What is a box/whisker plot/5-point summary plot and how does it helps to find an outlier.

What are all data types in R?
Numerical, integer, character, logical and complex

Data structures- list, vector, matrix, array, factor

Organization of data:

Rows- examples
Columns- features

1. **Multivariate**- rows of columns, low dimensional and dense
2. **Basket** - set of the things-market basket, keyword list, high dimensional and sparse
3. **Bag** - weighted set of things, a bag of words, a bag of visual things, high dimensional and sparse.

Feature scaling of data?

How features are changed?
- Standardization z-score
- Mean normalization
- Min-max scaling
- Unit vector

Python library - sklearn.preprocerssing.scale - for implementing the z-score

Mean normalization:- $x' = (x - \text{mean}(x)) / \max(x) - \min(x)$

PCA- principal component analysis, PCA tries to get the features with maximum and variance is high for high magnitude.

Min-max scaling:- one of the popular scaling

$$x' = (x - \min(x)) / \max(x) - \min(x)$$

Unit vectors:

$$X' = x/||x||$$

Some techniques will automatically take care of normalization- LDA

Univariate data:-

Analysis of any single column: To summarize OR describe it.

Data Set used for Analysis
http://lib.stat.cmu.edu/datasets/boston

BiVariate Data:

- Data that helps analysis of the relationship between two variables.

In applied mathematics, the logarithmic application does not work,
as **log(0) = undefined**

A logarithm is applied to get the normal distribution

MultiVariant Data: An analysis to understand the relationship between multiple variables.

BOXpLOT helps the comparison of vectors on respective min, mean, max, outliers.

##will discuss scaling later - SCALING

Covariance:

It is the measure of the joint change of two variables of interest.

$$\mathbf{Cov\ (X,\ Y) = E(\ [X - E(X)]\ [Y - E(Y)]\ )}$$

Simply, measure how much two variables change together
Expected values of the product of deviation of variable x from its mean and deviation of variable Y from its mean.

$Cov(X, Y) = E[\ (X - E(X))\ (Y - E(Y))\ ]$

$Cov(X, Y) = E[\ XY - XE(Y) - YE(X) + E(X)E(Y)\ ]$

$Cov(X, Y) = E(XY) - E(X)E(Y)$ Since $E(E(X)) = E(X)$

$Cov(X, Y) = ((\sum xy) / n) - [\ \sum x \sum y / n^2]$

Cannot compare any CoVariance directly, should standardize it.

Correlation:
It is interpreted as the degree of the linear relationship between two variables x and y.
Correlation always lies between **-1 and +1.**
It is always good to have either a positive correlation or a negative correlation because we can infer some insights only when we have some relation.

In the case of the Non-Linear trend: NO RELATION exists.

If P_{xy} is zero then there will be no relationship between variables.

$$P_{xy} = Cov\ (X, Y) / \sigma_x \sigma_y$$

$cov(x,x) = 1$

cov(x,x) = var(x)
var(x+y) = var(x) + var(y) + 2cov(x,y)
var(x-y) = var(x) + var(y) - 2cov(x,y)

Basis	**Covariance**	**Correlation**
Meaning	Covariance is a measure indicating the extent to which two random variables change in tandem.	Correlation is a statistical measure that indicates how strongly two variables are related.
What is it?	Relation b/w two variables	It is the scaled version of covariance.
Values	Lie between -ve Infinity and +ve infinity	Lie between -1 and +1
CHange in scale	Affects covariance	Does not affects the correlation
Unit free Measure	No	Yes

https://rdata.pmagunia.com/dataset/r-dataset-package-datasets-faithful

Identify other possibilities for correlation block method

Methods are circle, square, ellipse, ellipse, number, shade, color, pie.

Predictive analytics and modeling:-

Predictive analytics is the domain utilizing various aspects of stats techniques including ML, data modeling, analyzing current and historical data to make predictions for the future.

It deals with various transactional and historical data for forecasting the future with a certain degree of accuracy.

- Machine Learning
- Statistical Model
- Patterns
- Scoring
- Decision Making
- Visualization
- Consumer Behaviour
- Communication

NETFLIX and AMAZON

1. Netflix learns which movie viewers are likely to enjoy.
2. Amazon predicts what a customer will buy.

Regression:

It is a formula predicting the response as a function of the regressors making allowance for unobserved regressors.

- Response is a Dependent Variable
- Regressors are Independent Variables.

To PREDICT, STUDY, VERIFY, USE, CALIBRATE.

ε

Linear regression:-

It is both statistical and machine learning technique and is studies as a model for understanding the relationship between input and output numerical variables.

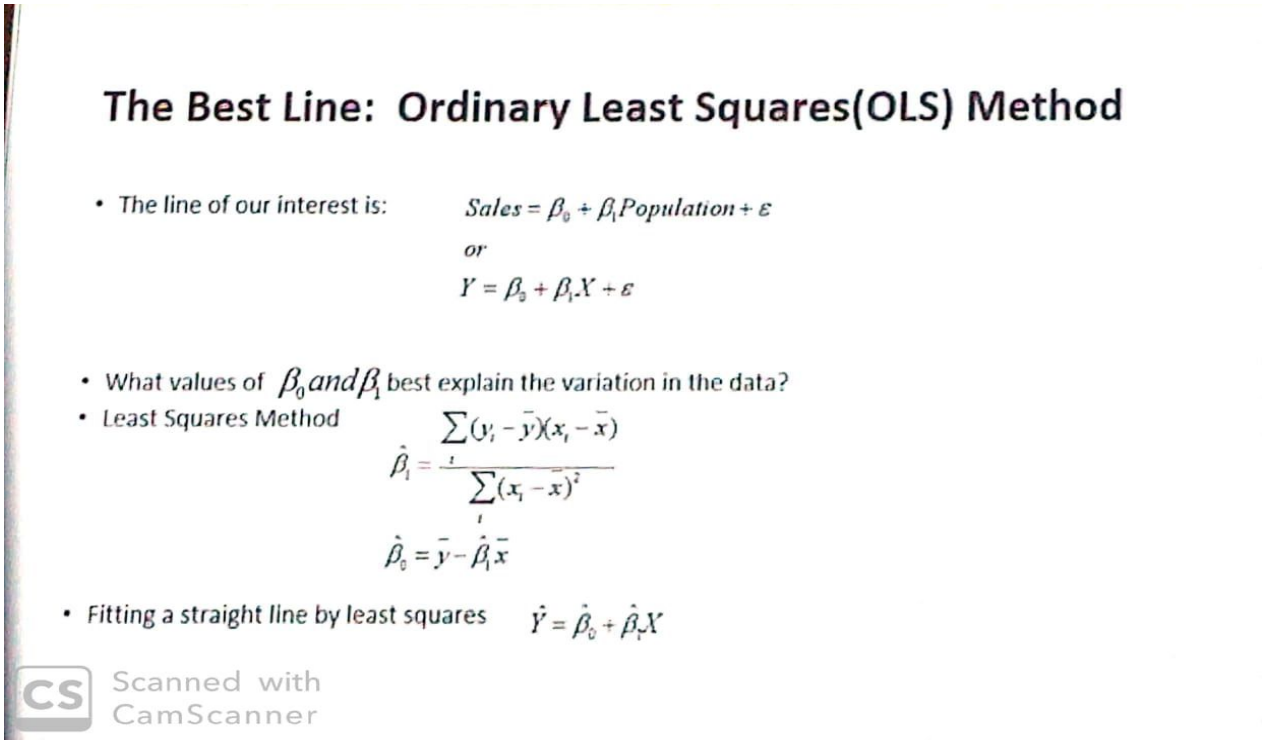
When there is a single input variable(x), the method is referred to as SIMPLE LINEAR REGRESSION

Formula:

$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

β - linear coefficient

NOTE:
Some times, **a model not in linear** may lead to **Linear Regression Model** with suitable transformations on regressors or response variables or both.



The Best Line: Ordinary Least Squares(OLS) Method

- The line of our interest is: $Sales = \beta_0 + \beta_1 Population + \varepsilon$

  or

  $Y = \beta_0 + \beta_1 X + \varepsilon$

- What values of $\beta_0 and \beta_1$ best explain the variation in the data?
- Least Squares Method

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Fitting a straight line by least squares $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$Y = \beta_o + \beta_1 x_1 + \varepsilon$

Example: sales = coefficient estimate for intercept + *(population) + ε

Residuals:-

Residuals mean errors. This a five-point summary likewise a box plot.

Coefficient estimate:-

```
Call:
lm(formula = SALES ~ POP, data = wis)

Residuals:
    Min      1Q  Median      3Q     Max
-6046.7 -1460.9  -670.5   485.6 18229.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 469.70360  702.90619   0.668    0.507
POP           0.64709    0.04881  13.258   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3792 on 48 degrees of freedom
Multiple R-squared:  0.7855,    Adjusted R-squared:  0.781
F-statistic: 175.8 on 1 and 48 DF,  p-value: < 2.2e-16

> |
```

Coefficients -Intercept:

Are constant values before considering any variable. (When x = 0)

Coefficient estimate:

This is one unit increase in X than expected change in Y, in this case, one unit change in the population then 0.64 unit change in sales.

```
POP           0.64709    0.04881  13.258   <2e-16 ***
```

Coefficient-Std Error:

- Measures how precisely the model estimates unknown value.
- Always Positive.

Note: Low value of Std Error, helps analysis and calculating Confidence

Coefficient-t Value:

.

The standard error is always positive.

T-value - estimate/std.error

Lower P values allow rejecting a Null Hypothesis.
High t-value will be helpful for our analysis

Degree of freedom will be total no.of variables - variables in the model

Error ----variation-----standard deviation
More data points we have fewer errors.
Explanatory variables/Independent/Regressor:- this is used to calculate how well the model is doing to explain the things when we increase no.of variables then it will also increase and there is no proper limit to define how much we can increase.

Multiple R squared and Adjusted R squared:

- Lies between 0 and 1.
- A higher value will be for higher variation.
- No proper limitation for how much we increase a variable.
- Only significant variables are considered in adjusted R squared.

	Intercept	Intercept Error	Multiple R^2	Adjusted R^2

POP	469.70	702.906	0.785	0.781
POP + MEDHVL	$-2.681*10^3$	$1.78* 10^3$	0.80	0.79
POP + MEDHVL + MEDSCHYR	$4.180 * 10^4$	$1.5 * 10^4$	0.8303	0.81

In the linear model, Increase in R^2 values also increases its intercept error.

It is not only concentrating about R^2 value we have to check with all the parameters like the minimum error and standard error

Homework:- f-statistics

p-value : $p < 0.05$ indicate that overall model is significant.

Multivariant linear regression: more than one independent variable

Collinearity:-

Symptoms:- when all the independent variables are correlated

If the variables have a positive correlation or negative correlation then our model can have a multi-collinearity problem.
We want our model's dependent variables should be predicted only by all the independent variables.

VIF - Variance Inflation Factor
How to detect the multi-collinearity:

- Correlation matrix
- VIF

ohoiho
$VIF_i = 1 / 1 - R_i^2$

###Font change??

Logistic regression:

It is a technique that is used to predict the Y variable when it is binary categorical. It can only take values like 1 or 0. The goal is to build the mathematical equation which predicts event 1

Why not linear regression?

- The dependent variable is not a normal distribution
- The model prediction is not continuous.
- Y ~ Binomial(n,p)

Linear Regression, can predict beyond a range of 0 to 1.

Logistic Regression lies always within a range of 0 to 1.

Sigmoidal function:

It is a mathematical function having a characteristic "s-shaped curve" or sigmoid curve.

Logistic function:

$S(x) = 1 / 1+e^{-x} = e^x / e^x + 1$

Where it is used?

- Spam detection
- Health
- Credit card fraud
- Marketing
- Banking

Odds of success is defined as the ratio of the probability of success over the probability of failure.

Example : success = 0.8, failure = 0.2

$Odds_{SUCCESS}$ = P(success) / P(failure)

= 0.8 / 0.2

= 4

P(success) = Odds / Odds + 1

Log-odds:

The logarithm of odds

This transformation is an attempt to get around the restricted range problem. It maps probability ranging between 0 and 1 to log-odds ranging from negative infinity to positive infinity.

The ordinal variable should be a factor in **R**.

For logistic we use **only binomial**.

To evaluate the performance of logistic:

1. AIC
2. Null deviance and residual deviance
3. Confusion matrix
4. ROC curve (receiver operator characteristic)

AIC(Akaike information criteria):

It always deals with the trade-off between the goodness of fit of the model and the simplicity of the model.

$AIC = 2k - 2\ln(\bar{L})$

K: Estimated no.of parameters

$\overline{L}$: Max value of the Likelihood function

Null deviance:

- Indicates the response predicted by a model with nothing but an intercept
- Lower the value better the model.

Residual Deviance:

- Indicates the response predicted by a model on adding independent variables.
- Lower the value, the better the model.

All the categorical variables are flattened to n-1 features.

Based on confusion matrix please understand the below;

1. **Accuracy**
2. **Sensitivity**
3. **Specificity**
4. **True positive rate(Sensitivity)**
5. **False-positive rate(100-Specificity)**
6. **Precision**
7. **Disspecificity**

Accuracy = (True positive + True Negative) / (TP + TN + FP + FN)

		Predicted	
		GOOD	BAD
Actual	GOOD	$GOOD_{\text{True Positive}}$	False Negative
	BAD	False Positive	$BAD_{\text{True Negative}}$

ROC:-

- evaluates the trade-off between true positives and false-positive rates.
- Summarizes the predictive power for all possible values of p > 0.5

- Higher the area under the ROC curve, the better prediction power of the model.

Steps:

- Data cleaning
- Data Validation
- E Data Analysis
- Created a logistic model
- Predicted value
- We compare the predicted value with the actual value
- Created the confusion matrix
- Calculate the accuracy

<u>Exercises to DO</u>

1. CLASS: Do a logistic regression on all columns
2. HW: Choose some columns to do the logistic regression and see if accuracy is increasing

23/ 07

PYTHON

Numpy, pandas, SciPy, matplotlib, IPython, Statsmodels, BeatifulSoup SciKits, Rpy

Write a program

Panel Data Analysis - PanDAs -> Pandas

MAchine learning <--- Artificial Intelligence

Artificial Intelligence

- Design an intelligent that perceives its environment and makes a decision.

ML_Machine LEarning

A study that gives computers the ability to learn and from experience without explicitly taught or programmed.

ML Solves many,

- dECISIONS Which offer to send to which customer?
- Basis On Past purchase behavior/ similar customer?
- Success WHat fraction of offers get converted?
- Evaluate Did customers redeem the coupons- how often?
- Improve Point of sales data, Social Data, Reviews.
- Model More data, More Features, Better Modelling, more….

Unsupervised Learning	To Find Structure in data. (clusters, Density, Patterns)
Supervised Learning	To Find Mapping between features to labels.
Semi- SUpervised	Using unlabeled data to improve supervised learning models.
Active Learning	Find which model to get labeled next to maximize value.
Reinforcement Learning	Learn local/early strategies from global/delayed "rewards".

Unsupervised Learning

- No particular targets set. Just keeps learning.
- NO predictions made.
- Observation of data happens.

Dimensionality Reduction: Reduction reduces space to store

- Decreasing the no.of variables to be considered.
- It mainly aims to represent data more compactly.
- Features in data may be "Correlated".
- Involves feature selection and feature EXTRaction.
- Feature extraction involves transforming high dimensional data into spaces of fewer dimensions.

<u>Feature Extraction</u>

A new vector from existing to easy learning.

Dimensionality Reduction	Reduction to reduce space without losing intent.
Clustering and Segmentation	

Density Estimation	
Pattern Mining	

----------*PRINCIPLES OF Edward R.Tufte*

In a 2-D plot, what is the maximum number of dimensions we can represent?

```
from google.colab import files
file = files.upload()

iris = pd.read_csv("Iris.csv")
```

Principal Component Analysis:

It is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences
PCA is a powerful tool for analyzing data

Purpose of PCA:

- Visualize
- Reduce the dimension, without loss of Information.

PCA refers to the process by which principal components are computed and used to better understand the data

What are the principal components?

With PCA, we can find the low-dimensional representation of the dataset that contains as much as possible of the variation. Therefore we only get the most interesting features.

PCA: Helps considering the fewer vectors/features that have higher variance.

PCA is USED?

- When data is multi-variate and numeric
- When the number of features is large
- When data is the unimodal-having one max mode
- We need to find the eigenvalues and eigenvectors
- When class labels are not present
- To visualize the data
- To reduce dimensions for the next stages

Summarize:

- Standardize the data
- Covariance matrix
- Obtain the eigen values and eigen vectors
- Sort the eigenvalues in descending order
- Select the k eigenvectors with the largest eigenvalues were k is the number of dimensions used in the new feature space(k <= d)

Unlabeled data cant be used for supervised learning

Unsupervised learning:

Feature extraction methods:

- Linear Method - PCA Principal Component Analysis
- Non-linear method - Multidimensional scaling

Multidimensional scaling:

- It is based on similarity or dissimilarity data

- Objects can be color, faces, map coordinates, political persuasion
- Graphs are a powerful mechanism for representing relationships between objects
- The basic assumption in MDS is that distance matrix.
- Euclidean distance between data points.
- MDS tries to position data points in Low Dimensional Space.

NOTE: PCA may not perform well with nonLinear data.

Don't decide the similarity-based only on the distance we have other techniques too.

Clustering :

The process of partitioning the given data into homogenous groups based on similar objects or given features in a group.

Most important unsupervised learning algorithm:

- **Understand**/discover structure in data
- **Summarize data points** by their cluster center
- **Compress the data** variability into "representative vector"
- **Extract features** from data for supervised learning
- **Generate class labels** when not know
- DO **Bulk Labeling** in Active Learning

Conditions for clustering:

- Scalability
- Deals with different attributes
- Cluster data with arbitrary shape
- Interpretability and usability
- High Dimensionality
- Insensitive to noise and outliers

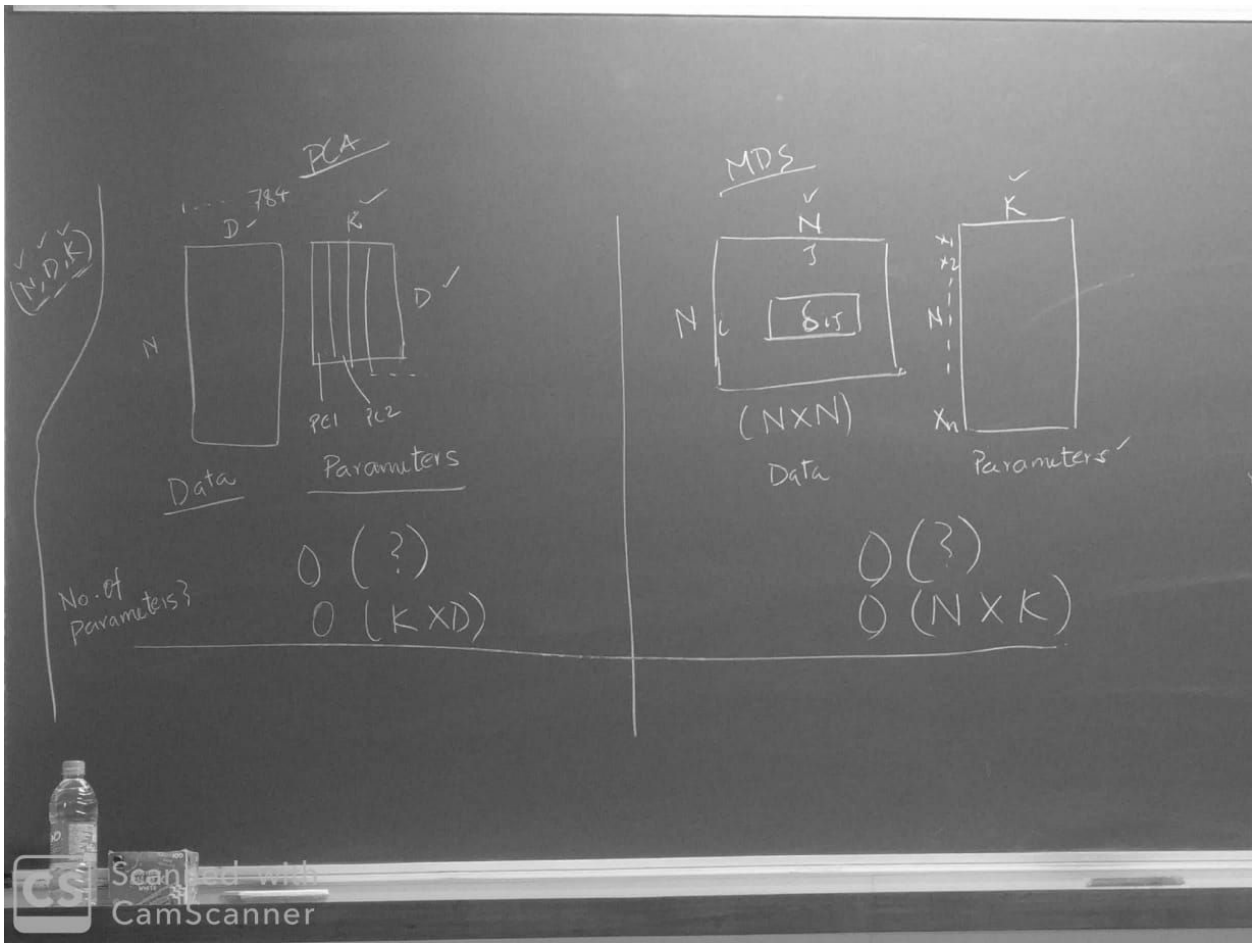
Applications:

- Recommendation engines
- Market segmentation
- Social network analysis
- Search result grouping

- Medical imaging
- Image segmentation
- Anomaly detection

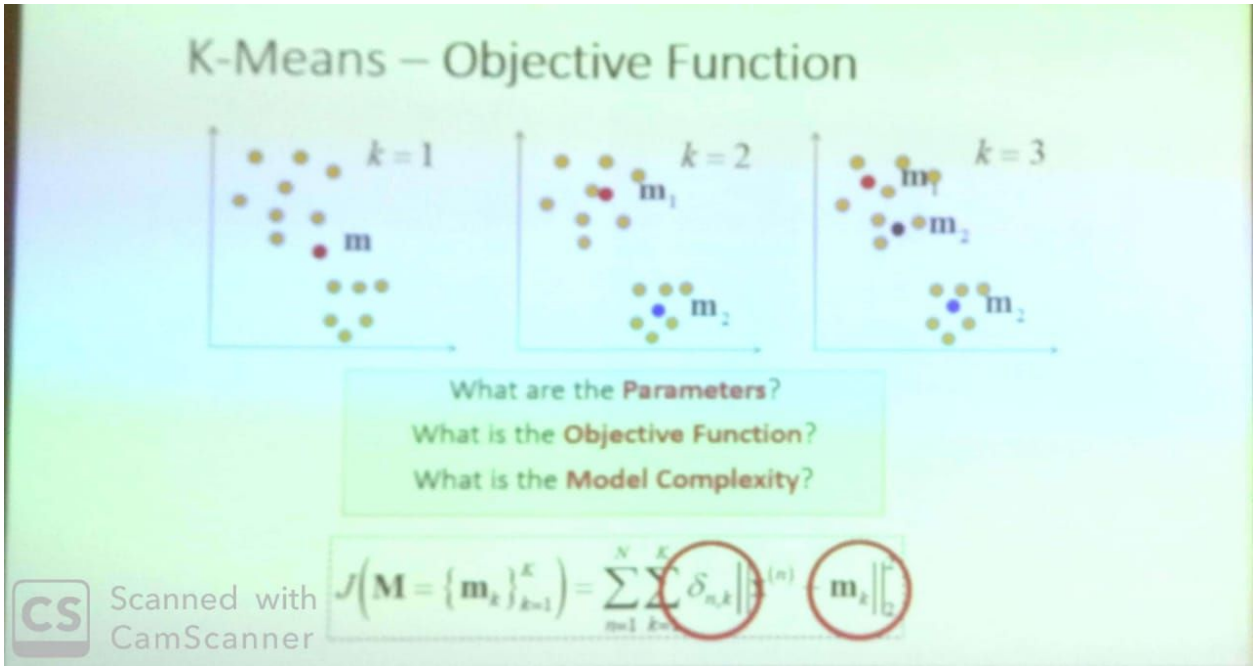
Homework:- spectral clustering
Except k-means what are other e-means algorithms?
Bias -variance trade-off?
Clustering methods:

- Partitional- hard, soft
- Spectral
- Hierarchical- agglomerative(Bottom-up), divisive(Top-down)



Partitional clustering:

K-means:
How to decide k?
What are the parameters?
What is the objective function?
What is model complexity?



Steps TO Clustering

1. **Expectation**
   Cluster centers -> cluster assignments
   O(NKD)
   N- number of parameters
   D- dimension

$$\delta_{n,k}^{(t+1)} = (k == \arg_{j = i..n} \min \{ \Delta (X^n, M_j^{(t)}) \})$$

2. **Maximization**
   Cluster assignments -> cluster centers

   O(DN)

Random initialization
Farthest first point Initialization:

The point should be marked possible far away from the existing points and so on.

- Pick the first point which is farthest from the mean of whole data
- Pick the second point which is the farthest point from the first point we have already selected
- Pick the third point which is farthest from the first and second.

The k-centers are placed in such a way ---

How many clusters?

- What is the maximum or a minimum number of clusters?
- How do we know the right number of clusters?

Hierarchical clustering:

- It groups the data **on the basis of the nearest measure** of all pairwise distance between the data points

Two types:

- top-down or divisive clustering
- Bottom-up or Agglomerative

We have to check:

The SImilarity between CLUSTER? **Sim** ({A,B}) *ON Data Points*

Advantages:

- **No prior information** on clusters is required.
- **Easy** to implement.

Disadvantages:

- It **can never undo** what was done previously
- Time complexity
- Sensitive to noise and outliers.
- Difficult to identify the correct number of clusters by the dendrogram.

Metric, distance, and similarity:

Distance from 0 to infinity then :similarity is [0,1]
Distance is 0 then :similarity is 1
Distance is infinity then :similarity is 0

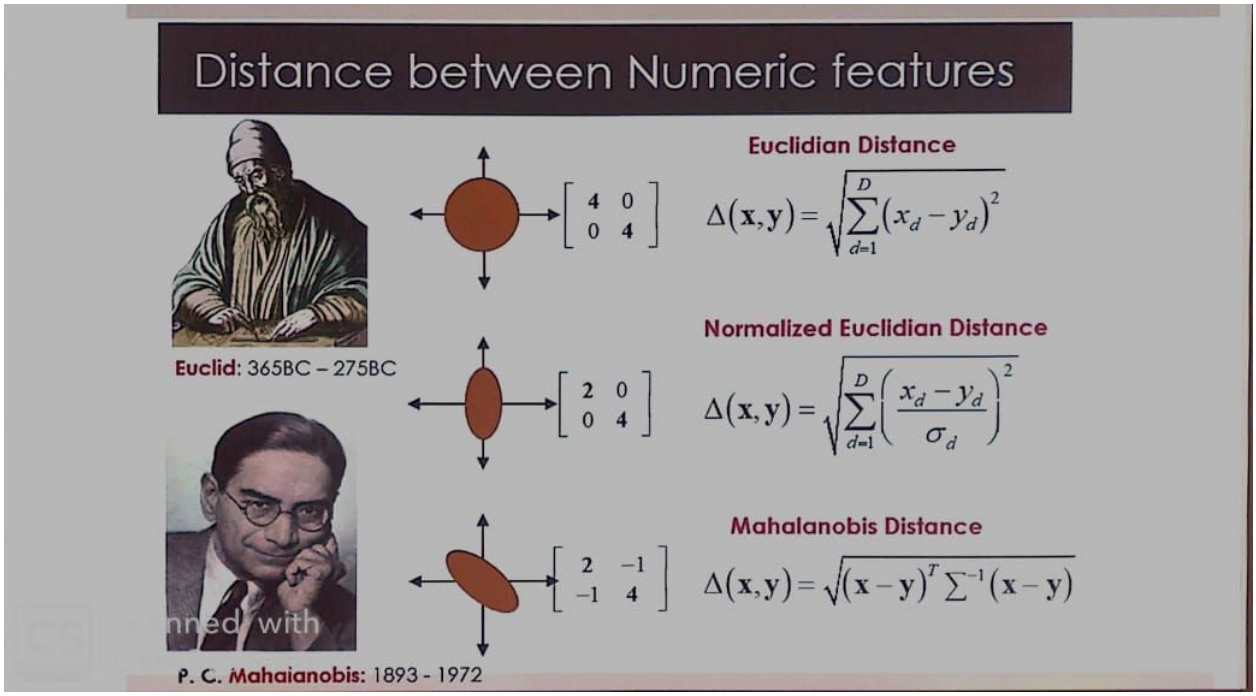
Non - Negativity	$\Delta(x,y) >= 0$
Coincidence	$\Delta(x,x) = 0$
Symmetric	$\Delta(x,y) = \Delta(y,x)$
Triangular Inequality	$\Delta(x,z) <= \Delta(x,y) + \Delta(y,z)$

Distance between numeric features:

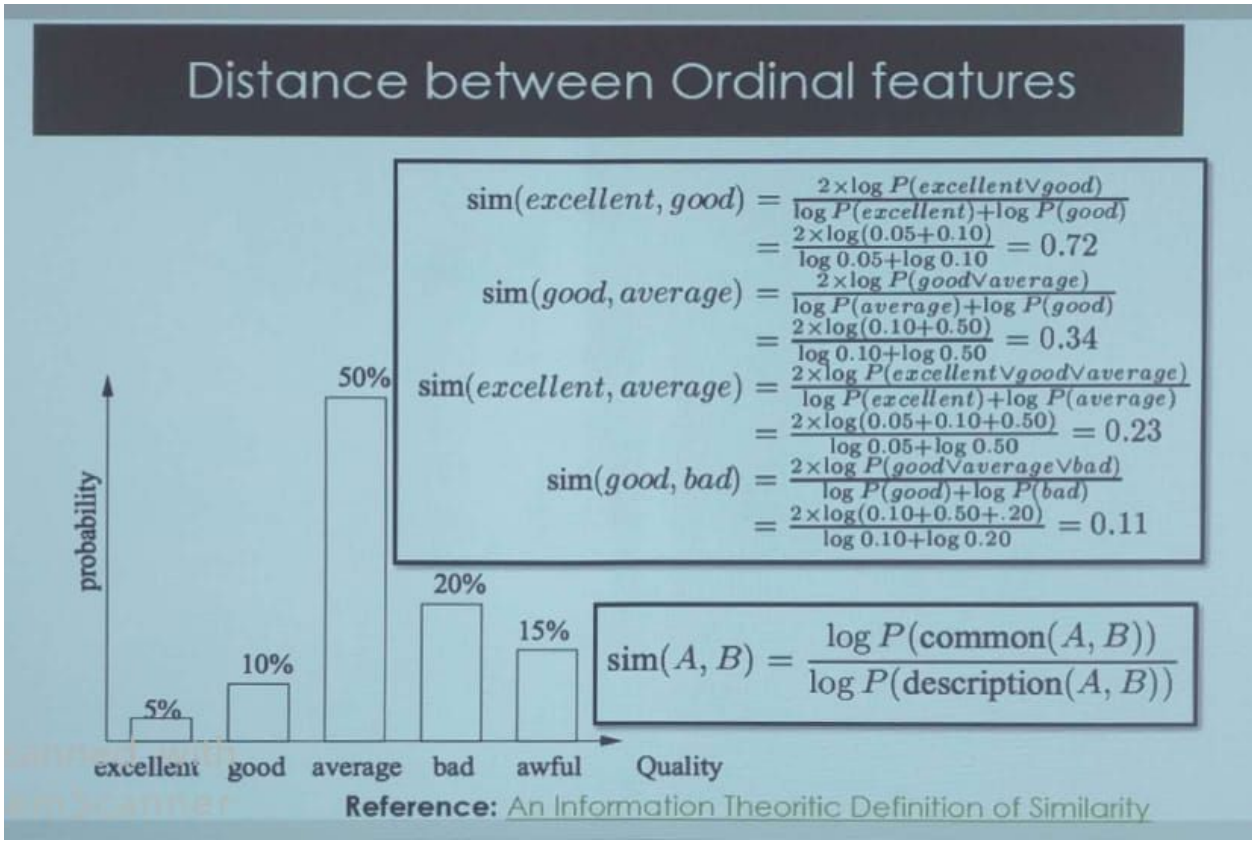
- Euclidean distance
- Normalized euclidean distance
- Mahabalanobis distance

For categorical data:

- Categorical Data Similarity depends on similarity among values of variables.



Distance between Ordinal Features:



Un-Weighted Sets

Given a two string find out the levestein distance between them?

Jaccard coefficient : Intersection / Unioin
Range: 0 to 1

Bag of words:

The similarity between two strings

- Hamming distance
- Levenshtein distance

There are five kinds of linkage possible such as single, average, complete and two other

The ML process:

PCA:- project data to preserve as much variance as possible, maximize variance

K-means centroid: - summarize the entire data using a single number

M and x
Distance between the data point and centered should be minimum
Distance between two centroids should be maximum

K-means clustering?

Find the k-representations of the entire data set
m,z

EXERCISE

N coin tosses.

H : n(H)

T : n(T)

WHat is P(H)?

- What would be an objective ?
- What

Unsupervised applications:

- Projection: PCA, MDS
- CLustering: k-means, hierarchical
- Density estimation:
- Itemset mining:

Data organization:

Item -set - basket data:

- Retail

Market basket analysis:

Apriori algorithm:

- This algorithm operates on a database Containing transactions.
- Useful for frequent itemsets and relevant association rules.
- Any subset of a large itemset must be large.
- Significant Components:
    - Support
    - Confidence
    - Lift

Limitations:

- Accurate data source in appropriate data format
- Time efficiency
- Selection of appropriate values of interesting parameters
- User specified constraints

Market Basket Analysis:

Support:- this says how much popular an itemset is. As measured by the proportion of transactions in which an itemset appears

Confidence:- this says how likely item y is purchased when item x is purchased
This is measured by the proportion of transactions with item x, in which item yhow also appears.

Lift:-

this says how likely item Y is purchased when X is purchased.
While controlling for how popular item Y is..
A lift value greater than 1 means that item Y is likely to be bought if item X is bought

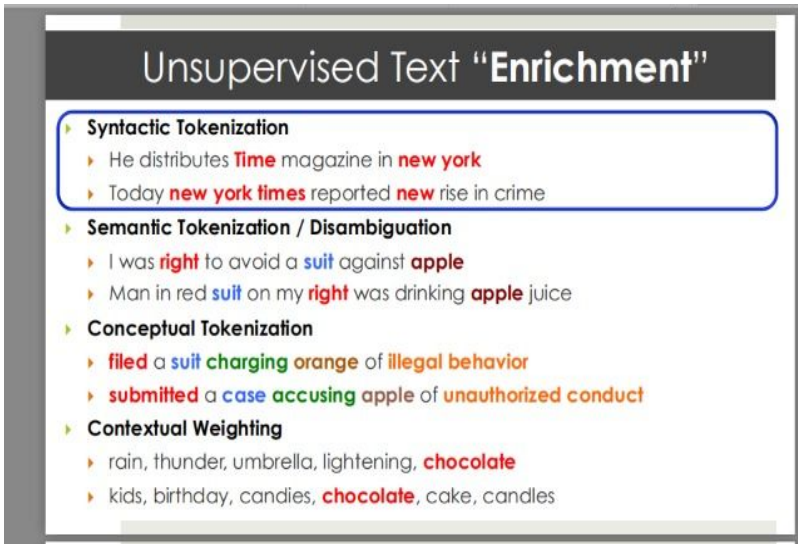
Lift < 1 that means item y is likely to be bought if the item y is unlikely to be bought if item x is bought

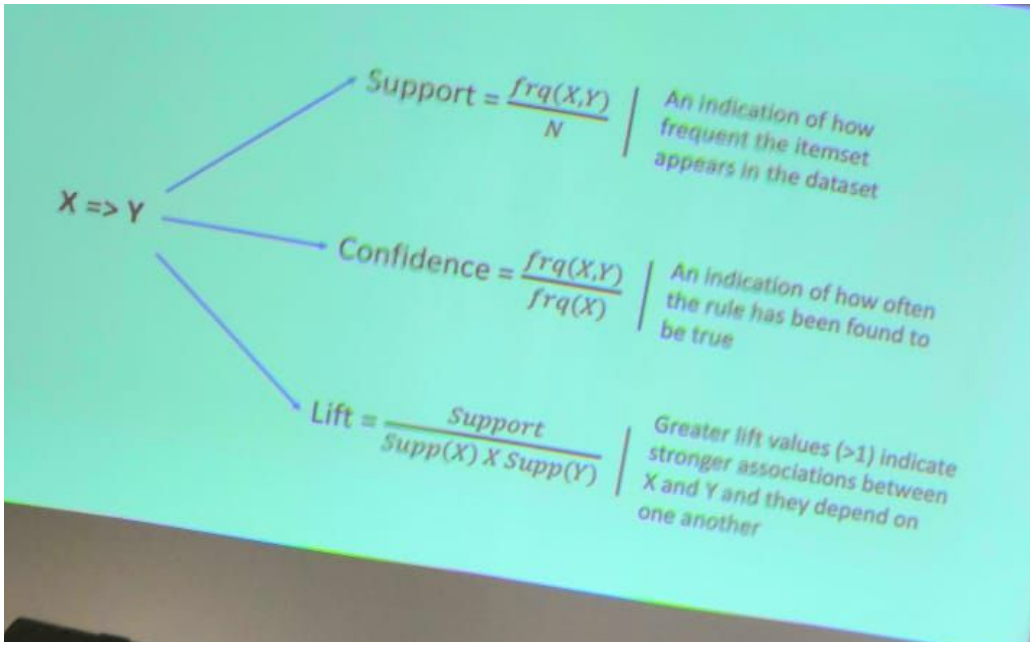
Co-occurrence analysis:

- Context-> nature of co-occurrence, market basket
- Co-occurrence -> definition of co-occurrence
- Consistency -> strength of co-occurrence
- Coherence -> tightness of a group of entities
- Community -> locally optimal groups of entities

Unsupervised, language independent text "enrichment"

- syntactic tokenization
- Semantic tokenization
- Conceptual tokenization
- Contextual weighting





Supervised learning:

Works on data which are labelled
Data with desired solution or label
Two types of techniques- regression, classification

Supervised learning paradigms:

- Regression - predict a numerical value

- Classification - predict a categorical class
- Retrieval -> predict relevance of an entity to a query

Recommendation algorithms:

- Recommendation predict user preference from a large pool of options

HomeWork:Day14

Find out the difference between parametric and non parametric classifier?

K- nearest neighbour Non-parametric classifier:

- Similar things that exist in close proximity
- Does not learn any model
- Need to store the entire data set
- Predictions are made just-in-time by calculating the similarity between an input sample and each training instance
- Many distance measures to choose from to match the structure of the input data
- Most popular distance measure used is euclidean distance

1- nn classifier:

K should be always odd number for good classification
Parameters involved?
Distance/similarity function is the key
Choosing k:

Higher the noise in the data, more k is better.

Better lookup times using KD-Trees and other tricks
Brittle in presence of noisy data
Losing the actual data
Regression the prediction is based on the mean or the median of the k-most similar instances
Classification, the output can be calculated as the class with the highest frequency i.e mode from the k-most similar instances

Advantages:

- Simple and easy to understand
- There is no need to build a model

- It can be used for classification , regression and search

Disadvantages:

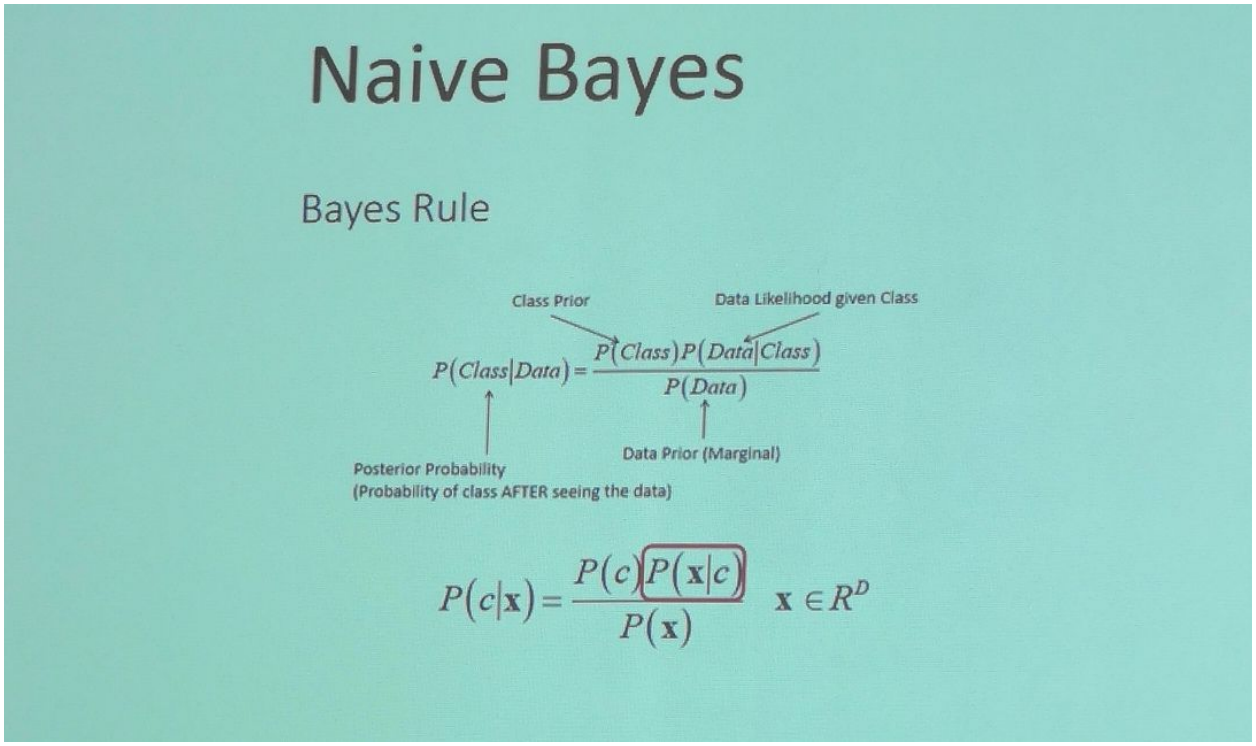
Slower as the number of examples increases

Bayesian classifiers:

Counting how many number of times each feature co-occurs with each class.
Special form of discriminant analysis
Its generative model and returns probabilities
The fundamental Naive bayes assumption is that each feature makes an:

-Independent, equal

Every pair of features being classified is independent of each other



Advantages:

- Easy to perform and fast to predict
- Performs well in multi-class prediction
- Less training and when assumption of independence holds better compared to logistic regression
- Better for categorical variable
- Text classification/ spam filtering / sentiment analysis

- Recommendation system

Disadvantage:
- Assumption of zero probability
- Assumption of independent variables

Other popular naive bayes classifiers:
- Multinomial naive bayes: follows the normal distribution
- Bernoulli naive bayes:

Decision trees:
- To visually and explicitly represent the decisions and decision making
- Covers both classification and regression
- A decision tree is a flow chart-like structure
- Each internal node represents test on a feature
- Each node represents a class label
- Branches represent conjunctions of features that lead to those
- A non-parametric algo for both classification tree and regression tree
- Target variable - discrete set of values are called classification trees
- Target variable - continuous set of values are called regression trees
- CART is the general term
- Decide on which features to choose and what conditions to use for splitting
- Know when to stop
- Entropy
- Information gain
- Gini impurity

Advantages:
- Easy to use and understand
- Can handle both categorical and numerical data
- Resistant to outliers, hence require little data pre-processing

Disadvantages:
- Prone to overfitting
- Require some kind of measurements as to how well they are doing
- Need to be careful with parameter testing
- Can create biased learned trees if some classes dominate

Pruning:

- Decision tree from overfitting the branches are removed that make

Support vector machines:

- Used for classification and regression
- Perform non-linear regression

Advantages:

- Accuracy
- Works well on smaller cleaner datasets
- It can be more efficient because it uses a subset of training points

Disadvantages:

- Not suitable for large data sets
- Less effective on noiser data sets

Which technique to be used when?

Parallel computing basics:
Data parallelism (distributed computing)

- Lots of data
- Process each "chunk" of data in parallel
- Combine results from each chunk
- Map reduce - data parallelism

Process parallelism(data flow computing):

- Lots of stages
- Pass data through all stages
- All stages running in parallel on different data
- Assembly line = process parallelism

Map-reducer:
map-----> combine ---> shuffle ----> reduce

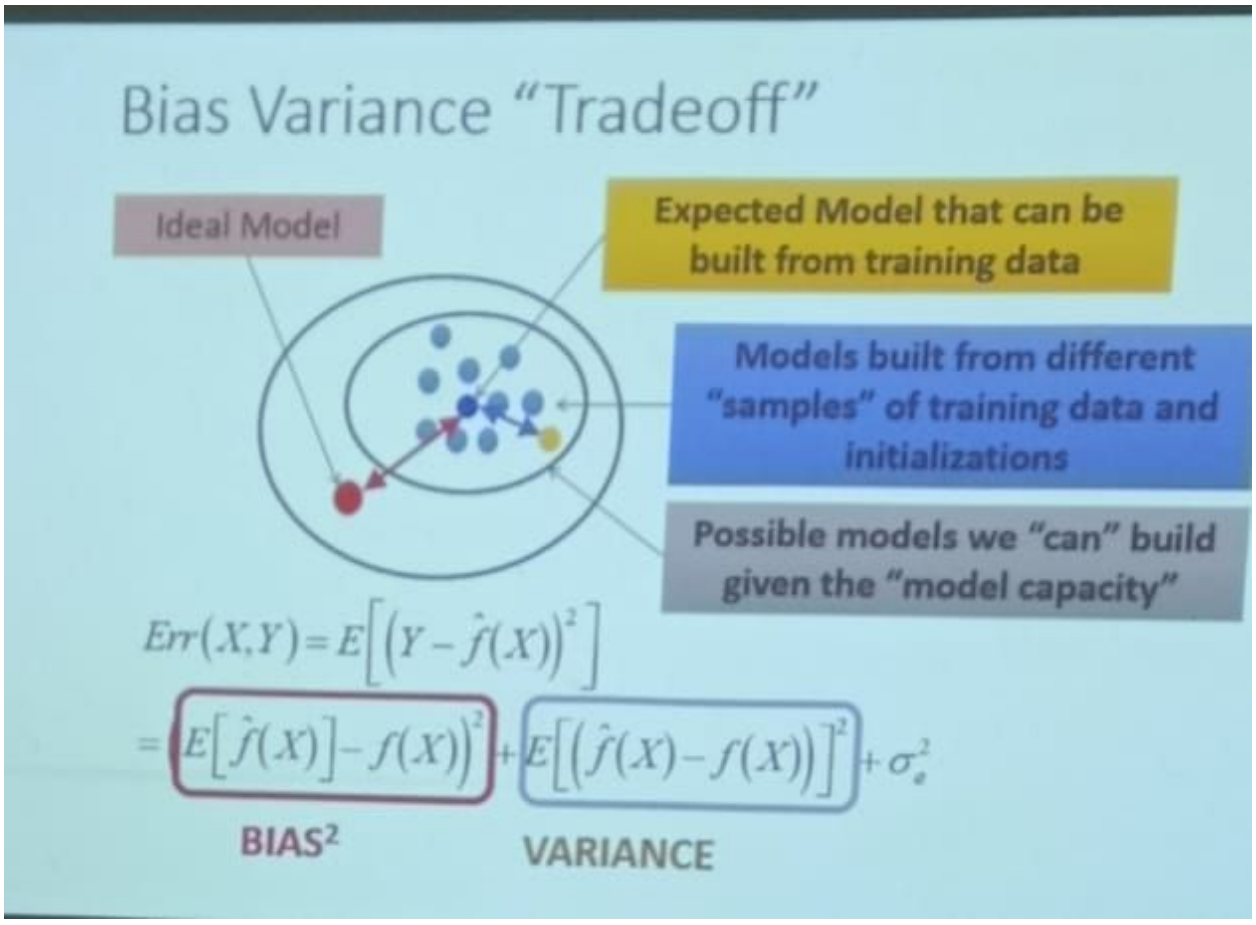
Two types of classifiers:
Descriptive:

- Learn class shapes
- Bayesian classifiers
- Nearest neighbours
- Parzen window

Discriminative:

- Learn class separators
- Decision trees
- Neural networks
- SVM

Bias variance "trade off":



All learning error can be broken down into two errors:

- Bias error
- Variance error

Bias error:

Bias refers to simplifying assumptions made by the algorithm to make the problem easier to solve
Low bias:- suggests more assumptions about the form of target function
High bias:- suggests less assumptions about the form of the target function

Variance error:
Variance is an amount that the estimate of the target function will change if different training data was used

Low variance:- suggests small changes to the estimate of the target function with changes to the training data set

High variance:- suggests large changes to estimate of the target function with the changes to the training data set

The goal of any unsupervised learning algo is to achieve low bias and low variance

Parametric or linear machine learning algo often have a high bias but a low variance

Non-parametric or non-linear ML algo will often have a low bias but a high variance

KNN- low bias and high variance, trade off can be changed by increasing the value of K.

SVM- low bias and high variance, trade off can be changed by increasing c parameter that influences the number of violations of the margin allowed in training data

There is no escaping relationship between bias and variance in machine learning
Increasing the bias will decrease the variance and vice versa

Some examples of parametric machine learning algorithms are:

- Linear Regression.
- Linear Support Vector Machines.
- Logistic Regression.

Some examples of nonparametric models are:

- Decision Trees.
- K-Nearest Neighbor.
- Support Vector Machines with Gaussian Kernels.
- Artificial Neural Networks.

Natural language processing?
With advent of internet, there is a lot of generated each day
Most of data generated is in next format and need to be defined to identify patterns
Text processing is a branch of machine learning which handles unstructured text data

Uses cases of text processing:

- Text analytics

Text analytics:

- Efficient way of analyzing unstructured data
- Adds context and color to information received from CX metrics
- Enables you to respond quickly to service issue

Information extraction: mail and google calendar
Machine translation: google translator
Document categorisation:

Is natural language processing is difficult to understand?

Yes,
Non-standard english
Segmentation issues

Idioms
Neologisms
World knowledge
Tricky entity names

NLP:- the ability of computational technologies and computational linguistics to process human natural language

- Interactions between computers and human language
- Automatic or semi-automatic
- What tools do we need?
  - Knowledge about language
  - Knowledge about world
  - A way to combine knowledge sources

  Branches of NLP:
  - Computational linguistics---type of grammar, structure of sentence
  - Machine learning---
    - Supervised ML- naive bayes, linear regression, SVM, decision tree,
    - Semi-supervised ML- graph based method
    - Unsupervised - Neural networks like RNN, CNN and soon
  - Toolbox --- scikit learn, tensor flow, no sql, hadoop

Errors:-
- False positive--matching a string that we should not have matched
- False negative-- not matching things that we should have matched

For better accuracy we have to reduce false positive and false negative

Regex are expressions are used as features in the classifiers

Basic text processing:
Word tokenization

Corpus :- It is a systematic computerized collection of authentic language that is used for linguistic analysis.
Simple terms large collection of data - corpus
Following formats:- text data, speech data

Corpus is required to perform:

- Statistical analysis such as frequency distributions co-occurrence of words

Pre-processing of text:

- Every NLP task needs to do text

Normalization:

1. segmenting/tokenizing words in running text
2. Normalizing word formats
3. Segmenting sentences in running text

Stop words removal
Important for sentiment analysis, text summarization and so on

Normalization:-

Deleting periods in a term
Deleting irrelevant numbers/ characters from data

Word normalization and stemming:-
Potentially more powerful but less efficient

Case folding

- Applications like information retrieval: reduce all the letters to lower case

For sentiment analysis, MT, Information extraction

Lemmatization:-
Reduce the inflections or variant forms to base form
Have to find correct dictionary headword form

Morphology:
Morphemes:

- Small meaningful units that make up words

- Stems: the core meaning-bearing units
- Affixes: bits and pieces that adhere to stems
    - Often with grammatical functions

Stemming:
Reduce the terms to their stems in information retrieval
Stemming is a crude chopping of affixes

- Language dependent
- Ex:- automate, automatic, automation all are reduced to automat

Porter's algorithm:

Understanding the components the components of NLP?

- Two major components of NLP
- Natural language understanding
- Natural language generation

Methods of dependency parsing?
DP
Graph
Constraint
Deterministic parsing

Language, grammar, and parsing:
Pos tag,
Parse tree,
Typed dependency

Information Extraction:

- IE systems extract clear , factual information

Name entity recognition:-

- Find and classify names in text

The ML sequence model approach to NER:

Training:

1. Collect a set of representative training documents
2. Label each token for its entity class or other(O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing:

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Word embeddings:

- A variety of word representation to allow similar words to have similar representation
- Technique to represent words as vectors in a predefined vector space
- Each word in the vocabulary defined as sparse vectors is called as one hot encoding
- One hot encoding need to have thousands and the millions of dimensions to handle entire vocabulary which is not feasible

  Ex:- bag of words, TFIDF, glove, word2vec

Bag of words:

- Describes the occurrence of words in the documents
- Steps:
  - Collect data
  - Design vocabulary
  - Create document vectors

TF: term frequency

Number of times a term appears in a particular document

IDF- Inverted Document Frequency

Number of documents in which the term is present

Word2vec:

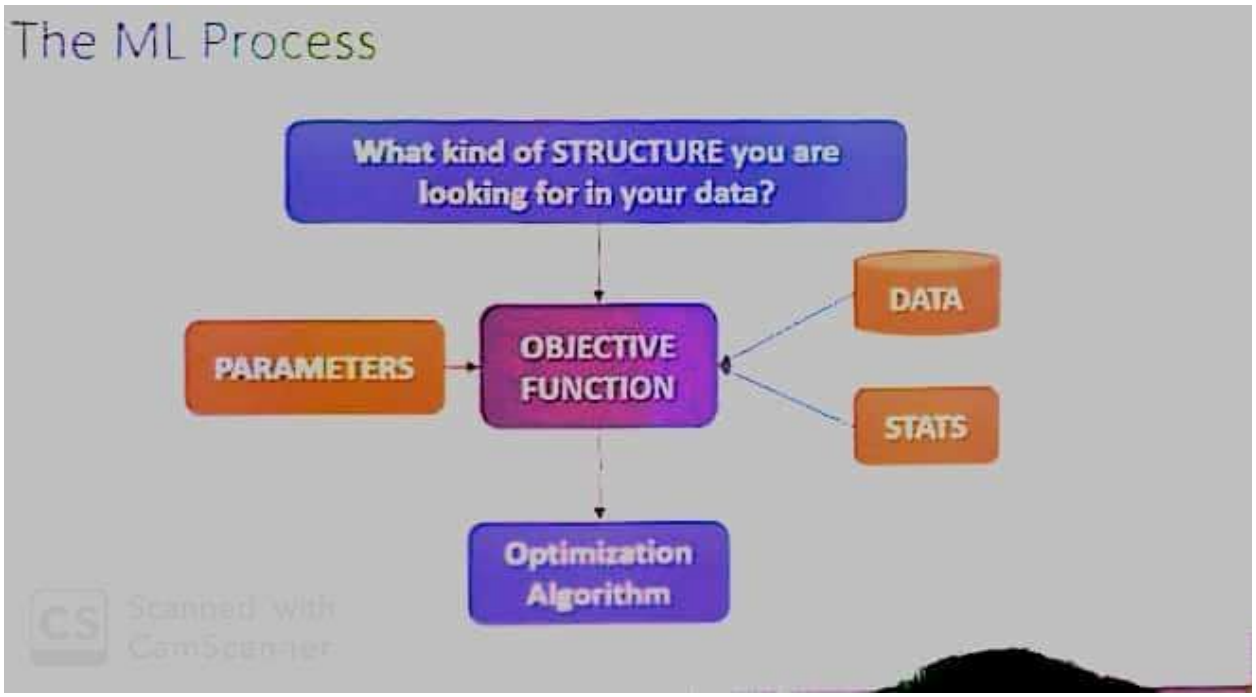
- It is a two layered neural network model

Take two data sets :- gensim package find tfidf, cosine similarity

Automated content extraction
How to build relation extractors:
Hand-written pattern
Supervised ML
Semi supervised and unsupervised:
Boot-straping, distant supervision, unsupervised learning from web

Extracting richer relations using rules:

CFG:- A context-free grammar (CFG) is a set of recursive rewriting rules (or *productions*) used to generate patterns of strings.



Text classification:-

Supervised learning:- classification model
A training set of m hand-labeled documents along with the data points

Naive bayes Intuition:
Relies on very simple representation of document

- Bag of words
- For a document d in a class c

$$P(c \mid d) = p(d \mid C)\, p(c) / p(d)$$

Use cases:
-spam filtering
Advantages:-

- Very fast, low storage requirements
- Robust to irrelevant features

If the data is very less we can use this algorithm because of its high bias nature

Sentiment analysis:
Has many other names-- opinion extraction, opinion mining, sentiment mining, subjectivity analysis

Neural network:

- Network of neurons connected together to form a computer inspired by human brain
- Artificial neuron is called perceptrons.
- Basic unit of information processing is neuron.

Artificial Neuron

Input vector → integrate function → activation function → output
(Aggregating function) ()

Integration function:

- Aggregates the input signals from other neurons.
- Positive - exhibitary
- Negative - inhibitory

Activation Function a(x):

- The activation function of a node defines the output of that node given an input or set of inputs.

  F(x) = sigm(x) = 1 / (1 + exp(-x))
- There can be different activation function:- step, ramp function, sigmoid

Types of neural networks:

- Single layer feedforward
- Multilayer feedforward:- inputs → hidden layer → output
- Feedback network
  -

Training a perceptron:

Perceptron Algorithm: An Iterative Algorithm to learn the weight vector. Update weights in proportion to the error contributed by inputs.

η → learning rate

Randomly initialize weight vector w_0
Repeat until error is less than a threshold

Choose Neural:

- When data is heavy
- Don't know which features to choose.

NOTE: low difference in weights indicates chance of convergence.

With change/Correction in weights turns out to be a rotation of

Descriptive -----> curves
Discriminative ----> lines

Learning Non linear patterns:
Non -linear → change integration function
→ multi-layered perceptrons

Type of Integration function: Quadratic Function and Spherical Functions.

Layerwise activation regions;
Helps classifying the vectors/features on a graph, for which multiple linear equations are plotted.

<u>Propagation types:-</u>

Forward propagation:-

- x no.of inputs are processed by the multiple hidden neurons and gives the output from the output layer
- This results estimation process is called forward propagation.

Backward propagation:- the loss at the output is propagated

Weight adjustment is always proportion to the error generated

Gradient descent algorithm:

- Training algorithm for a neural network (to minimize the error)
- Full batch GDA
- Stochastic GDA

Gradient in the sense slowly we are minimising the error

Full batch GDA:-

Complete data is used to compute the gradient. It uses all the training data points to update each of the weights once.

Stochastic GDA:-

Sample data is considered to update weights but not the entire data. Sample can be as low as 1 to update the weights of the MLP.

How choose number of hidden layers and how many nodes should be there is a hidden layer?

If there are more number of features then increase the number of hidden layers

Data increases the increase the number of nodes in a hidden layer

Rows(data or records) → increases then increase number of nodes
Columns or features → increases the increase the hidden layers