we can redefine the probability as:**

**p = e^y/ 1 + e^y**

**If p is the probability of success, 1-p will be the probability of failure which can be written as:**

**q = 1 - p = 1 - (e^y/ 1 + e^y) *where* q is the probability of failure**



**p = 1/(1+e^-X)**

* **Maximum likelihood estimation (MLE) is an estimation method that allows to use a sample to estimate the parameters of the probability distribution that generated the sample.**

**Parameter estimation:**

* **Estimate the parameter values that best explain the observations.**
* **Calculate probability of new observations given data.**

**\*A loss function is for a single training example. It is also sometimes called an error function. A cost function, on the other hand, is the average loss over the entire training dataset. The optimization strategies aim at minimizing the cost function.**

**PCA:**

* **Assumptions: Correlation among features.**
* **Pca is linear transformation technique used for dimensionality reduction.**
* **Converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (principal components).**
* **Explains a large part of the variation in data using a small number of principal components.**
* **It will be identifying the directions of largest to smallest variance.**
* **Once we transform the data into the new coordinate system, we can choose to drop the variables which do not have much variance.**
* **Eigenvectors are the Principal Component directions.**
* **Eigenvalues are the magnitude of stretch.**
* **Eigenvalues represent the variance along those (new) directions.**
* **PCA is done only on predictor numerical variables.**
* **Largest variance principal component is called as dominant component.**
* **PCA does not always guarantee better performance. The main use is dimensionality reduction.**
* **Linear Discrement Analysis is also similar to Pca but It will consider classes into account. Lda Maximizes variance between classes and minimizes variance with in the class.**

**REGULARIZATION:**

* **Under fitting (Bias)**

**• High Bias problem**

**• High training error**

**• Low difference between training and test/validation errors**

* **Overfitting (Variance)**

**• High Variance problem**

**• Low training error**

**• High difference between training and test/validation errors**

* **Reducing overfitting by penalizing large coefficients is called Regularization.**
* **Regularization parameter λ controls the extent of penalization.**
* **The main idea behind Regularization is to find a new line that does not fit the training data as well.**

**Ridge regression:**

* **Penalization of large coefficients through L2 norm is a ridge regression ((another name for Least Squares Error).**
* **The penalty term “regularizes” the problem when there is perfect multicollinearity.**
* **Reduces the magnitude of coefficients (does not make them zero)**
* **When the sample sizes are relatively small, then ridge regression can improve predictions from new data (reduce variance) by making predictions less sensitive to the training data.**

**LASSO Regularization:**

* **Penalization of large coefficients through L1 norm is a ridge regression ((another name for Least Absolute Error)**
* **Least Absolute Shrinkage and Selection Operator (LASSO).**
* **Values of irrelevant coefficients set to zero**
* **Good for feature selection**

**Elastic-net:**

* **A mix-between Ridge and LASSO regression.**
* **If some of your covariates are highly correlated, you may want to look at the Elastic Net instead of the LASSO.**
* **• α=0 is Ridge**

**• α=1 is LASSO**

**• α=0.5 is a equal mix of the two**

* **The big difference between ridge and lasso regression is that ridge can only shrink the slope asymptotically close to 0 and while lasso can shrink the slope all the way to 0.**
* **Lasso better when the model contains a lot useless variables than ridge .**
* **Ridge regression better when most variables are useful to build model.**

**Discriminative Vs Generative:**

* **All classification problems are essentially equivalent to evaluating conditional probability.**
* **P(Yi | X), i.e., given certain evidence X, what is the probability that this is from class Yi.**
* **To solves this problem by modelling the probabilistic relationship between X and Y (sigmoid function, linear in X, etc.) directly.** **Such models are called Discriminative models.**
* **In Discriminative approaches model will assign weight to each feature.**
* **A positive weight votes that this configuration is likely correct.**
* **A negative weight votes that this configuration is likely incorrect.**

**Ex: Logistic Regression**

* **Computes P(Yi | X) by using Bayes theorem and instead computes the inverse conditional probability P(X | Yi ).** **These type of methods are called Generative Learning Models.**
* **A simple classifier that performs surprisingly well on a large class of problems.**

**Ex: Naïve Bayes Algorithm**

**Naïve-Bayes:**

* **All Features in dataset will independently contribute to the probability that is why it is known as ‘Naive’.**
* **Assumptions:**
* **Conditional probability of each feature given the class is independent of all other features.**
* **It works best when all the predictor variables are categorical variables.**
* **Very frequently used in text mining, character image analysis problems.**
* **Still Naïve Bayes does surprisingly well in a lot of situations.**
* **For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).**

**.**

**P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).**

**P(c) is the prior probability of class (Marginal likelihood)**

**P(x|c) is the likelihood which is the probability of predictor given class.**

**P(x) is the prior probability of predictor.**

* **Algorithm will use normal distribution for calculating likelihood for a new value of x (continuous variable).**
* **Limitations:**
* **It performs well in case of categorical input variables compared to numerical variable(s).**
* **If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”.** **To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.**
* **Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.**

**Laplace Smoothing:**

**Parameter Estimation with add 1 smoothing:**

**Let cc refer to a class (such as Positive or Negative), and let ww refer to a token or word.**

**The maximum likelihood estimator for P(w|c)P(w|c) is**

**count(w,c)count(c)=counts w in class ccounts of words in class c.count(w,c)count(c)=counts w in class ccounts of words in class c.**

**This estimation of P(w|c)P(w|c) could be problematic since it would give us probability 00 for documents with unknown words. A common way of solving this problem is to use Laplace smoothing.**

**Let V be the set of words in the training set, add a new element UNKUNK (for unknown) to the set of words.**

**Define**

**P(w|c)=count(w,c)+1count(c)+|V|,P(w|c)=count(w,c)+1count(c)+|V|,**

**where VV refers to the vocabulary (the words in the training set).**

**In particular, any unknown word will have probability**

**1count(c)+|V|+1.**

**Note:For More information visit** [**https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/**](https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/)

**Decision Tree:**

* **Decision tree is nothing but a nested if-else conditions.**
* **Tree Structured classifier. (Follows the nested if –else architecture)**
* **Continuous splitting of the feature space.**
* **Recursive Partitioning of the feature space into regions.**
* **Assumptions:**
* **At the beginning, the whole training set is considered as the root.**
* **Feature values are preferred to be categorical.**
* **If the values are continuous then they are discretized prior to building the model.**
* **Order to placing attributes/Features as root or internal node of the tree is done by using some statistical approach.**
* **Records are distributed recursively on the basis of attribute values.**
* **Split measure/Impurity Measure**

**• Classification Error**

**• Gini Index**

**• Entropy**

* **Advantages:**
* **Easily understandable.**
* **Interpretability**
* **Impurity measure used by CART in order to choose the attribute to split. While building a decision tree, at each level, the algorithm tries to select the attribute, which on splitting, will give pure sets (or relatively pure sets). The attribute which gives minimum on splitting is chosen. For a perfect classification, i.e. pure nodes, Impurity measure would be zero.**
* **Information gain measures how well a given attribute separates the training examples according to their target classification.**
* **Pruning Types:**

1. **Reduce Error pruning (Merging leafs under nodes)**
2. **Cost Complexity pruning(Collapsing inner nodes)**

**Ensemble Methods:**

**1.Bagging:**

* **Bagging: Bootstrap Aggregation.**
* **Bootstrap sampling is nothing but sampling with replacement.**
* **In Bagging Every sample will contain 63% of unique observations.**
* **Parallelization Approach.**
* **Reduce variance.**
* **Particularly used for when we are working with unstable classifier.**
* **Same estimator and different data sets.**

**Ex:** **RandomForestClassifier, Bagging Classifier**

* **The fundamental difference is that in Random forests, only a subset of features is selected at random out of the total and the best split feature from the subset is used to split each node in a tree, unlike in bagging where all features are considered for splitting a node.**
* **The base estimator in RandomForestClassifier is Decision Tree. In Bagging classifier, we can use any supervised learning algorithm as an estimator.**

**2.Committie Machine:**

* **In this we are training multiple algorithms on same data and will assign same weight to each classifier.**

**3.Stacking:**

* **It is similar to committee machine but instead of assigning weights to estimators as arbitrarily, it will learn weights using predictions.**

**4.Boosting:**

* **Boosting is stage wise process or sequential process.**

**Ex: 1. AdaBoost ------>Exponential loss**

**2. Logit boost -->logistic loss**

**3. Gradient boosting-->**

* **Reduce Bias.**

**KNN:**

* **KNN is a instance based learning, it is also called as lazy learning algorithm.**
* **Assumptions:**

**1.k-NN performs much better if all of the data have the same scale.**

**2.k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large.**

**3.k-NN makes no assumptions about the functional form of the problem being solved.**

* **Every algorithm has two phases 1. Training and 2. Prediction/Testing.**
* **In the Training phase KNN will save examples.**
* **Prediction time is more than Training time.**

**Drawbacks of Perceptron:**

* **Only separates or classifies linearly separable data.**
* **Classification or Discriminant boundary depends on the initialization of weights.**

**Or**

* **It will not provide same hyper plane every time.**

**SVM:**

* **The Assumption of svm is all points are linearly separable by linear decision surface.**
* **Svm able to handle certain cases where there is nonlinearity by using nonlinear basis functions(Kernels)**
* **The samples which are the closest to the line are support vectors.**
* **The perpendicular distance between hyperplane and closest sample is called as margin.**
* **The motive of svm is to find the W should classify all the training examples correctly and have a maximum margin.**
* **The distance between Positive closest point to the decision surface Is equal to negative closest point to the decision surface.**
* **Since the closest samples are only supporting the margin hence they are called as support vectors.**
* **Every sample have LaGrange multiplier calculated by loss function using quadratic programming.**
* **Which samples having non zero value LaGrange multiplier those are called as support vectors.**
* **Svm is classifier with the maximum margin width is robust to outliers and this has strong generalization ability.**
* **Such models which are try to estimate probability P(X) from a large number of samples are called generative models.**
* **Choosing which kernel transformation to apply depends a lot on your dataset. Unfortunately, there is no perfect recipe, and it will come with experience via trial and error.**
* **A kernel is a function that returns the result of a dot product performed in another**

**space.**

* **SVMs are also called sparse kernel machines. It is because they only need to compute the kernel function on the support vectors and not on all the vectors.**
* **Kernel types:**

1. **Linear kernel:**
   * + - * **Linear kernel works well for text classification.**
         * **It is simply defined by: K(x,x’) = x.x’**

**where x and x’ are two vectors.**

1. **Polynomial kernel:**
   * + - * **More generic version of the kernel:**

**K(x,x’) = (x.x’+c)^d**

* + - * + **It has two parameters: c which represents a constant term, and d which represents the degree of the kernel**
        + **A polynomial kernel with a degree of 1 and no constant is simply the linear kernel.**
        + **When you increase the degree of a polynomial kernel, the decision boundary will become more complex leads to overfitting.**

1. **RBF or Gaussian kernel:**
   * + - * **Sometimes polynomial kernels are not sophisticated enough to work. When you have a difficult dataset.**
         * **More complicated, kernel: The Gaussian kernel. It is also named RBF kernel, where RBF stands for Radial Basis Function.**
         * **A radial basis function is a function whose value depends only on the distance from the origin or from some point.**
         * **The RBF kernel function is:**

**K(x,x’) = exp(-r||x-x’||^2)**

* + **A kernel is a function that returns the result of a dot product performed in another space.**
  + **A kernel is a measure of the similarity between two vectors**
  + **When gamma is too small, the model behaves like a linear SVM. When gamma**

**is too large, the model is too heavily influenced by each support vector.**

**LSI:**

* **In tf idf and count vectorizer computing similarity based on words but lsi is based on topics or clusters of similar words.**
* **Lsi will capture association between words**
* **The concept goal of lsi add missing synonyms to the documents.**
* **Assumption of lsi is words that are close in meaning will occur in similar pieces of text.**
* **The resulting new dimensions or topics might be difficult to interpret.**

**Clustering:**

* **Grouping data points together based on similarity is called as clustering.**
* **Clustering maximize the similarity between data points in same cluster and minimize the similarity between clusters.**
* **Useful at**

1. **Categorization**
2. **Preprocessing Tool**
3. **Visualization (Understand something about data in a feature space)**
4. **Outlier Mining.**

* **Types of clustering:**

1. **Partitional (Directly Search through the in the space of all Portions)**
2. **Hierarchical**
3. **Density Based**

* **Clustering Measures:**

1. **Diameter (Average/max pairwise distance in same cluster)**
2. **Radius (Average distances to the centroid)**
3. **Purity (Large portion of data points to the same class)**
4. **Entropy**
5. **Rand Index (Which is used when you have reference clustering & What Fraction of pairs classified correctly)**

**K-Means/K-Mediods/PAM:**

* **Repeat steps until the centroids no longer changes.**
* **Repeat Multiple Times.**
* **Assumptions:**

1. **There are 'k' clusters in the data**
2. **Each point in the cluster is closest only to its cluster center**
3. **Clusters are spatially grouped—or “spherical”**
4. **Clusters are of a similar size**

* **When should I use:**
  + - **Your data is numeric .It does not work with categorical features .we are computing distance between real numbers**
    - **useful when you have an idea of how many clusters exists in your space.**
    - **If you don’t have labels for your data.**

**Hierarchical Clustering:**

* **Produced a nested sequence of clusters. This clustering structure is called Dendogram.**
* **Entire graph generated in a single shot .**
* **Two types 1. Bottom- up approach (Agglomerative Clustering)**

**2. Top-Down Approach (Devise Clustering)**

**Distance Measures:**

**1.Centroid based**

**2.Single link (Smallest sample pair wise distance of two clusters)**

**3.Average Link**

**4.Complete link (maximum of pair wise distance)**

**5.Radius Based**

**6.Daimeter**

**Density Based:**

* **The Assumption of dB scan is clusters are dense regions in space separated by regions of lower density.**
* **Two parameters 1. Min points(How many points would consider as a dense)**

**2.Epsilon(area which you will perform the count)**

* **Core point is point that is lies in a high density region.**
* **A point is density reachable if there is a core point from which we can reach this point by traversing through only core points.**
* **Two points I and j density connected if there exists a core point k from which both of them are density reachable.**

**\*K-means is an unsupervised learning algorithm used for clustering problem whereas KNN is a supervised learning algorithm used for classification and regression problem. This is the basic difference between K-means and KNN algorithm.**

**Feature Engineering:**

* **A variable is any characteristic, number, or quantity that can be measured or counted.**
* **There are two types of variables**

**1.Numerical**

**a. Continuous**

**b. Discrete**

**2.Catgorical**

**a. Nominal**

**b. Ordinal**

**c. Date/Time**

* **A variable which values are whole numbers (counts) is called discrete.**

**Examples:**

**• Number of items bought by a customer in a supermarket (10, 50, …)**

**• Number of active bank accounts of a borrower (1, 4, 7, ...)**

**• Number of pets in the family**

**• Number of children in the family**

* **A variable that may contain any value within some range is called continuous.**

**Examples:**

**• Amount paid by a customer in a supermarket ($32.50, $12, $5.20, …)**

**• House price (GBP 350,000, GBP 57000, GBP 1,000,000, ...)**

**• Time spent surfing a website (3.4 seconds, 5.10 seconds, ...)**

**• Total debt as percentage of total income in the last month (0.2, 0.001, 0, 0.75, ...)**

* **Categorical variables in which categories can be meaningfully ordered are called ordinal. Examples:**

**• Student's grade in an exam (A, B, C or Fail)**

**• Days of the week (Monday = 1 and Sunday = 7)**

**• Educational level, with the categories: Elementary school, High school, College graduate and PhD ranked from 1 to 4**

* **No intrinsic order of the labels are called as Nominal**

**Examples:**

**• Country of birth (Argentina, England, Germany)**

**• Postcode**

**• Vehicle make (Citroen, Peugeot, ...)**

* **Variable Characteristics:**

1. **MISSING DATA (Lack of information for some observations within a variable.)**
2. **CATEGORICAL VARIABLES (Variables contain strings instead of numbers)**
3. **LINEAR MODEL ASSUMPTIONS (Do the variables fulfill those assumptions?)**
4. **DISTRIBUTIONS (Normal, skewed and others)**
5. **OUTLIERS (Unusual or unexpected values)**
6. **FEATURE MAGNITUDE(Scale of the different features)**

**Missing Data Imputation:**

* **Missing data, or missing values, occur when no data is stored for a certain observation in a variable.**
* **Missing data are a common occurrence in most datasets**
* **Missing data can have a significant effect on the conclusions that can be drawn from the data.**
* **Understanding missing data mechanisms may help us to choose the right missing data imputation technique.**
* **There are three mechanisms to lead missing data They are**

**1.Missing data completely at random (MCAR)**

**2. Missing data at random(MAR)**

**3. Missing data not at random(MNAR)**

* **Imputation is the act of replacing missing data with statistical estimates of the missing values.**
* **The goal of any imputation technique is to produce a complete dataset that can be used to train machine learning models.**

**MCAR:**

* **The probability of being missing is the same for all the observations.**
* **There is absolutely no relationship between the data missing and any other values, observed or missing, within the dataset.**
* **Disregarding those cases would not bias the inferences made**

**MAR:**

* **The probability of an observation being missing depends on available information.**

**MNAR:**

* **There is a mechanism or a reason why missing values are introduced in the dataset.**

**CATEGORICAL VARIABLES:**

* **The number of different labels in a categorical column is known as cardinality.**
* **For highly cardinal variables:**

**• Some labels may appear only in train set -->over-fitting**

**• Some labels may appear only in test set -->model will not know how to interpret the values**

* **Variables with too many labels tend to dominate over those with fewer labels, particularly in tree based algorithms.**
* **A big number of labels within a variable may introduce noise with little, if any, information.**
* **Reducing cardinality may help improve model performance.**

**Rare Labels:**

* **Rare labels are those that appear only in a tiny proportion of the observations in a dataset.**
* **Rare labels may cause over-fitting and operationalization problems.**
* **Hard to understand the role of the rare label on the outcome prediction.**
* **Removing rare labels may improve model performance.**

**Feature Magnitude:**

* **Features with bigger magnitudes dominate over features with smaller magnitudes.**
* **The regression coefficient is directly influenced by the scale of the variable.**
* **Gradient descent converges faster when features are on similar scales.**
* **Feature scaling helps decrease the time to find support vectors for SVMs.**
* **Euclidean distances are sensitive to feature magnitude.**
* **The machine learning models affected by the magnitude of the feature:**

**• Linear and Logistic Regression**

**• Neural Networks**

**• Support Vector Machines**

**• KNN**

**• K-means clustering**

**• Linear Discriminant Analysis (LDA)**

**• Principal Component Analysis (PCA)**

* **Machine learning models insensitive to feature magnitude are the ones based on Trees:**

**• Classification and Regression Trees**

**• Random Forests**

**• Gradient Boosted Trees**

**Imputation Techniques:**

1. **Complete case Analysis:**
   * + **CCA Discarding observations where values in any of the variables are missing.**
     + **Suitable for categorical and numerical variables.**
     + **Assumptions:**

**Data is missing at completely random.**

* + - **Advantages**
      * **Simple**
      * **No data manipulation required**
      * **Preserves the distribution of the variables**
    - **Limitations:**
      * **It can exclude a large fraction of the original dataset (if missing data is abundant)**
      * **Excluded observations could be informative for the analysis (if data is not missing at random)**
      * **When using our models in production, the model will not know how to handle missing data**
    - **When to use:**
      * **Data is missing completely at random**
      * **No more than 5% of the total dataset contains missing data**

1. **Mean / Median imputation:**
   * + **Mean / median imputation consists of replacing all occurrences of missing values (NA) within a variable by the mean or median.**
     + **Suitable numerical variables**
     + **If the variable is normally distributed the mean and median are approximately the same**
     + **If the variable is skewed, the median is a better representation**
     + **Assumptions:**
       - **Data is missing at random.**
       - **The missing observations, most likely look like the majority of the observations in the variable (aka, the mean / median)**
     + **Advantages:**
       - **Easy to implement**
       - **Fast way of obtaining complete datasets**
       - **Can be integrated in production (during model deployment)**
     + **Limitations:**
       - **Distortion of the original variable distribution.**
       - **Distortion of the original variance**
       - **Distortion of the covariance with the remaining variables of the dataset**
       - **The higher the percentage of NA, the higher the distortions.**
     + **When to use:**
       - **Data is missing at random**
       - **No more than 5% of the variable contains missing data.**
     + **The mean or median value should be calculated only in the train set and used to replace NA in both train and test sets to avoid overfitting.**
2. **Arbitrary value imputation:**
   * + **Arbitrary value imputation consists of replacing all occurrences of missing values (NA) within a variable by an arbitrary value.**
     + **Suitable numerical and categorical variables**
     + **Categorical: “Missing”**
     + **Typically used arbitrary values are 0, 999, -999 (or other combinations of 9s) or -1 (if the distribution is positive).**
     + **Assumptions:**
       - **Data is not missing at random.**
       - **If this is the case, we want to flag the missing values with a different (arbitrary) value, instead of replacing those occurrences with the mean or the median, which represent the most common value.**
     + **Advantages:**

**• Easy to implement, Fast way of obtaining complete datasets**

**• Can be integrated in production (during model deployment)**

**• Captures the importance of being "missing" if there is one**

* + - **Limitations:**

**• Distortion of the original variance, Distortion of the original variable distribution**

**• Distortion of the covariance with the remaining variables of the dataset**

**• If the arbitrary value is at the end of the distribution it may mask or create outliers**

**• Need to be careful not to choose an arbitrary value too similar to the mean or median (or any other common value of the variable distribution)**

**• The higher the percentage of NA, the higher the distortions**

* + - **When to use:**

**Data are not missing at random**

1. **End of tail imputation:**

* **End of tail imputation is equivalent to arbitrary value imputation, but automatically selecting arbitrary values at the end of the variable distributions.**
* **If the variable is normally distributed, we can use the mean plus or minus 3 times the standard deviation.**
* **If the variable is skewed, we can use the IQR proximity rule.**
* **Suitable numerical variables**
* **The values to replace missing data should be calculated only on the train set.**

1. **Frequent Category imputation:**

* **Mode imputation consists of replacing all occurrences of missing values (NA) within a variable by the mode, or the most frequent value.**
* **Suitable numerical and categorical variables.**
* **In practice, we use this technique with categorical variables.**
* **Assumptions:**
  + - * + **Data is missing at random**
        + **The missing observations, most likely look like the majority of the observations (aka, the mode)**
    - **Advantages:**
      * **Easy to implement**
      * **Fast way of obtaining complete datasets**
      * **Can be integrated in production (during model deployment)**
* **Limitations:**

**Distortion the relation of the most frequent label with other variables within the dataset**

**• May lead to an over-representation of the most frequent label if there is a big number of NA**

**• The higher the percentage of NA, the higher the distortions**

* + - **When to use:**
      * **Data is missing at random**
      * **No more than 5% of the variable contains missing data.**

1. **Missing Category imputation:**

* **This method consists in treating missing data as an additional label or category of the variable**
* **Missing observations are grouped in the newly created label 'Missing'.**
* **This is the most widely used method of missing data imputation for categorical variables.**
* **Suitable for categorical variables**
* **Advantages:**
  + - * + **Easy to implement**
        + **Fast way of obtaining complete datasets**
        + **Can be integrated in production (during model deployment)**
        + **Captures the importance of "messiness" if there is one**
        + **No assumption made on the data**
* **Limitations:**

**If the number of NA is small creating an additional category is in essence adding another rare label to the variable.**

1. **Random sample imputation:**
   * + **Random sampling consists in taking a random observation from the pool of available observations of the variable, and using that randomly extracted value to fill the NA.**
     + **Suitable for both numerical and categorical variables**
     + **Assumptions:**
       - **Data is missing at random**
       - **The idea is to replace the population of missing values with a population of values with the same distribution of the original variable.**
     + **Advantages:**
       - **Easy to implement**
       - **Fast way of obtaining complete datasets**
       - **Can be integrated in production (during model deployment)**
       - **Preserves the variance of the variable**
     + **Limitations:**
       - **Randomness (Every time we score the same observation, we may obtain a different prediction)**
       - **The relationship of imputed variables with other variables may be affected if there are a lot of NA**
       - **Memory heavy for deployment, as we need to store the original training set to extract values from and replace the NA in coming observations.**
     + **The population of values used to replace NA should be the train set To avoid over-fitting.**
2. **Missing indicator:**
   * + **A Missing Indicator is an additional binary variable, which indicates whether the data was missing for an observation (1) or not (0).**
     + **Suitable for numerical and categorical variables**
     + **The Missing Indicator is used together with methods that assume data is missing at random:**

**• Mean, median, mode imputation**

**• Random sample imputation**

* + - **Assumptions:**
      * **Data is NOT missing at random**
      * **Missing data are predictive**
    - **Advantages:**
      * **Easy to implement**
      * **Captures importance of missing data**
      * **Can be integrated in production (during model deployment)**
    - **Limitations:**
      * **Expands the feature space**
      * **Original variable still needs to be imputed**
      * **Many missing indicators may end up being identical or very highly correlated**
    - **When to use:**

**if the data was missing completely at random, this would be captured by the mean, median or mode imputation, and if it wasn't this would be captured by the additional "missing indicator" variable.**

**Categorical variable encoding:**

* **Categorical encoding refers to replacing the category strings by a numerical representation.**
* **Categorical Encoding Techniques:**

1. **Traditional techniques:**
   1. **One hot encoding**
   2. **Count / frequency encoding**
   3. **Ordinal / Label encoding**
2. **Monotonic relationship:(** **Improve the performance of linear models)**
   1. **Ordered label encoding**
   2. **Mean encoding**
   3. **Weight of evidence**
3. **Alternative techniques:**
   1. **Binary encoding**
   2. **Feature hashing**
   3. **Others**

**Categorical Encoding Techniques:**

* **One hot encoding:**
  + **One hot encoding, consists in encoding each categorical variable with a set of Boolean variables which take values 0 or 1, indicating if a category is present for each observation.**
  + **More generally, a categorical variable should be encoded by creating k-1 binary variables, where k is the number of distinct categories.**
  + **One hot encoding into k-1 binary variables takes into account that we can use 1 less dimension and still represent the whole information**
  + **Advantages:**

**• Makes no assumption about the distribution or categories of the categorical variable**

**• Keeps all the information of the categorical variable**

**• Suitable for linear models**

* + **Limitations:**

**• Expands the feature space**

**• Does not add extra information while encoding**

**• Many dummy variables may be identical, introducing redundant information**

* **One hot encoding of top categories:**

**Does not keep the information of the ignored labels**

* **Label / integers encoding:**
  + **Integer encoding consist in replacing the categories by digits from 1 to n (or 0 to n-1, depending the implementation), where n is the number of distinct categories of the variable..**
  + **Advantages:**

**• Straightforward to implement**

**• Does not expand the feature space**

**• Can work well enough with tree based algorithms**

* + **Limitations:**

**• Does not add extra information while encoding**

**• Not suitable for linear models**

**• Does not handle new categories in test set automatically**

* **Count / frequency encoding:**
  + **Categories are replaced by the count or percentage of observations that show that category in the dataset.**
  + **Very popular encoding method in Kaggle competitions.**
  + **Advantages:**

**• Straightforward to implement**

**• Does not expand the feature space**

**• Can work well enough with tree based algorithms**

* + **Limitations:**

**• Not suitable for linear models**

**• Does not handle new categories in test set automatically**

**• If 2 different categories appear the same amount of times in the dataset, that is, they appear in the same number of observations, they will be replaced by the same number: • may lose valuable information.**

* **Ordered ordinal encoding:**
  + **Categories are replaced by integers from 1 to k, where k is the number of distinct categories in the variable, but this numbering is informed by the mean of the target for each category.**
  + **Advantages:**

**• Straightforward to implement**

**• Does not expand the feature space**

**• Creates monotonic relationship between categories and target**

* + **Limitations:**

**• May lead to over-fitting**

**• Difficult to implement together with cross validation with current libraries**

* **Mean encoding:**
  + **Mean encoding implies replacing the category by the average target value for that category**
  + **Advantages:**

**• Straightforward to implement**

**• Does not expand the feature space**

**• Creates monotonic relationship between categories and target**

* + **Limitations:**

**• May lead to over-fitting**

**• Difficult to implement together with cross validation with current libraries**

**• If 2 categories show the same mean of target, they will be replaced by the same number => potential loss of value**

* **Weight of evidence:**
  + **WoE = ln (p(1) / p(0) )**
  + **Limitations:**

**• May lead to over-fitting**

**• Not defined when the denominator is 0**

* + **Advantages:**
    - **Creates a monotonic relationship between the target and the variables.**
    - **It orders the categories on a "logistic" scale which is natural for logistic regression**
    - **The transformed variables can then be compared because they are on the same scale. • Therefore, it is possible to determine which one is more predictive.**
* **Rare Label Encoding:**
  + **Rare labels are those that appear only in a tiny proportion of the observations in a dataset**
  + **Grouping Rare Labels**
  + **This way, categories that are new in test set, will be treated as rare and put into the Rare group**
* **Feature hashing:**
  + **Use a hashing method, any of choice, to encode the variables**
  + **May lead to different labels taking the same value**

**Variable Transformation:**

* **Variables follow a Gaussian Distribution**
* **Normality can be assessed with histograms and Q-Q plots**
* **Transformations:**
* **Logarithmic Transformer**
* **Reciprocal Transformer**
* **Exponential Transformer**
* **Power Transformer**
* **BoxCox Transformer**
* **Yeo-Johnson Transformer**

**Discretization:**

* **Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values.**
* **Discretization is also called binning, where bin is an alternative name for interval.**
* **Outliers are placed into the lower or upper intervals, together with the remaining inlier values at the ends of the distribution.**
* **Discretization approaches:**
  + **Unsupervised:**

1. **Equal-width**
2. **Equal frequency**
3. **K means**
   * **Supervised:**
4. **Decision Trees**

**Feature-Scaling:**

* **The regression coefficient is directly influenced by the scale of the variable**
* **Variables with bigger magnitude / value range dominate over the ones with smaller magnitude / value range.**
* **Gradient descent converges faster when features are on similar scales.**
* **Feature scaling helps decrease the time to find support vectors for SVMs**
* **Euclidean distances are sensitive to feature magnitude.**
* **The machine learning models affected by the magnitude of the feature:**

**• Linear and Logistic Regression**

**• Neural Networks**

**• Support Vector Machines**

**• KNN • K-means clustering**

**• Linear Discriminant Analysis (LDA)**

**• Principal Component Analysis (PCA)**

* **Machine learning models insensitive to feature magnitude are the ones based on Trees:**

**• Classification and Regression Trees**

**• Random Forests**

**• Gradient Boosted Trees**

* **Feature scaling refers to the methods used to normalize the range of values of independent variables.**
* **In other words, the methods to set the feature value range within a similar scale.**
* **Feature scaling is generally the last step in the data preprocessing pipeline, performed just before training the machine learning algorithms.**
* **Scaling methods:**

**• Standardization**

**• Mean normalization**

**• Scaling to maximum and minimum**

**• Scaling to absolute maximum**

**• Scaling to median and quantiles**

**• Scaling to unit norm**

**1. Standardization:**

* **Centers the mean of variable at zero and sets the variance to 1.**
* **Z-score = (𝑋 −𝑚𝑒𝑎𝑛(𝑋) )/𝑆𝑡𝑑(𝑋)**
* **Preserves the shape of the original distribution**
* **Preserves outliers.**

**2. Mean Normalization:**

* **Centers the variable at 0 and re-scales the variable to the value range.**
* **X\_scaled = (𝑋 −𝑚𝑒𝑎𝑛(𝑋))/ (𝑚𝑎𝑥 𝑋 −min(𝑋))**
* **Variance varies**
* **May alter shape of the original distribution**
* **Minimum and maximum values within [-1;1]**
* **Preserves outliers**

**3. MinMaxScaling:**

* **Scales the variable between 0 and 1**
* **X\_scaled= (𝑋 −𝑚𝑖𝑛(𝑋))/ (𝑚𝑎𝑥 𝑋 −min(𝑋))**
* **Mean varies**
* **Variance varies**
* **May alter shape of the original distribution**
* **Minimum and maximum values within [0;1**
* **Preserves outliers**

**4. MaxAbsScaling:**

* **Scales the variable between -1 and 1**
* **X\_scaled= 𝑋/ 𝑚𝑎𝑥 (𝑋)**
* **Mean not centered**
* **Variance not scaled**

**5.** **Robust Scaling:**

* **Scaling to median and IQR**
* **X\_scaled= (𝑋 −𝑚𝑒𝑑𝑖𝑎𝑛(𝑋) )/(75𝑡ℎ 𝑞𝑢𝑎𝑛𝑡 𝑋 −25𝑡ℎ 𝑞𝑢𝑎𝑛𝑡(𝑋))**
* **Median centered at zero**
* **Handles outliers**

**6.** **Scaling to unit norm:**

* **In scaling to unit norm, we divide each feature vector by either the Manhattan distance (l1 norm) or the Euclidean distance of the vector (l2 norm).**

**Mixed Variables:**

**Two types:** 1. **Numbers or labels in different observations**

**2.** **Numbers and labels in the same observation**

**Feature Selection:**

* **Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in machine learning model building.**
* **Feature Selection is a very critical component in a Data Scientist’s workflow. When presented data with very high dimensionality, models usually choke because**

**1.Training time increases exponentially with number of features.**

**2.Models have increasing risk of overfitting with increasing number of features**

* **Computationally expensive**
* **Feature Selection methods helps with these problems by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance.**
* **Feature selection techniques:**

1. **Filter Methods**
2. **Wrapper Methods and**
3. **Embedded Methods**

**1.Filter Methods:**

* **Filter Methods considers the relationship between features and the target variable to compute the importance of features.**
* **Select variables independently of the machine learning algorithm**
* **Rank features according to a certain criterion**
* **May select redundant variables because they do not consider the relationships between features.**

1. **F Test:** **There are some drawbacks of using F-Test to select your features. F-Test checks for and only captures linear relationships between features and labels. A highly correlated feature is given higher score and less correlated features are given lower score.** **Correlation is highly deceptive as it doesn’t capture strong non-linear relationships.**
2. **Variance Threshold:** **This method removes features with variation below a certain cutoff.**
3. **Chi –sqaure: evaluate only categorical variables in classification task.**
4. **Univariate roc\_auc: Build model per each feature**
5. **Mutual Information:** **Mutual Information between two variables measures the dependence of one variable to another. If X and Y are two variables, and If X and Y are independent, then no information about Y can be obtained by knowing X or vice versa. Hence their mutual information is 0.**

**If X is a deterministic function of Y, then we can determine X from Y and Y from X with mutual information**

**When we have Y = f(X,Z,M,N), 0 < mutual information < 1**

**Advantage of using mutual information over F-Test is, it does well with the non-linear relationship between feature and target variable.**

**2.Wrapper Methods:**

* **Evaluate the features in the light of a specific machine learning algorithm.**
* **Detect interactions between variables**
* **Find the optimal feature subset for the desired classifier**
* **Computationally expensive**

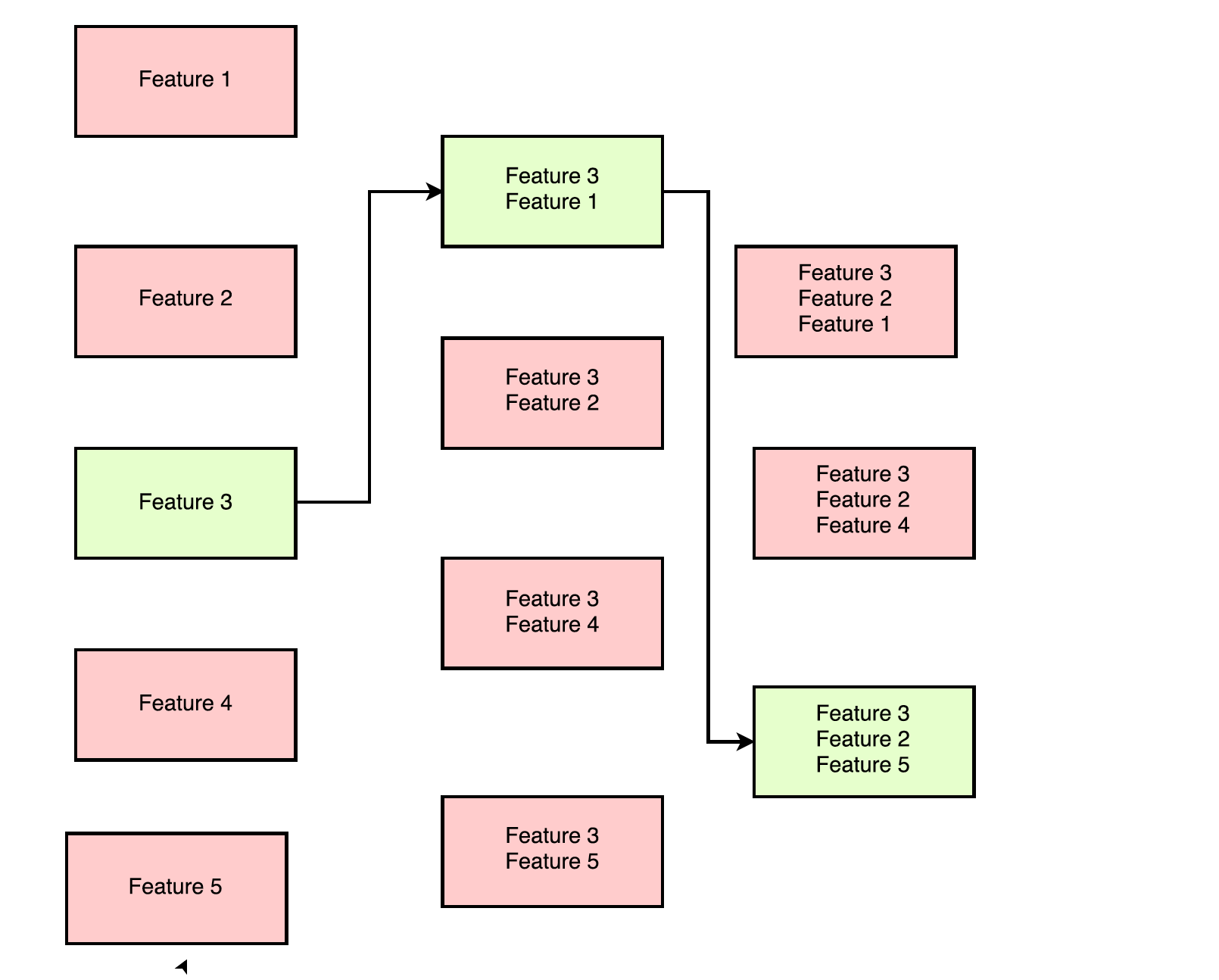
**1.Forward search:** **This method allows you to search for the best feature w.r.t model performance and add them to your feature subset one after the other.**

**For data with n features,**

**->On first round ‘n’ models are created with individual feature and the best predictive feature is selected.**

**->On second round, ‘n-1’ models are created with each feature and the previously selected feature.**

**->This is repeated till a best subset of ‘m’ features are selected.**



**2.Backward feature elimination : Removes 1 feature at a time**

**3.** **Exhaustive feature search : Searches across all possible feature combinations**

* **Computationally expensive than filter methods**
* **Better predictive accuracy than filter methods**

**EMBEDDED METHODS:**

* **Perform feature selection during the modelling algorithm's execution.**
* **Faster than wrapper methods**
* **More accurate than filter methods**
* **Detect interactions between variables**
* **Find the feature subset for the algorithm being trained**
* **Select features based on coefficients of linear or logistic regression.**
* **By using permute concept in tree based algorithms**
* **Consider the interaction between features and models**
* **They are less computationally expensive than wrapper methods, because they fit the machine learning model only once.**

**Recursive Feature Elimination:**

**As the name suggests, this method eliminates worst performing features on a particular model one after the other until the best subset of features are known.**

**For data with n features,**

**->On first round ‘n-1’ models are created with combination of all features except one. The least performing feature is removed**

**-> On second round ‘n-2’ models are created by removing another feature.**

**Interpolation Vs Extrapolation:**

* **Interpolation is an estimation of a value within two known values in a sequence of values.**
* **Extrapolation is an estimation of a value based on extending a known sequence of values or facts beyond the area that is certainly known.**

**Approach to handling Imbalanced Datasets:**

* **The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets.**
* **Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.**

**Data Level approach: Resampling Techniques:**

1. **Random Over-Sampling:**

* **Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.**
* **Unlike under sampling this method leads to no information loss.**
* **Outperforms under sampling**
* **It increases the likelihood of overfitting since it replicates the minority class events.**

1. **Random Under-Sampling:**

* **Random Under sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out.**
* **It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.**
* **It can discard potentially useful information which could be important for building rule classifiers.**

1. **Cluster-Based Over Sampling:**

* **In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size.**

1. **Informed Over Sampling: Synthetic Minority Over-Sampling Technique:**

* **This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.**
* **While generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise**
* **SMOTE is not very effective for high dimensional data**
* **Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances**
* **No loss of useful information**

1. **Modified synthetic minority oversampling technique (MSMOTE)**

**Deep Learning:**

* **Perceptron/Neuron:** **The perceptron may eventually be able to learn, make decisions, and translate languages.**
* **Universal Approximation Theorem: A multilayered network of neurons with a single hidden layer can be used to approximate any continuous function to any desired precision.**
* **Sequences: They are everywhere, Time series, speech, music, text, video, each unit in the sequence interacts with other units Need models to capture this interaction**
* **dendrite: receives signals from other neurons**
* **synapse: point of connection to other neurons**
* **soma: processes the information**
* **axon: transmits the output of this neuron**
* **Neural networks capture information as hierarchal manner.**
* **A single McCulloch Pitts Neuron can be used to represent Boolean functions which are linearly separable**
* **The thresholding logic used by a perceptron is very harsh!**
* **perceptron function behaves like a step function**
* **A step function Not smooth, not continuous (at w0), not diﬀerentiable.**
* **For most real world applications, we would expect a smoother decision function which gradually changes from 0 to 1**
* **sigmoid neurons where the output function is much smoother than the step function called the logistic function**
* **sigmoid is Smooth, continuous, diﬀerentiable.**
* **Gradient tells us the responsibility of parameter towards the loss.**
* **Gradient Descent Rule: The direction u that we intend to move in should be at 180° w.r.t. the gradient In other words, move in a direction opposite to the gradient**
* **The gradient w.r.t. a parameter is proportional to the input.**
* **The learning rate is a hyper parameter that controls how much to change the model in response to the estimated error each time the model weights are updated.**
* **Backpropagation is for calculating the gradients efficiently, while optimizers is for training the neural network, using the gradients computed with backpropagation.**
* **Quantities of interest:**

**Gradient w.r.t. output units**

**Gradient w.r.t. hidden units**

**Gradient w.r.t. weights and biases**

* **Logistic function:**

**g(z) = σ(z)**

**g’(z) = g(z)(1−g(z))**

* **Tanh:**

**g(z) = tanh(z)**

**g’(z) = 1−(g(z))^2**

* **When the curve is steep the gradient (∆y1 /∆x1 ) is large**
* **When the curve is gentle the gradient (∆y2/ ∆x2 ) is small**
* **Recall that our weight updates are proportional to the gradient w = w− η∇w**
* **Hence in the areas where the curve is gentle the updates are small whereas in the areas where the curve is steep the updates are large**
* **In 3d graph A small distance between the contours indicates a steep slope along that direction A large distance between the contours indicates a gentle slope along that direction.**
* **Gradient descent takes a lot of time to navigate regions having a gentle slope.**
* **If I am repeatedly being asked to move in the same direction then I should probably gain some conﬁdence and start taking bigger steps in that direction is momentum based gradient descent.**
* **Update rule for momentum based gradient descent**

**Update t = γ ·update t−1 + η∇wt**

**wt+1 = wt –update t**

* **In addition to the current update, also look at the history of updates.**
* **Even in the regions having gentle slopes, momentum based gradient descent is able to take large steps because the momentum carries it along.**
* **Momentum based gradient descent oscillates in and out of the minima valley as the momentum carries it out of the valley.Takes a lot of u-turns before ﬁnally converging. Despite these u-turns it still converges faster than vanilla gradient descent.**

**Nesterov Accelerated Gradient Descent:** **Can we do something to reduce these oscillations ?.**

* **Why not calculate the gradient (∇wlook ahead) at this partially updated value of w (wlook ahead = wt−γ·updatet−1) instead of calculating it using the current value wt.**

**Update rule for NAG**

**wlook ahead = wt −γ ·updatet−1**

**updatet = γ ·updatet−1 + η∇wlook ahead**

**wt+1 = wt –updatet**

* **Hence the oscillations are smaller and the chances of escaping the minima valley also smaller**

**Stochastic And Mini-Batch Gradient Descent**

* **We see many oscillations. Why ? Because we are making greedy decisions.**
* **Each point is trying to push the parameters in a direction most favorable to it (without being aware of how this affects other points)**
* **A parameter update which is locally favorable to one point may harm other points (its almost as if the data points are competing with each other)**
* **Can we reduce the oscillations by improving our stochastic estimates of the gradient**
* **Notice that the algorithm updates the parameters after it sees mini batch size number of data points.**
* **1 epoch = one pass over the entire data 1 step = one update of the parameters**

**Tips for initial learning rate ?:**

* **Tune learning rate [Try diﬀerent values on a log scale: 0.0001, 0.001, 0.01, 0.1. 1.0]**
* **Run a few epochs with each of these and ﬁgure out a learning rate which works best Now do a ﬁner search around this value [for example, if the best learning rate was 0.1 then now try some values around it: 0.05, 0.2, 0.3] Disclaimer: these are just heuristics ... no clear winner strategy**
* **Step Decay: Halve the learning rate after every 5 epochs or Halve the learning rate after an epoch if the validation error is more than what it was at the end of the previous epoch**
* **Exponential Decay: η = η−kt 0 where η0 and k are hyper parameters and t is the step number**
* **1/t Decay: η = η0 1+kt where η0 and k are hyper parameters and t is the step number**
* **line search :In practice, often a line search is done to ﬁnd a relatively better value of η Update w using diﬀerent values of η Now retain that updated value of w which gives the lowest loss Esentially at each step we are trying to use the best η value from the available choices.**

**Gradient Descent with Adaptive Learning Rate:**

* **If there are n points, we can just sum the gradients over all the n points to get the total gradient**
* **What happens if the feature x2 is very sparse? (i.e., if its value is 0 for most inputs)**
* **∇w2 will be 0 for most inputs (see formula) and hence w2 will not get enough updates**
* **If x2 happens to be sparse as well as important we would want to take the updates to w2 more seriously**
* **Intuition: Decay the learning rate for parameters in proportion to their update history** **(more updates means more decay)**
* **By using a parameter speciﬁc learning rate it ensures that despite sparsity w gets a higher learning rate and hence larger updates**
* **Update rule for Adagrad**

**vt = vt−1 + (∇wt)2 wt+1 = wt − η √vt + ∗∇wt**

* **Adagrad decays the learning rate very aggressively (as the denominator grows)**
* **As a result after a while the frequent parameters will start receiving very small updates because of the decayed learning rate**
* **To avoid this why not decay the denominator and prevent its rapid growth**

**Update rule for RMSProp**

**vt = β ∗vt−1 + (1−β)(∇wt)2 wt+1 = wt − η √vt + ∗∇wt**

* **RMSProp+** **cumulative history of the gradients ==🡺** **Adam**
* **Update rule for Adam**

**mt = β1 ∗mt−1 + (1−β1)∗∇wt**

**vt = β2 ∗vt−1 + (1−β2)∗(∇wt)2**

**ˆ mt = mt /1−βt 1**

**ˆ vt = vt /1−βt 2**

**wt+1 = wt − η √ˆ vt + ∗ ˆ mt**

* **The eigenvectors of a matrix A ∈Rn×n having distinct eigenvalues are linearly independent.**
* **The eigenvectors of a square symmetric matrix are orthogonal.**
* **We have seen 3 diﬀerent interpretations of PCA**

**It ensures that the covariance between the new dimensions is minimized**

**It picks up dimensions such that the data exhibits a high variance across these dimensions**

**It ensures that the data can be represented using less number of dimensions**

* **Simple models trained on diﬀerent samples of the data do not diﬀer much from each other.** **However they are very far from the true sinusoidal curve (under ﬁtting)**
* **On the other hand, complex models trained on diﬀerent samples of the data are very diﬀerent from each other (high variance)**
* **Simple model: high bias, low variance**
* **Complex model: low bias, high variance**
* **There is always a trade-oﬀ between the bias and variance**

**Diﬀerent forms of regularization :**

1. **l2 regularization**
2. **Dataset augmentation**
3. **Parameter Sharing and tying**
4. **Adding Noise to the inputs**
5. **Adding Noise to the outputs**
6. **Early stopping**
7. **Ensemble methods**
8. **Dropout**

**\*L1 and L2 regularizations are constraints the weights to certain regions of the parameter space.**

**\*L-1 regularization: constrains most weights to be 0**

**\*L-2 regularization: prevents most weights from taking large values.**

**\*Dropout is prevent the neural network from over fitting ,it is equivalent to the ensembling.**

**\*The main objective of unsupervised pretraining is constrains the weights to lie in regions where the characteristics of the data are captured well.**

**Activation Functions:**

* **Activation function decides, whether a neuron should be activated or not by calculating weighted sum and further adding bias with it.**
* **A neural network without an activation function is essentially just a linear regression model.**
* **In a neural network, we would update the weights and biases of the neurons on the basis of the error at the output. This process is known as back-propagation.**

**\*Sigmoid:**

**1.sigmoid function compresses all its inputs to the range [0,1]**

**Sigma(X) = 1/(1+e^X)**

**2.** **Sigmoids are not zero centered(Essentially, either all the gradients at a layer are positive or all the gradients at a layer are negative)**

**3.** **Sigmoids are computationally expensive (because of exp (x))**

**4. The gradient vanishes at saturation.**

**\*Tanhx:**

**1.Compress all inputs to the range [-1,1]**

**2.Zero centered.**

**3.** **Also computationally expensive.**

**4.** **The gradient still vanishes at saturation.**

**tanh(x) = sinh(x)/cosh(x) = ( ex - e-x )/( ex + e-x )**

**\*Relu:**

**f(X) = max(0,X)**

**\*Does not saturate in the positive region.**

**\*Computationally efficient**

**\*In practice converges much faster than sigmoid/tanhx**

**\*No saturation in leaky relu.**

**\*Relu create dead environment**

**Parametric ReLU**

**f(x) = max(ax; x)**

**a is a parameter of the model will get updated during backpropagation.**

**\*ELU (Exponential Linear Unit):**

**f(x) = x if x > 0**

**= aex -1 if x <= 0**

**1.Expensive**

**Deep Learning has evolved :**

1. **Better optimization algorithms**
2. **Better regularization methods**
3. **Better activation functions**
4. **Better weight initialization strategies**

* **If all the weights in a network are initialized to the same value ,then all weights will get the same update and remain equal(Symmetry breaking problem.**
* **Let’s try to initialize the weights to small random numbers, all the activations in a layer are very close to 0 then**  **gradient of the weights connecting this layer to the next layer will be close to zero(vanishing gradient problem)**
* **Let us try to initialize the weights to large random numbers,** **Most activations have saturated, gradients will all be close to 0 (vanishing gradient problem).**

**Drawbacks of ANN & CNN:**

**Outputs are independent of previous inputs**

**Input is of a fixed length**

**RNN:**

**\*** **In feedforward and convolutional neural networks the size of the input was always fixed.**

**\* And also each input to the network was independent of the previous or future inputs.**

**\*** **Make sure that the function executed at each time step is the same.Sharing parameters at every time step**

**\* Rnn have recurrent connections between time steps which account for dependence between inputs.**

**\*** **The state (*si*) of an RNN records information from all previous time steps.**

**\*** **At each new timestep the old information gets morphed by the current input.**

**\*** **One could imagine that after *t* steps the information stored at time step *t-k* (for some *k < t*) gets completely morphed so that it would be impossible to extract the original information**

**stored at time step *t – k***

**\*** **A similar problem occurs when the information flows backwards(backpropagation)**

**\*** **It is very hard to assign the responsibility of the error caused at time step *t* to the events that occurred at time step *t – k***

***\**Padding was only ensures that input matrix is of uniform size**

**\* Computations are done only for the required number of time steps.**

**LSTM:**

**Three Gates :**

**1.Selective Write (output gate ):What fraction of previous state information send to next state .**

**2.Selective Read (Input gate) : What fraction of information need to capture from current state .**

**3.Selective Forget(Forget gate):What fraction of information need to remove form previous state.**

\***In LSTM backward propagation will control the how much amount of gradient flows backwards.**

**GRU:**

**Two gates:**

**1.Selective write**

**2.Selective Read**

**\*\*\*Attention Mechanism:**

**\*Essentially at each time step we come up with a distribution on the input words.** **This distribution tells us how much attention to pay to each input words at each time step.**

**\*Ideally, at each time-step we should feed only this relevant information (i.e. encodings of relevant words) to the decoder.**

**\*Intuitively this should work better because we are not overloading the decoder with irrelevant information(about words that do not matter at this time step).**

**Word Embedding:**

**One -hot Encoding:**

**\*** **Euclidean distance between any two words in the vocabulary in 2^1/2.**

**\*** **Cosine similarity between any two words in the vocabulary is 0**

**\*** **Count based methods (SVD) rely on global co-occurrence counts from the corpus for computing word representations.**

**1.Does not capture semantic relation ship between words**

**2.Matrix is too sparse**

**3.Wont consider the ordering of the words**

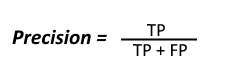
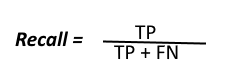
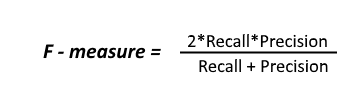
**\*** **Predict based methods learn word representations using co-occurrence information.**

**\* We are injecting noise into the model so that we need call it as negative sampling.**

**Confusion Matrix:**

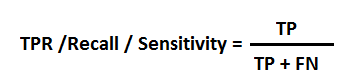
* **Confusion matrix is table that is often used to describe the performance of the classification model.**
* **A confusion matrix is a summary of prediction results on a classification problem.**
* **Positive (P) : Observation is positive**
* **Negative (N) : Observation is not positive**
* **True Positive (TP) : Observation is positive, and is predicted to be positive.**
* **False Negative (FN) : Observation is positive, but is predicted negative.**
* **True Negative (TN) : Observation is negative, and is predicted to be negative.**
* **False Positive (FP) : Observation is negative, but is predicted positive.**
* **Accuracy or classification rate**



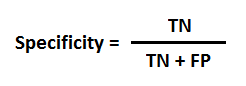
* **Recall: Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).**
* Precision: **To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP).** 
* **High recall, low precision**: **This means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.**
* **Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)**
* **F-measure:  
  Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more.  
  The F-Measure will always be nearer to the smaller value of Precision or Recall.  
  **
* **Based on the application only we are giving priority to the either FP or FN.**
* **The F-Measure will give equal importance to FP and FP.**
* **The F-beta score is the weighted harmonic mean of precision and recall, reaching its optimal value at 1 and its worst value at 0.**
* **So the F-Beta Score is defined as:**
* Fβ=(1+β^2) \* (precision \*recall) /((β^2⋅precision) +recall)
* **The beta parameter determines the weight of recall in the combined score. beta < 1 lends more weight to precision, while beta > 1 favors recall (beta -> 0 considers only precision, beta -> +inf only recall).**
* **Beta value is depends on type of business problem.**

## **TPR (True Positive Rate) / Recall /Sensitivity**

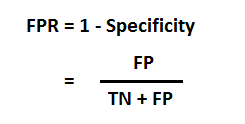
* **Precision is more important than recall when you would like to have less False Positives in trade off to have more False Negatives. Meaning, getting a False Positive is very costly, and a False Negative is not as much.**

****

## **Specificity**

****

## **FPR**



* **Precision: How many selected items are relevant**
* **Recall:How many relevant items are selected**

**Q) How to calculate confusion metric for multi-class classification ?**

True Positive Rate(All classes): sum of all diagonal elements

False positive rate(Type-I Error) (Each class): sum of column x (without main diagonal), sum(cm(:, x))-cm(x, x)

False negative rate (Type-II Error) (Each class): sum of row x (without main diagonal), sum(cm(x, :), 2)-cm(x, x)

True negative rate (Each Class): confusion\_matrix.values.sum() - (FP + FN + TP)

**Q)What is the advantage of hinge loss?**

**Hinge loss will penalizes incorrectly classified examples and also correctly classified examples that lies within margin.**

**Speed over accuracy use hinge loss**

**Accuracy over speed matters go with log loss or cross entropy loss.**

* **Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.**

**Cross-entropy:**

* **Cross-entropy is commonly used to quantify the difference between two probability distributions.**
* **Sum of multiplication of probability associated with random variable and value associated with random variable.**

**Expected value = sum(p(i) \* v(i))**

* **Information of an event is function of probability of an event .**

**IC(X=S) = log (1/ p(X=S))**

* **Entropy is expected information content of a random variable.**

**H(x)=-Sum(p(x)log2(p(x))**

* **No of bits required to transmit any message is equal to entropy of message.**
* **Information gain measures how well a given attribute separates the training examples according to their target classification.**

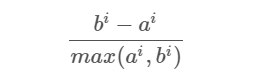
**WORD MOVER DISTANCE :**

* **The assumption is that similar words should have similar vectors.**
* **WMD uses the word embedding’s of the words in two texts to measures the minimum distance that words in one text need to travel in semantic space to reach the words in the other text.**
* **WMD use word embedding to calculate the distance between two texts even though there is no common word.**
* **Make the distribution of one document to look like distribution of another document .**

**Silhouette Analysis:**

**\*Silhouette analysis can be used to determine the degree of separation between clusters. For each sample:**

* **Compute the average distance from all data points in the same cluster (ai).**
* **Compute the average distance from all data points in the closest cluster (bi).**
* **Compute the coefficient:**

****

**The coefficient can take values in the interval [-1, 1].**

* **If it is 0 –> the sample is very close to the neighboring clusters.**
* **It it is 1 –> the sample is far away from the neighboring clusters.**
* **It it is -1 –> the sample is assigned to the wrong clusters.**

**Edit distance :**

**How many operations would it take to convert one string to another string.**

**Operations:**

**1.Insertion**

**2.Deletion**

**3.Substitution**

**Syntax: It will refers to the way words arranged together in a context and relationship between them.**

**\*word net is hierarchically organized lexical database, Synset is set of synonymous representing a sense.**

**Path based similarity: \*two words are similar if they are near by in the hypernym graph.**

**Pathlength(c1,c2)=number of edges in shortest path (In hypernym graph) between senses c1 and c2.**

**Simpath(c1,c2) = 1/(1+pathlen(c1,c2)**

**Sim(w1,w2)=max(sim(c1i,c2i)**

**Leacock-chodorow (L-C) similarity :**

**simLC(c1,c2) = -log(pathlen(c1,c2))/2d**

**d=maximum depth of the hierarchy**

**Bias-Variance Tradeoff:**

* **Total error is composed of Bias, Variance and a Random irreducible error. Bias and Variance can be managed.**
* **If the model performance on training and testing data sets is inconsistent, it indicates a problem either with Bias or Variance.**
* **Bias arises when you make assumptions preventing you from finding relevant relationships between inputs (independent variables) and outputs (dependent variable). This causes the model to under fit the data.**
* **Variance arises due to the model being overly sensitive to small fluctuations in the training data. Such a model over fits the data.**
* **Bias is the accuracy of our predictions.**
* **Variance is the difference between many model’s predictions. Or Variation of different model predictions.**

**Ways to Decrease Bias and Variance:**

* **Outliers and Influential Observations can cause statistical bias. Bias cannot be increased by increasing training sample size**
* **Adding Features (Independent variables or Predictors) tends to decrease bias**
* **Variance and standard error are can be minimized by increasing training sample size**
* **Dimensionality reduction and feature selection methods reduce variance**

**Parameters can be tuned in supervised models to control bias and variance:**

* **Increasing K in KNN can be decrease the variance at the cost of increasing bias**
* **In CART Increasing length of the tree increases variance. Pruning is used to control variance.**
* **IN DNN addition of hidden units will decrease bias at the cost of increasing variance**
* **Regularization methods will decrease variance at the cost of increasing bias .It can be applied to other techniques (Not only linear models )**
* **Boosting methods combine many “weak” (high bias) models in an ensemble that lowers bias compared to individual models.**
* **Bagging (bootstrap aggregating) techniques combine “strong” models to minimize variance.**

**A good model will have**

**• a small mean of the errors (low bias, i.e., the model accurately captures the behaviour of the data), and**

**• a small standard deviation of the errors (low variance, i.e., error does not vary much based on the choice of the dataset)**

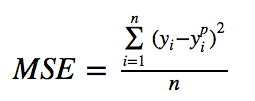
**Gain & Lift Chart:**

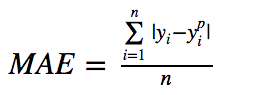
* **Confusion matrix evaluates model on the whole population, gain or lift charts evaluates the model performance in a portion of population.**
* **Lift is dependent on total response rate of the population. Hence, if the response rate of the population changes, the same model will give a different lift chart.**
* **ROC curve on the other hand is almost independent of the response rate.**
* **AUC ROC, it only takes into account the order of probabilities and hence it does not take into account the model’s capability to predict higher probability for samples more likely to be positive.**
* **Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.**
* **A parameter is a variable that is internal to the model and whose value is estimated from the training data. They are often saved as part of the learned model. Examples include weights, biases etc.**
* **A hyper parameter is a variable that is external to the model whose value cannot be estimated from the data. They are often used to estimate model parameters. The choice of parameters is sensitive to implementation. Examples include learning rate, hidden layers etc.**

**Loss Functions:**

* **Loss function is a method of evaluating how well your algorithm models on your dataset.**
* **Finding the best parameters (weights) for your data by using loss function is nothing but a optimization.**

**Regression Losses:**

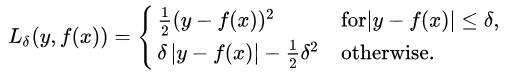
1. **Mean Square Error/Quadratic Loss/L2 Loss:**
   * + **Mean square error is measured as the average of squared difference between predictions and actual observations.**
     + **It’s only concerned with the average magnitude of error irrespective of their direction.**
     + **However, due to squaring, predictions which are far away from actual values are penalized heavily in comparison to less deviated predictions.**
     + **MSE has nice mathematical properties which makes it easier to calculate gradients.**
     + **The range is 0 to ∞.**
     + 
2. **Mean Absolute Error/L1 Loss:**
   * + **Mean absolute error, on the other hand, is measured as the average of sum of absolute differences between predictions and actual observations.**
     + **Like MSE, this as well measures the magnitude of error without considering their direction.**
     + **Unlike MSE, MAE needs more complicated tools such as linear programming to compute the gradients.**
     + **MAE is more robust to outliers since it does not make use of square.**
     + **The range is 0 to ∞.**



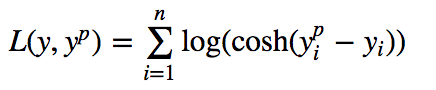
**MSE VS MAE:**

* **Model with MSE loss give more weight to outliers than a model with MAE loss.**
* **The model with RMSE as loss will be adjusted to minimize that single outlier case at the expense of other common examples, which will reduce its overall performance.**
* **MAE loss is useful if the training data is corrupted with outliers.**
* **MAE will give large Gradient even for small loss values.**

1. **Mean Bias Error:**
   * + **This is same as MAE with the only difference that we don’t take absolute values.**
     + **Clearly there’s a need for caution as positive and negative errors could cancel each other out.**
     + **Although less accurate in practice, it could determine if the model has positive bias or negative bias.**
2. **Huber Loss, Smooth Mean Absolute Error:**
   * + **Huber loss is less sensitive to outliers in data than the squared error loss.**
     + **It’s also differentiable at 0. It’s basically absolute error, which becomes quadratic when error is small.**
     + **How small that error has to be to make it quadratic depends on a hyper parameter, 𝛿 (delta), which can be tuned.**
     + **Huber loss approaches MAE when 𝛿 ~ 0 and MSE when 𝛿 ~ ∞ (large numbers.)**
     + **Residuals larger than delta are minimized with L1 (which is less sensitive to large outliers), while residuals smaller than delta are minimized “appropriately” with L2.**
     + **Why use Huber Loss?**

**One big problem with using MAE for training of neural nets is its constantly large gradient, which can lead to missing minima at the end of training using gradient descent. For MSE, gradient decreases as the loss gets close to its minima, making it more precise.**

1. **Log-Cosh Loss:**
   * + **Log-cosh is another function used in regression tasks that’s smoother than L2. Log-cosh is the logarithm of the hyperbolic cosine of the prediction error.**
     + **It is twice differentiable everywhere, unlike Huber loss.**



1. **Quantile Loss:**
   * + **Quantile loss functions turn out to be useful when we are interested in predicting an interval instead of only point predictions.**
     + **Knowing about the range of predictions as opposed to only point estimates can significantly improve decision making processes for many business problems.**

<https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0>

**Classification Losses:**

**1. Exponential loss:**

* **The exponential loss is convex and grows exponentially for negative values which makes it more sensitive to outliers.**
* **The exponential loss is used in the AdaBoost algorithm.**

**2. Logistic loss:**

* **The logistic loss is convex and grows linearly for negative values which make it less sensitive to outliers.**
* **The logistic loss is used in the Logit Boost algorithm.**

**3. Savage loss:**

* **The Savage loss is quasi-convex and is bounded for large negative values which makes it less sensitive to outliers.**
* **The Savage loss has been used in gradient boosting and the Savage Boost algorithm.**

**4. Tangent loss:**

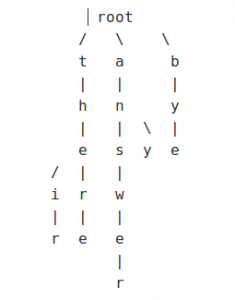
* **The Tangent loss is quasi-convex and is bounded for large negative values which makes it less sensitive to outliers.**
* **Interestingly, the Tangent loss also assigns a bounded penalty to data points that have been classified "too correctly".**
* **This can help prevent over-training on the data set.**
* **The Tangent loss has been used in gradient boosting, the Tangent Boost algorithm and Alternating Decision Forests.**

**5. Hinge loss:**

* **Correctly classified points lying outside the margin boundaries of the support vectors are not penalized, whereas points within the margin boundaries or on the wrong side of the hyperplane are penalized in a linear fashion compared to their distance from the correct boundary.**
* **hinge loss function is both convex and continuous, it is not smooth (is not differentiable) at y f(x) = 1. Consequently, the hinge loss function cannot be used with gradient descent methods or stochastic gradient descent methods which rely on differentiability over the entire domain. However, the hinge loss does have a subgradient at y f(x) = 1, which allows for the utilization of subgradient descent methods.**
* **SVMs utilizing the hinge loss function can also be solved using quadratic programming.**

**Flashtext:**

* **Flashtext is a Python library created specifically for the purpose of searching and replacing words in a document.**
* **It requires word or list words and a string to searching or replacing.**
* **Words are called as keywords.**
* **keywords are stored as a Trie Data Structure which is very efficient at Retrieval tasks.**
* **Both searching and replacing operations happen over a single pass**
* **Trie Data Structure :**



* **An example of single pass replacement looks like this-**

**String = “spamham sha”**

**Replace “spam” with “eggs” and “sha” with “md5”**

**Now let’s see how does the String look like with and without a single pass.**

**Single-pass**

**String = “eggsham md5”**

**Without Single-pass**

**String = “eggmd5m md5”**

**Topic Modeling:**

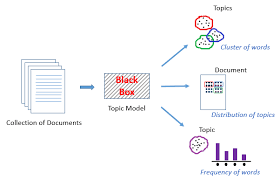
* **Topic modeling is a form of dimensionality reduction. Rather than representing a text in its feature space, we can represent the text in its topic space.**
* **By doing topic modeling we build clusters of words rather than clusters of texts.**
* **Document classification is assigning a single category to a text, topic modeling is assigning multiple tags to a text.**
* **Topic modeling can improve classification by grouping similar words together in topics rather than using each word as a feature.**
* **Topic Modeling Algorithms:**

**1.** **LDA – Latent Dirichlet Allocation –Its foundations are Probabilistic Graphical Models**

**2.LSA or LSI – Latent Semantic Analysis or Latent Semantic Indexing – Uses Singular Value Decomposition (SVD) on the Document-Term Matrix. Based on Linear Algebra----Matrix Factorization approach.**

**3.NMF – Non-Negative Matrix Factorization – Based on Linear Algebra**

* **All of them output 2 matrices: WTM (Word Topic Matrix) and TDM (Topic Document Matrix). The matrices are significantly smaller and the result of their multiplication should be as close as possible to the original DWM matrix.**



**Singular Value Decomposition:**

* **SVD is method to factorize a rectangular square matrix into three matrices.**

**A*nxp*= U*nxn* S*nxp* VT*pxp***

**U = Left singular matrix**

**V = Right Singular matrix**

**S = Singular Values (Singular matrix)**

* **Row Vectors of U form an orthogonal set**
* **Row vectors of VT form an orthogonal set**
* **Singular matrix is diagonal matrix those singular values are arranged in the descending order.**
* **Patterns among the terms are captured by the left singular matrix**
* **Patterns among the documents are captured by right singular matrix**

**How LDA Works?**

* **Choose number of topics**
* **In the initialization stage, each word in the document is assigned to a random topic.**
* **The above step will gives the topic representations of all documents and word distributions of all the topics.**
* **Iteratively, the algorithm goes through each word and reassigns the word to a topic taking into consideration:**
* **What’s the probability of the word belonging to a topic**
* **What’s the probability of the document to be generated by a topic**
* **Vector which shrinks or stretches on a transformation without rotating is commonly known as Eigen Vector.**
* **Cosine similarity is a distance metric which is used when the magnitude of vectors does not matter, while their directions do. This metric is typically used in text data represented by word frequency or tf-idf. We assume that if one word, for example, panther, occurs more frequently in document 1 than it occurs in document 2, document 1 is more related to animals. However, the documents can be of uneven lengths. Then, there can be a possibility that 'panther' occurred more number of times in document 1 just because it was way longer than document 2. Cosine similarity corrects for this. Euclidean distance, however, would just calculate the squared distance of magnitude, and hence would not capture the direction, or sentiment, as such. If the cosine similarity is 1, this means that the angle between two document vectors is zero, which in turn means that both vectors are pointing in the same direction, but are of different magnitudes due to different document lengths. Similarly, if it is 0, it means that the documents are unrelated as document vectors in this case would be orthogonal.**
* **Multivariate regression refers to regressing the dependent variable on more than one independent variables, while multinomial regression refers to creating a regression model where there are more than two categories in the dependent nominal variable.**

**Anomaly Detection Methods:**

* **Density based: DBSCAN, LOF**
* **Distance Based: K-Means, KNN, Regression Hyper plane distance**
* **Parametric: Single Class SVM,GMM,Extreme Value theory**

**Isolation Forest:**

* **Isolation forest can works as supervised and unsupervised classifier**
* **It calculates anomaly score = 2^ -(E(h(x))/c(n))**

**Where h(x) is number of edges for a point x**

**C(n) is normalized constant for a dataset of a size n**

**H(i) is the harmonic number and it can be estimated by ln(i) + 0.5772156649 (Euler’s constant)**

**c(n) is the average of h(x) given n**

* **The anomaly score is a function of the number of splits**
* **Anomalies will be isolated in only a few steps but nominal points more**

**Angle-Based Outlier Detection (ABOD):**

* **It considers the relationship between each point and its neighbor(s). It does not consider the relationships among these neighbors. The variance of its weighted cosine scores to all neighbors could be viewed as the outlying score.**
* **ABOD performs well on multi-dimensional data.**
* **PyOD provides two different versions of ABOD:**

**1.Fast ABOD: Uses k-nearest neighbors to approximate**

**2.Original ABOD: Considers all training points with high-time complexity**

**k-Nearest Neighbors Detector:**

* **For any data point, the distance to its kth nearest neighbor could be viewed as the outlying score.**
* **PyOD supports three kNN detectors:**

**1.** **Largest: Uses the distance of the kth neighbor as the outlier score**

**2. Mean: Uses the average of all k neighbors as the outlier score**

**3. Median: Uses the median of the distance to k neighbors as the outlier score**

**Local Outlier Factor:**

* **Local Outlier Factor (LOF) is a score that tells how likely a certain data point is an outlier/anomaly.**
* **Local outlier factor is a density-based method that relies on k-nearest neighbors.**
* **The LOF method scores each data point by computing the ratio of the average of densities of the neighbors to the density of point itself.**
* **Basically it compares the local density of a point with local density of its k nearest neighbors.**

<https://www.youtube.com/watch?v=vnoBkTa7arI>

* **LRD is the most optimal distance in any direction from the neighbor to the individual point**

**Generative Vs Discriminative**

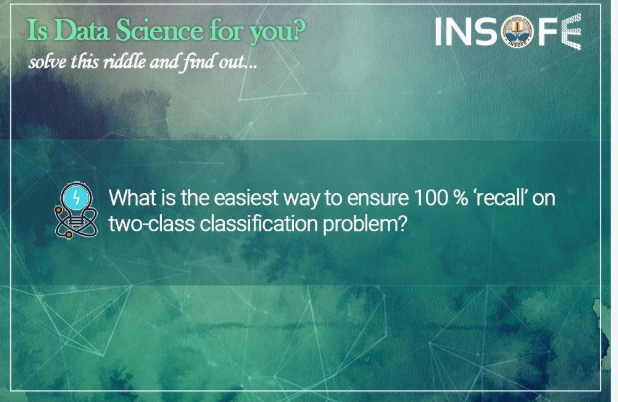
* **In General, A Discriminative model ‌models the decision boundary between the classes. A Generative Model ‌explicitly models the actual distribution of each class. In final both of them is predicting the conditional probability P(Animal | Features). But Both models learn different probabilities.**
* **Assume some functional form for P(Y|X).**
* **Estimate parameters of P(Y|X) directly from training data.**
* **A Generative Model ‌learns the joint probability distribution p(x,y). It predicts the conditional probability with the help of Bayes Theorem. A Discriminative model ‌learns the conditional probability distribution p(y|x).**
* **Assume some functional form for P(Y), P(X|Y)**
* **Estimate parameters of P(X|Y), P(Y) directly from training data**
* **Use Bayes rule to calculate P(Y |X)**

**Generative classifiers**

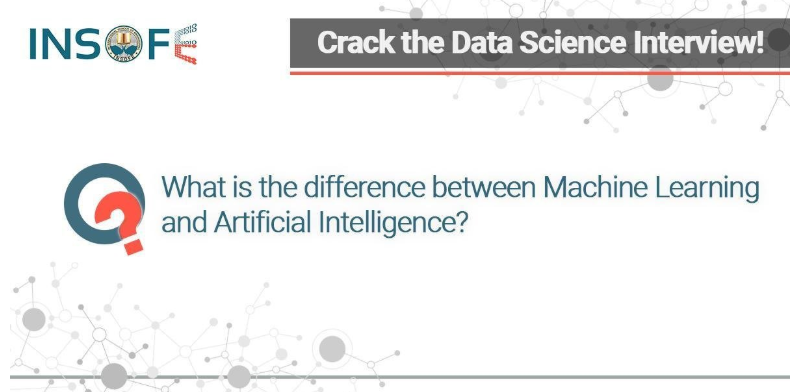
1. **‌Naïve Bayes**
2. **Bayesian networks**
3. **Markov random fields**
4. **‌Hidden Markov Models (HMM)**

**Discriminative Classifiers**

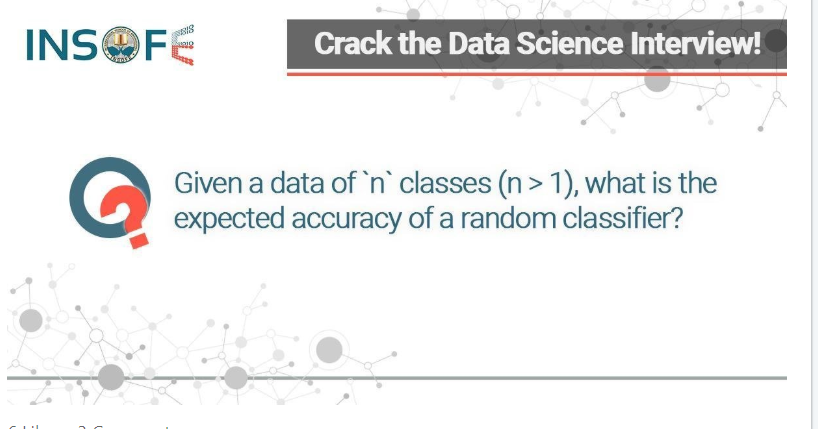
1. **‌Logistic regression**
2. **Scalar Vector Machine**
3. **‌Traditional neural networks**
4. **‌Nearest neighbor**
5. **Conditional Random Fields (CRF)s**



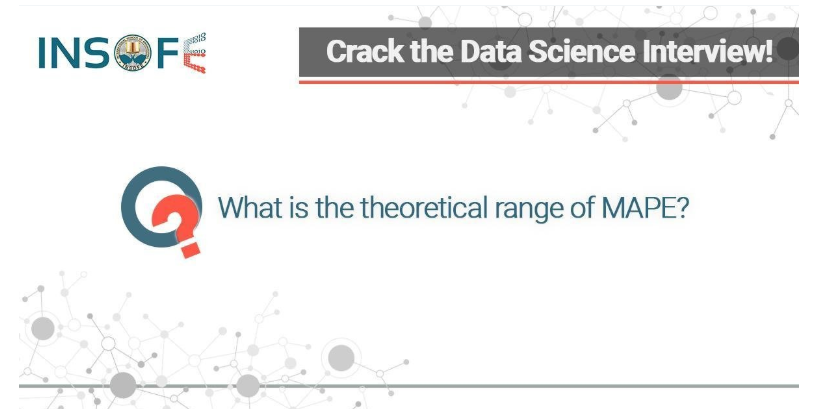
Answer: By predicting every single response as ‘positive’



Answer: Focus of machine learning is to build mathematical models, which can approximate the behavior of given data into a (sometimes complex) formula. This formula gives out an expression (usually this is called as the target) using the data as input to the formula AI on the other hand contains disciplines which focus on mimicking human intelligence as closely as possible (e.g. Siri, Sophia). ML is an application of AI.



Answer: The expected accuracy of a random classifier will be equal to the expectation of the Probability Mass Function (PMF) of the target class. In the special case of a uniformly distributed target, the expected accuracy will be '1/n'.



Answer: Theoretical range of MAPE is the set of all non-negative rational numbers.

**Topic: Normalization in Deep learning:**

**Types in normalization:**

**1.Batch Normalization**

**2.Weight Normalization**

**3.Layer Normalization**

**Batch normalization normalizes the input features across the batch dimension. The key feature of layer normalization is that it normalizes the inputs across the features**.

**\*\*\*\*Cause for Repetition in Decoder phase:**

**@The decoder state is pretty similar at two consecutive time steps**

**@The context vector is being fed to the decoder at different time steps are similar**

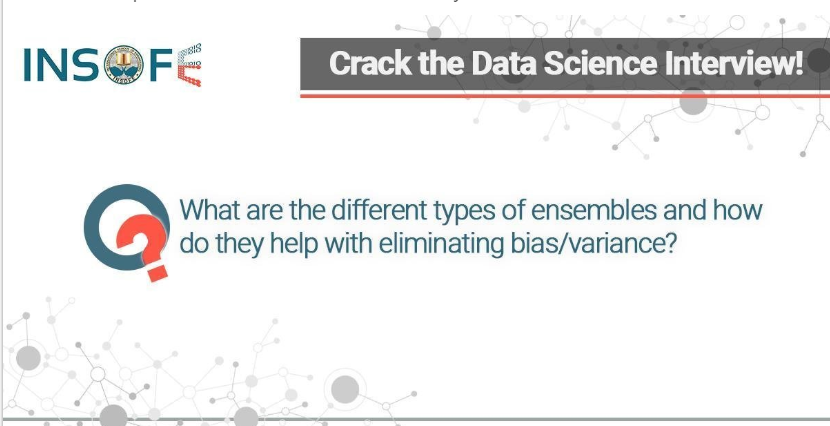
**@some times it will produce factual results inaccurately**

**@Repetition will happen**

**Coverage Mechanism: Penalize the repeatedly attending to the same parts in source text**

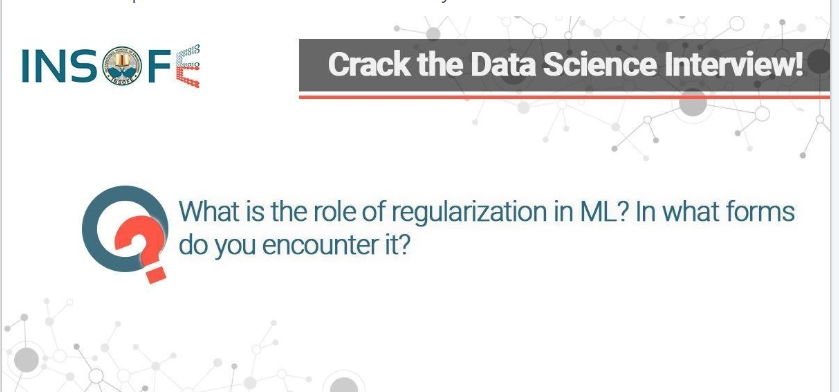
**Rouge Score:** ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation. It works by comparing an **automatically produced summary** or **translation** against a set of **reference summaries** (typically human-produced).

Recall in the context of ROUGE means how much of the reference summary is the system summary recovering or capturing



**Answer:** Stacking - reduces variance across multiple models predictions Bagging - reduces variance across the dataset by modeling on different parts of the data Boosting - reduces bias by systematically reducing the error

**Looking For Opportunity In AI./Data Scientist/ Deep Learning Er. / NLP Enthusiast @ Accenture Innovation Hub (RnD unit)**



**Answer:** Regularization is used to avoid over-fitting. The cost function to be minimized is a sum of errors and the model complexity. By doing this model is more generalized and avoids variance.

BLEU: Bilingual evaluation understudy

BP: Brevity penalty {1 if MT length = Reference Length else e(1-MT Length/Reference Length)

I suggest using [gensim](https://radimrehurek.com/gensim/) to do everything. You can read the file, and also benefit from having a lot of methods already implemented on this great package.

Suppose you generated [GloVe](https://nlp.stanford.edu/projects/glove/) vectors using the C++ program and that your "-save-file" parameter is "vectors". Glove executable will generate you two files, "vectors.bin" and "vectors.txt".

Use [glove2word2vec](https://radimrehurek.com/gensim/scripts/glove2word2vec.html) to convert GloVe vectors in text format into the word2vec text format:

from gensim.scripts.glove2word2vec import glove2word2vec

glove2word2vec(glove\_input\_file="vectors.txt", word2vec\_output\_file="gensim\_glove\_vectors.txt")

Finally, read the word2vec txt to a gensim model using [KeyedVectors](https://radimrehurek.com/gensim/models/keyedvectors.html):

from gensim.models.keyedvectors import KeyedVectors

glove\_model = KeyedVectors.load\_word2vec\_format("gensim\_glove\_vectors.txt", binary=False)

\*The character-level embeddings are generated by taking the final hidden states of a bi-directional recurrent neural network (RNN) applied to embeddings of characters in the token. Such character-level embeddings have been shown to be helpful to deal with out-of-vocab (OOV) tokens.

\*use pointer networks (Vinyals et al., 2015) to predict the start and end position of the answer.

**Pointer network:** instead of using attention to blend hidden units of an encoder to a context vector at each decoder step, it uses attention as a pointer to select a member of the input sequence as the output.

. Ptr-Nets not only improve over sequence-to-sequence with input attention, but also allow us to generalize to variable size output dictionaries.

name Pointer Networks (Ptr-Nets), can be trained to output satisfactory solutions to three combinatorial optimization problems – computing planar convex hulls, Delaunay triangulations and the symmetric planar Travelling Salesman Problem (TSP).

EM=82.3% and F1=91.2%

**Textual entailment** (**TE**) in [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) is a directional relation between text fragments. The relation holds whenever the truth of one text fragment follows from another text

**Pointer Generation Networks**:

Text summarization models have two problems1.Inaccurate Factual Details

2.Repetition of the words.

Two parts

1. hybrid pointer-generator network

that can copy words from the source text

via pointing, which aids accurate reproduction

of information

2. Coverage eliminates

Repetition

CNN is works based on tiling, tiling is nothing but a connectivity pattern between neurons inspired by connectivity pattern in animal visual cortex the individual neurons are arranged in such way that they responds to the overlapping region of the visual fields.

**Q)What is difference between time series dataset and normal machine learning datasets ?**

Time series adds an explicit order dependence between observations: a time dimension.

**Q)What is extrapolation?**

Making predictions about the future is called extrapolation.

**Q) What is Supervised learning ?**

Supervised learning is where you have input variables (X) and an output variable (y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

**Q)What is NLP?**

Statistics allows you to summarize data and shapes how we make decisions.

**Q) ANN VS RNN?**

In RNN the hidden state of each time step is fed to the next time step thereby allowing interactions between time steps in the sequence.

Advantage of padding is kernel is centered on each pixel of image

Text summarization is summarize the entire text into some reasonable extent

IE is nothing but a acquire structured information knowledge from unstructured text

**Text Summarization:** Text summarization is summarize the entire text into some reasonable extent.

Two types 1. Extractive

1. Abstractive

**Extarctive:**

**Steps:**

**1.Choose sentences to extract from the document**

**2.Chooose order to place sentences in the summary**

**3.Simplify the sentences**

**4.Increse diversification by removing redundant sentences**

**step -1: Unsupervised content collection:**

**1.give weight to each word in the document by using tf-idf**

**2.caluculate weight for sentences by using below formula**

**W(s) = sum(weights of all words in the sentences in the sentence)/length(sentence)**

**Lex rank Algorithm (Graph based method)**

1.find similarity matrix for all sentences

2.Convert similarity matrix into row stochastic matrix (It will give probability scores)

3.find informative score for each sentence by using PageRank algorithm

I(j)= u \*∑I(j)M(k,j)+(1-u)÷|S|

K!=j

Removing redundancy by using Maximal marginal relevance

\*\* We can also find out informative score or relavance by using how many named entities are involved in sentnces

**Q)What is the optimization function in Extractive summarization /**

Optimaztion function = Sum of selected sentences scores of document – Sum of redudency score between all sentnces

**Evalution fo summarization:**

**1.Rouge Evaluation**

**Text Classification:**

**Rule based clasiification:**Accuracy can be high if rules carefully refined by expert,but building and maintaining these rules is expensive

**Naïve Assumption: Each** feature is conditionally independent of each other

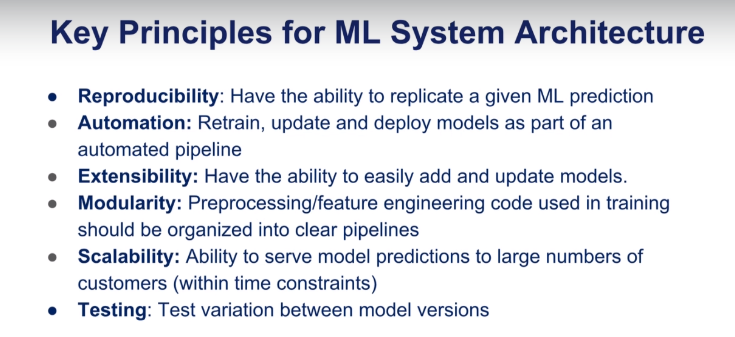
\*computing probability of sentence or sequence of words or probability of next word in the sentence is called as **language modeling.**

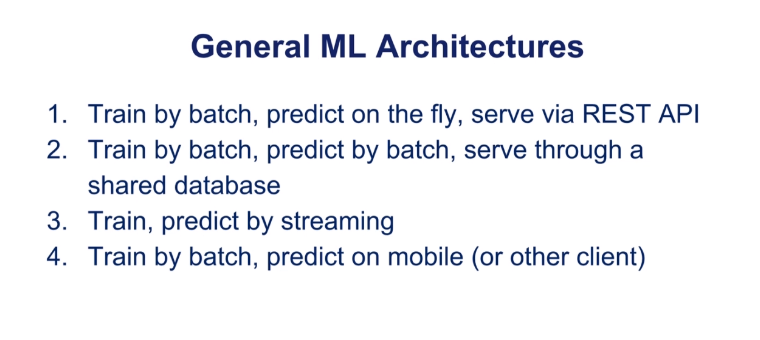
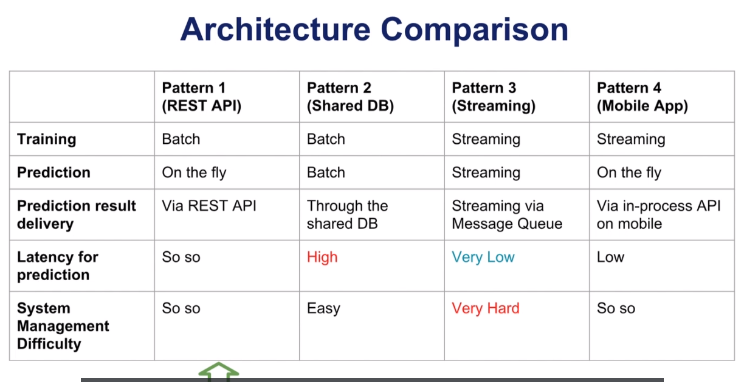
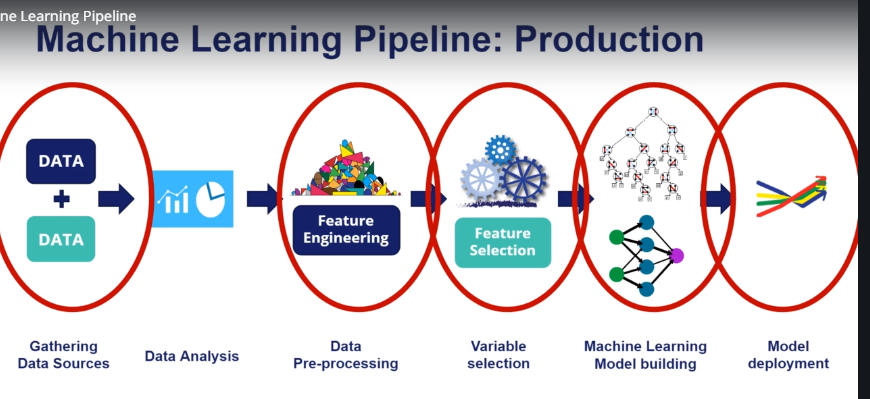
\***Perplexity** is the inverse probability of test data, normalized by the number of words

**Pyforest:**

**Pyforest-lazy is an open source library that imports all popular** [**hashtag#Python**](https://www.linkedin.com/feed/hashtag/?highlightedUpdateUrns=urn%3Ali%3Aactivity%3A6567249753204187136&keywords=%23Python&originTrackingId=vxBBefzvT8qd%2FQJDZhVjgQ%3D%3D) **Data Science libraries so that they are always there when you need them. If you don't use a library, it won't be imported. This is all done with a single line of code: from pyforest import \* And if you use Jupyter or IPython, you can even skip this line because pyforest adds itself to the auto start. Installation: pip install pyforest GitHub repository:** [**https://lnkd.in/f44-hSA**](https://lnkd.in/f44-hSA) **pyforest has been developed by** [**Tobias Krabel**](https://www.linkedin.com/in/ACoAAByOASQBr25JblA2Pnl1FPBYE-bP_VH-ZsQ/)**,** [**Florian**](https://www.linkedin.com/in/ACoAABXQGNQBN7uSP828oyfJAuiexTVJzJVKHEg/) **and Guido from 8080 Labs.**

* Model Deployment is the process of integrating a machine learning model into an existing product environment so that we can use it to make business decision based on live or feature data.



* 
* 
* 

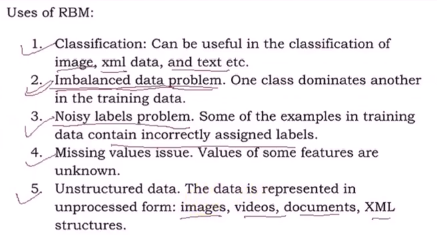
Predicting High Priority Tickets: To predict priority 1 & 2 tickets, so

that they can take preventive measures or fix the problem before it surfaces.

**RBM:**

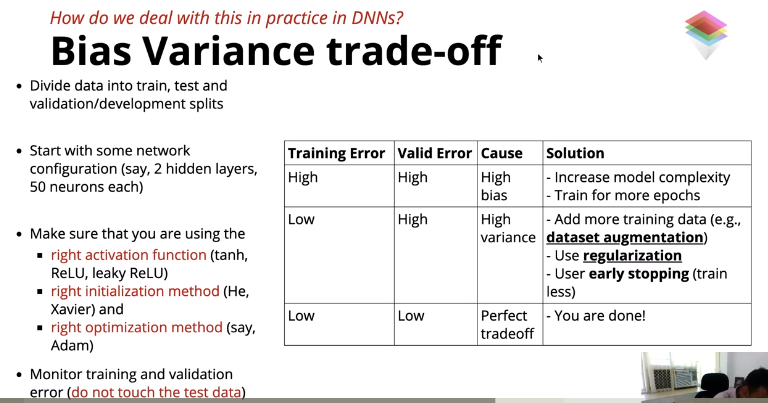
1.Rbm is symmetrical bipartite graph.

2.No Connections between visible nodes and latent or hidden nodes (i.e Restriction)



Why We need to take log likelihood instead of likelihood?

Ans) Natural logarithm is monotonically increasing function (If x value is increases the value of y also increases and this maximum value of log of probability occurs at the same point as the original probability function therefore it makes the work more simpler)



PYTORCH:

\*Very Efficient performance on GPU (Efficient Execution on GPU)

\*Numpy cannot be run on GPU

\*Auto grad which enables to do back propagation in a functional manner

\*Pytorch is very much faster than numpy

\*Pytorch + GPU is very much faster than pytorch cpu and numpy

\*K-fold cross validation more reliable estimate of out-of-sample performance than train/test split .

**Outliers Detection:**

* **An outlier is any data point which differs greatly from the rest of the observations in a dataset.**
* **Outliers are of two types: Univariate and Multivariate. A univariate outlier is a data point that consists of extreme values in one variable only, whereas a multivariate outlier is a combined unusual score on at least two variables.**
* **Isolation forest: The anomaly score is function of the number of splits**
* **Lot of splits is low anomaly score.**

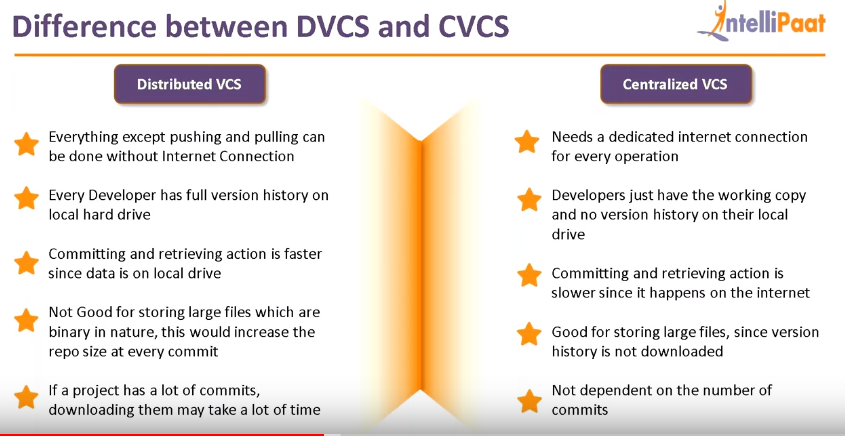
**GIT:**

* **Top 3 Distributed version control systems:**

**1.Git**

**2.Perforce**

**3.Mercurial**

* **Git is a distributed version control system for tracking changes in computer files and coordinating work on those files among multiple people.**
* **It is primarily used for source code management in software development ,but it can be used to keep track of changes in any set of files.**
* 
* **Top two cvss are 1. Subversion & 2. Helix Core**