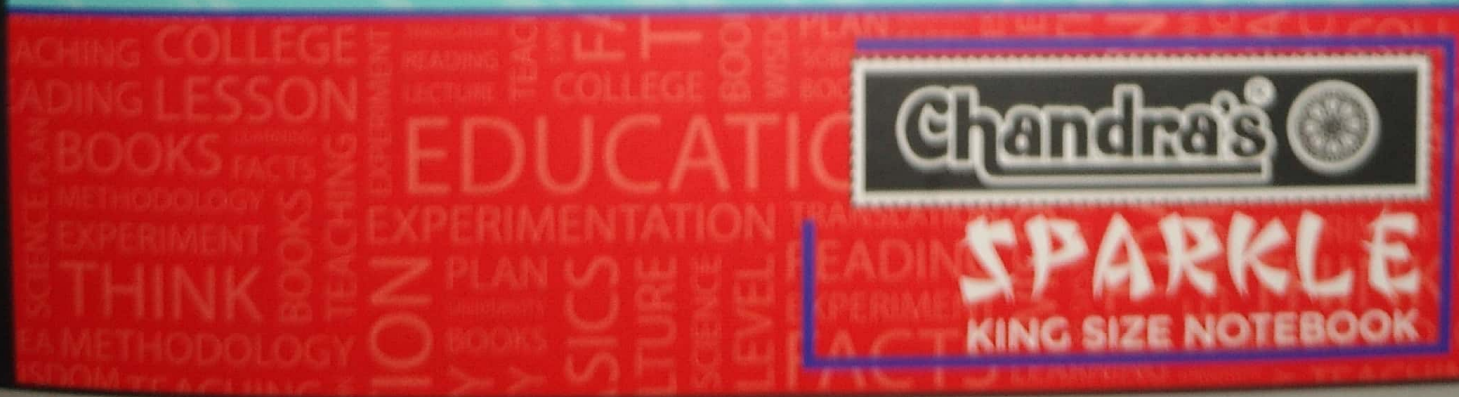
Book 2



INDEX

Name SANDEEP. A.C

Standard Section Roll No.

Subject

S. No.	Date	Title	Page No.	Teacher's Sign
5	Text Mining Data Preprocessing, TF-IDF Weighting, WordCloud		33-36	
6	Linear Algebra Matrix, Scalars, Vectors.			

Types of Machine Learning Problems

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Machine Learning.

Simple and general steps involved in creating an Machine Learning Model

1. Data Collection
2. Data Preprocessing
3. If necessary do Exploratory data analysis.
4. Choose a Model and Train it.
5. Evaluate the model.
6. Make predictions.

2.

Steps involved in Data Preprocessing

① Missing Value Analysis

→ Why Missing values

- Human Error
- Refuse to answer while surveying
- Optional box in questionnaire.

→ Should we ignore missing value or Should we impute them.

- understand why each value is missing.
- use bar graph to plot the percentage of missing values in each variable.
- Delete observations or variables where you do not intend to impute a value.
  - Drop a variable (Column)
  - Drop a observation (Row)
  - Consider variables to impute whose missing value is less than 30%.

How to impute missing values.

1. Fill with central statistics
   - → Mean
   - → Median (when skewed)
   - → Mode (Categorical data)

2. Distance based or Data mining method
   - → KNN imputation
     K Nearest Neighbour

3. Prediction method.

KNN Imputation

- use Euclidean or Manhattan distance
- Euclidean distance

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

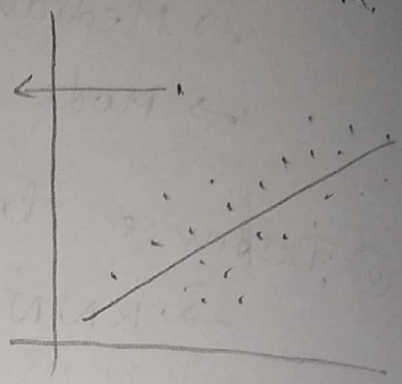
- Manhattan distance

$$d(p, q) = \sum_{i=1}^{n} |p - q|$$

(2). Outlier Analysis

Outlier is an inconsistent data when compared with rest of the data.

Sometimes we have to first perform outlier analysis and then perform missing value analysis.



Detect and replace with NA.

use Box plot to detect outliers.

(3) Feature Selection

It is extracting a subset of relevant features from the data for the use of model construction.

Also known as Dimensionality reduction

Correlation Analysis

- Correlation tells you the association between two continuous variables.
- Ranges from -1 to +1.
- Correlation is represented as 'r'
- Correlation is used when both the variables are numeric variables.

. correlation can be calculated as

$$r = \frac{\text{Covariance}(x,y)}{\text{Std Dev}(x) * \text{Std Dev}(y)}$$

. Co variance

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n}$$

ANOVA

. It is a statistical technique used to compare the means of two or more groups of observations.

. It is used to find the relation between 1 categorical variable and 1 numeric variable.

. Null Hypothesis

$H_0: \mu_1 = \mu_2 = \mu_3 \cdots = \mu_k$

μ - group mean

k - number of groups

Chi Square test

. This test is applied when you have two categorical variables from a single population

Mood / whether	Happy	Sad
Sunny	72%	28%
Rainy	72%	28%

which means no association between mood and the whether.

Mood / whether	Happy	Sad
Sunny	82%	18%
Rainy	60%	40%

There is an association between Mood and whether.

. Chi Square test can be calculated as

$$c^2 = \sum_{i=1}^{k} \left[\frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}} \right]$$

- Hypothesis Testing

H_0: Two variables are independent

H_1: Two variables are not independent.

- Degrees of freedom (v)

(number of rows - 1)*(number of columns - 1)

- Calculate critical value using table

v (vs) Confidence interval

- If chi Square statistic is greater than critical value the reject Null Hypothesis.

(4) Feature Scaling

It is a method to range the variables so that they can be compared on the common ground.

It is performed only on the continuous variable.

There are two methods to scale the data: (1) Normalization

(2) Standardization.

Normalization

It is the process of reducing unwanted variation either within or between the variables.

When variables are normalized, they will range between 0 to 1.

They are sensitive to outliers.

Formula for Normalization

$$Value_{new} = \frac{Value - minValue}{maxValue - minValue}$$

Standardization

- Also called as z-score.
- It transforms the data such that, it will have zero mean and unit variance.

mean = 0 variance = 1

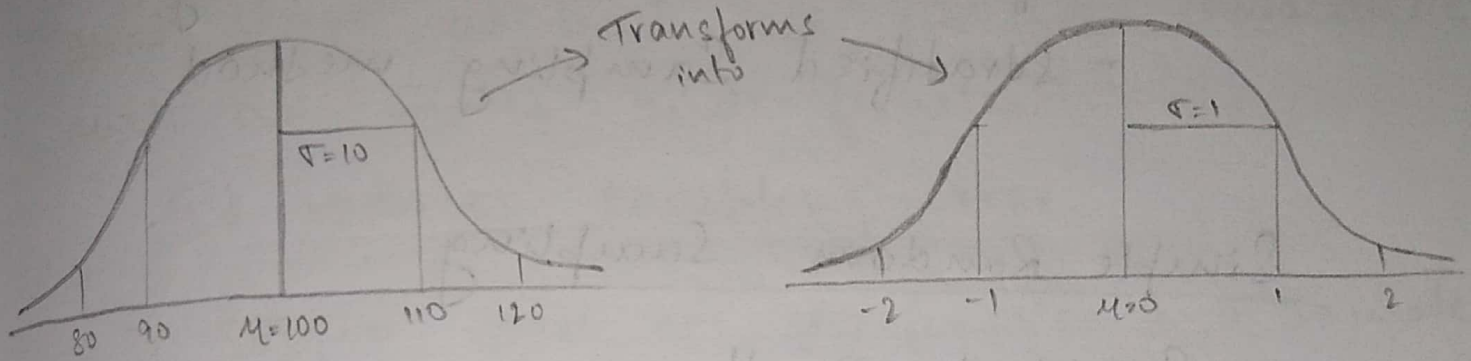
- The result is in the units of std dev. There is no scale/measure for the data.
- z is -ve when raw score is below mean and z is +ve when raw score is above mean.

- Formula to standardize

$$z = \frac{x - \mu}{\sigma}$$

μ - mean of the population
σ - Std dev of the population.

· Standardization should be applied only when the data is normally distributed.



(5) Data Sampling Techniques

· It involves extracting a subset of data from the whole population.

· The sample should represent the characteristic of whole population.

· Types of Sampling methods

Probability Samples

- Simple Random
- Systematic Random
- stratified
- Multi-stage cluster

Non Probability Samples

- Convenience
- Snowball
- Quota
- Theoritical
- ...

- Most prefered sampling techniques are
  - Simple random sampling
  - Systematic random sampling
  - Stratified Sampling method.

Simple Random Sampling

- Simplest of all.
- Each individual has equal chance of being selected.

Systematic Sampling Technique

- It is also called an k^{th} name selection technique.
- $K = \frac{N}{n}$

  N - Number of Observations

  n - Desired Sample

- Once the required sample size has been calculated, every k^{th} record is selected from a list of population.
- As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method

Stratified Sampling

- Stratum is a subset of the population that share at least one common characteristic such as males and females.
- It reduces sampling error.

Eg: Suppose there are 70 females and 30 males. And we need a sample size of 10. Now in Stratified sampling, we select 7 females and 3 males.

Evaluating the performance of the model

Error Metrics are used to evaluate the performance of the model.

- Choice of metrics depends on type of model.

Classification Metrics

① Confusion Matrix

- Describes the performance of classification Model.
- Row represents the actual class.
  Column represents the predicted class.

		Predicted class	
		Yes	No
Actual class	Yes	TP	FN
	No	FP	TN

② Accuracy

- How accurately the model can able to classify
- $$\frac{TP + TN}{\text{Total Observations}} = \text{Accuracy}$$
- Eg: Accuracy = (80 + 195) / 300 = 91.7%.

③ Misclassification Error

Error - classifying a record as belonging to one class when it belongs to another class.

Error rate - percent of misclassified records out of the total records in validation data.

$$\text{Error rate} = \frac{\text{Misclassified records}}{\text{Total Observations}}$$

$$\text{Accuracy} = 1 - \text{Error Rate}$$

④ Specificity

- The proportion of actual negative cases which are correctly identified.
- Also called as True Negative Rate.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

⑤ Recall

. Also known as Sensitivity or True Positive Rate.

. The proportion of actual positive cases which are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

⑥ False Positive Rate $= \frac{FP}{FP + TN}$

⑦ False Negative Rate $= \frac{FN}{FN + TP}$

Regression Metrics

. We want to know how well the model predicts the new data, not how well it fits the data it was trained.

. Key component of most measures is difference between actual value and predicted value (Error)

Some measures of error

① MAE (Mean Absolute Error)
MAD (Mean Absolute Deviation)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i|$$

n - number of observations
f_i - actual value
y_i - predicted value

② MAPE

- Mean Absolute percentage error
- Measures accuracy as a percentage of error.

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - P_t}{A_t} \right|$$

n - number of observations
A_t - Actual values
P_t - Predicted values

- when we need error in terms of percentage then it's better use MAPE.

③ RMSE / RMSD

- Root Mean Square Error / Deviation.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (A_t - P_t)^2}{n}}$$

A_t - Actual values
P_t - Predicted values

- If our data contains time series data or Time Series Analysis, it's better to go for RMSE.

MACHINE LEARNING MODEL'S

① Linear Regression - already explained

② Logistic Regression

- It is an classification model.
- Input can be continuous or categorical.
- Possible outcomes could be class and probabilities
- Predicts the probability of particular outcome.
- Target variable could be binomial, ordinal or multinomial.

binomial → Yes / No → 1 / 0

ordinal → low/medium/high → It is in some order

multinomial → More than one categories

Math behind Logistic regression

- Logistic function

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

· Logit Link Function

→ The logit link function transforms probabilities (between 0 and 1) to logit scores ($-\infty$ and $+\infty$)

→ Can be calculated as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 = \text{logit}(\hat{p})$$

Estimate intercept and parameters using OLS and MLE.

· In logistic regression, it will calculate the regression co-efficient for each category.

In Linear regression, it will calculate the regression co-efficient for each variable

Assumptions of Logistic Regression

1. Ratio of Number of Categories to Number of observations. should be balanced.
2. Absence of Multicollinearity
3. No Outliers
4. Independence of errors

③ Cluster Analysis

. It is used to group the observations based on independent variables.

. We will not have dependent/target variable.

. It is an unsupervised learning technique.

. clustering is the process of organizing objects into groups whose members are similar in some way

. A cluster is therefore a collection of objects which have similar characteristics. And are dissimilar to the objects belonging to other clusters.

Clustering Methods

- Distance based (K-Means)
- Hierarchical → Single link method
  - Average link method
  - Complete link method.
- Partitioning
- Probabilistic

#Important

K means clustering

The k means algorithm identifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

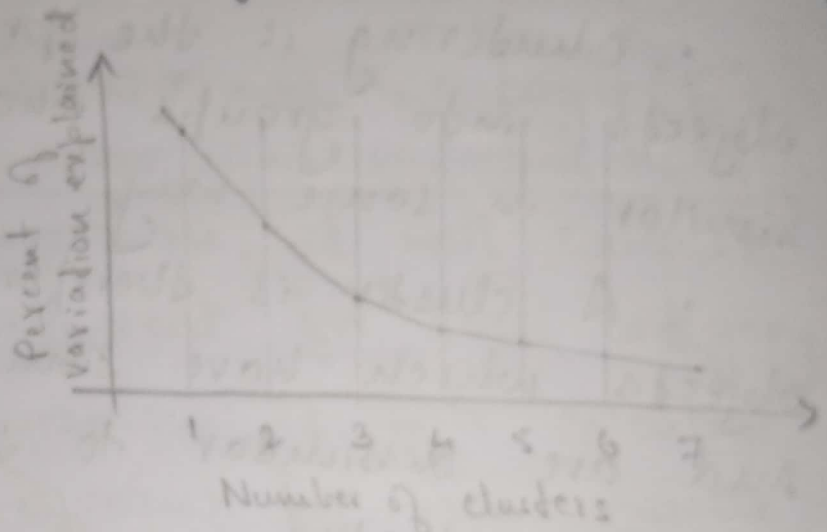
One disadvantage is, we have to define the number of centroids you need in the dataset (value k)

22

How to determine the number of clusters you need in the dataset (value k)

Answer: Elbow Method.

Elbow Method shows the percent of variation explained for each number of clusters.



---

① understand the difference between classification and Regression

(4) Decision Trees

. It is an supervised machine learning algorithm.

. It is a predictive model based on a branching series of Boolean tests.

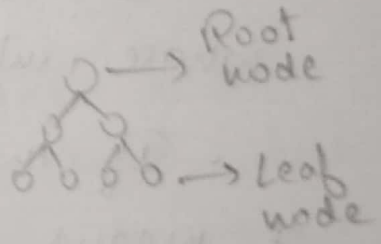
. Can be used for classification and Regression.

. Each node represents an variable and each leaf node represents an class label.

. Decision tree is a rule. Each branch connects nodes with 'and' and multiple branches are connected by 'or'.

. Extremely easy to understand by the buisness users.

. Two most popular decision tree algorithms.



- C5.0
  - Multi Split
  - Information gain
  - Rule based pruning

- CART
  - Binary split
  - Gini Index
  - Tree based pruning

Information Gain.

- We use C5.0 to build any decision tree.
- It represents the expected amount of information that would be needed.
- Measure of purity
- Loss of entropy

- **Information Gain** = Entropy of the system before split − Entropy of System after split.

- Entropy: Uncertainty in the data OR It is the measure of impurity.

It can be calculated as

$$H = -\sum_{i=1}^{c} p_i \log_2 p_i$$

- Selects the variable as parent node whose information gain is high.
- If information gain of a variable is high, it means the variable is contributing more information compared to other variables.

⑤ Random Forest

- Since we are building n number of decision trees, it is called as Random Forest. It is Random because, we are going to select randomly n number of variables and n number of observations to build each decision tree.

- Boosting: Feeding the error from one decision tree to another decision tree, is called Boosting. (To improve accuracy)

- Bagging: create several subsets of data from training data, and use of subsets of data to train the decision trees.

- Can be used for both classification and Regression.

- we can select the number of trees to be built based on trial and error method. Select the number which gives the highest accuracy.

Gini Index

- To build any decision tree in the random forest, we use CART.
- We use Gini Index to select the parent node.

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2$$

where P is probabilities of each category in a variable.

- Gini Index measures the impurity of the data.
- We select the variable whose Gini Index is low.
- A variable category with multiple target class.

$$Gini = \sum_{i \neq j} P(i)\, P(j)$$

i & j are levels of the target variable

- Internal working of algorithm.
  - $m = Sqrt(M)$

    M - Total number of variables

    m - number of variables used to determine the decision at a node of the tree.
  - Total number of observations are divided into 2 parts. 66% and 33%. (66% ↓ Train, 33% ↓ Test)

out of bag Sample method.
and error is called out of bag error.

(6) KNN

. KNN Stands K Nearest Neighbour.

. Supervised learning algorithm.

. Simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

. Can be used for classification and Regression.

. Lazy Learning, because it does not store any patterns from the Historical data.

Internal Working

. Pick a number of neighbours you want to use for classification or regression.

. Choose a method to measure distances.

. Keep a data set with records.

. For every new point, identify the number of nearest neighbours you picked using the method you choose.

. Let them vote if it is a classification or take a mean/median for regression.

Distance metric

- **Euclidean distance**

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- **Manhattan distance**

$$D = \sum_{i=1}^{n} |p - s|$$

- **weighted Distance**

$$dist(x_i, x_j) = \left((x_{i1} - x_{j1})^h + (x_{i2} - x_{j2})^h + \ldots + (x_{ir} - x_{jr})^h\right)^{1/h}$$

h - weight

Sometimes, we give more importance to some dimensions more than the other. We accomplish them with weighted distance.

(7) Naive Bayes

- classification Algorithm
- One of the most practical learning methods
- Probabilistic classification.
- It works on Bayes theorem of probability to predict the class of unknown data set.
- One of the important assumptions in Bayes theorem is, there is strong independence between variables.

Algorithm

- Provides a way of calculating posterior probability.
- $P(c|x) = \frac{P(x|c) * P(c)}{P(x)}$

$P(c|x) = P(x_1|c) * P(x_2|c) * P(x_3|c) \ldots * P(x_n|c) * P(c)$

$P(c|x)$ - Posterior probability of class given predictor.

$P(c)$ - Prior probability of class

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

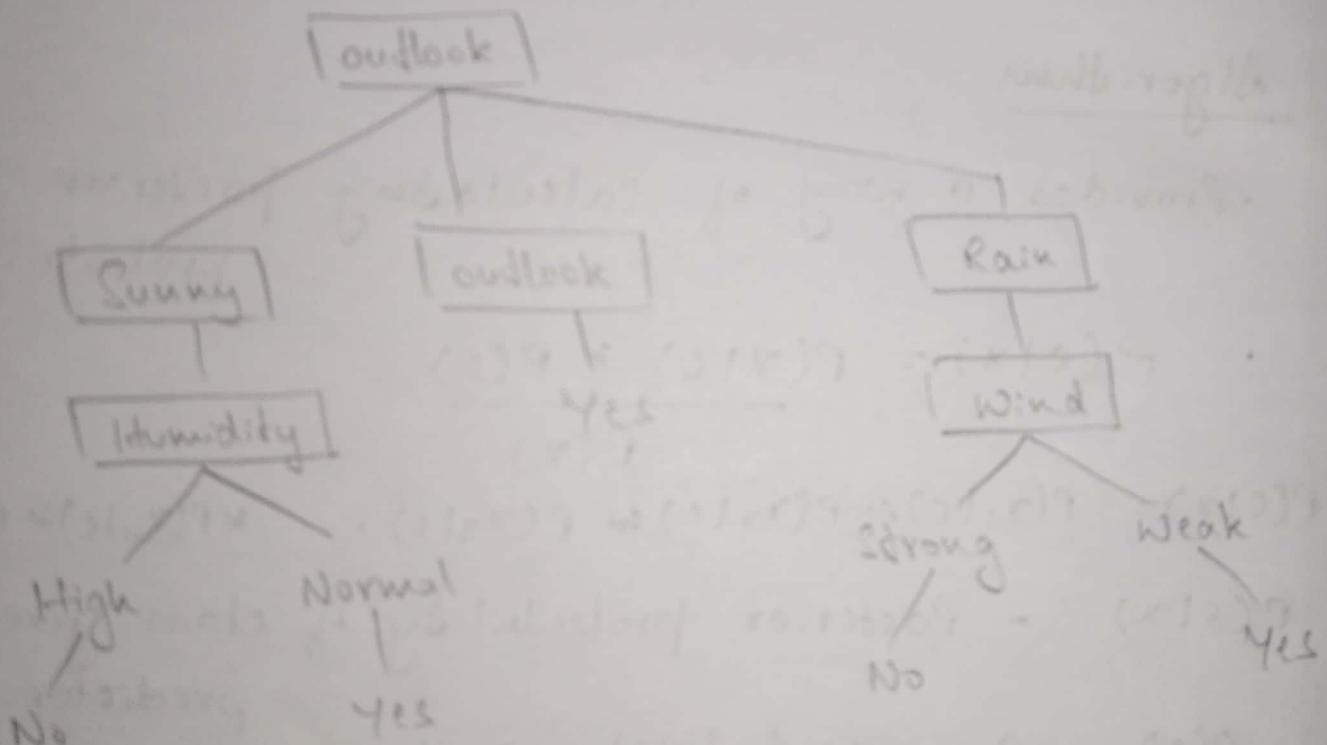
$P(x)$ - prior probability of predictor.

- If variables are continuous

$$P(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(v-\mu_c)^2}{2\sigma_c^2}}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$



Text Mining

. Collection of texts is called Corpus

Data Pre processing

. Punctuation marks
. Numbers
. Case folding (upper case and Lower case)
. Stop words
. Lemmatization

am, are, is → be

car, cars, cars, ... → car

. Stemming

computer, computational, computation → compute

TF-IDF weighting

→ TF (Term Frequency)

TF = (No of terms term t appears in a document)

(Total no of terms in the document)

→ IDF (Inverse Document Frequency)

- Tells how important a term is.

$$IDF = \log_{10}\left[\frac{\text{Total no of Documents}}{\text{No of documents with term } t \text{ in it}}\right]$$

$$\text{(TF-IDF weighting) } W = TF * IDF$$

Word cloud

Larger the value of the frequency of a word, Larger it's size in the cloud.

MATHEMATICS - LINEAR ALGEBRA

Matrix - It is collection of numbers ordered in rows and columns.

$$A = \begin{bmatrix} 5 & 12 & 6 \\ -3 & 0 & 14 \end{bmatrix}$$

2 rows

3 columns

A is a 2×3 matrix

. A matrix can only contain numbers, symbols and expressions.

$$A_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & - & - & \cdots & \vdots \\ \vdots & & & & \vdots \\ a_{m1} & - & - & - & a_{mn} \end{bmatrix}$$

In programming start with, zero(0).

- A matrix is a collection of vectors.
- Dimension = m x n
- Rank 2

Scalars

All numbers we know from algebra are referred to as scalars in linear algebra.

Eg: [15], [12], etc. — Dimension = 1 X 1

Scalars are objects with No Dimension.

Even a point is Scalar.

Rank 0

Vectors

A vector is practically the simplest linear algebraic object.

They sit somewhere between scalars and matrix.

$$\begin{bmatrix} 15 \\ 12 \\ -7 \end{bmatrix}$$ is a vector.

Matrix is a collection of vectors.

Vectors have dimension.

Rank 1

They are 2 types of vectors.

① Row Vector

$$\begin{bmatrix} 9 & 7 & 3 & 9 \end{bmatrix}$$

Dimension = 1 x m

② Column vector

$$\begin{bmatrix} 9 \\ 7 \\ 3 \\ 9 \end{bmatrix}$$

Dimension = m x 1

Operations using Matrices.

① Addition of matrices (278)

② Subtraction of matrices (278)

③ Transpose of a matrix (280)

④ Dot Product. (281)

⑤ Error when adding matrices. (279)

Tensor Flow.

- Rank 3
- It is collection of matrices.
- Dimension : $k \times m \times n$

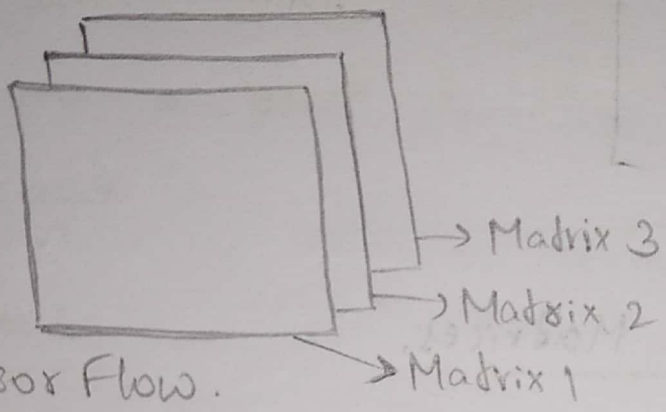


Fig: Tensor Flow.

DEEP LEARNING - NEURAL NETWORKS

Training the model is a central concept in machine learning as this is the process through which the model learns how to make sense of the input data.

4 steps involved in training an algorithm.

① Data - Readily available
② Model - Any ML model.
③ Objective Function
④ optimization Algorithm.

→ * Objective Function — It estimates how correct the models outputs are.

It is the measure used to evaluate how well the model's output's match the desired correct values.

2 types of objective functions

① Loss functions
② Reward functions.

DEEP LEARNING - NEURAL NETWORKS

Training the model is a central concept in machine learning as this is the process through which the model learns how to make sense of the input data.

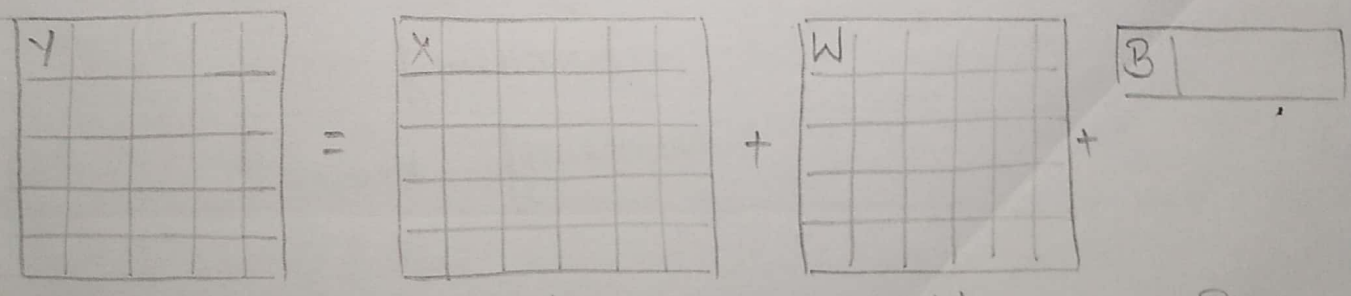
4 steps involved in training an algorithm.

1. Data - Readily available
2. Model - Any ML model.
3. Objective Function
4. optimization Algorithm.

Linear model with Multiple Inputs and Multiple outputs

K - Number of Inputs
M - Number of Outputs
N - Number of Observations

$N \times M = N \times K + K \times M + 1 \times M$



Y - Observed/predicted. X - Variable's W - weights B - Bias

Loss functions

- Also called as Cost functions.
- Lower the loss function
  Higher the level of accuracy.
- We normally encounter loss functions when dealing with supervised learning.

Common Loss functions

Regression

outputs are Numbers

↓

L2-Norm

$$L(y,t) = \sum_i (y_i - t_i)^2$$

y_i - predicted value
t_i - target value.

Lower the sum of N2-Norm, lower the error of prediction.
∴ Lower the Cost function.

classification

outputs are Categories

↓

Cross Entropy

$$L(y,t) = -\sum_i t_i \ln y_i$$

y - predicted value
t - target value

Lower the sum of Cross Entropy, lower the error of prediction.

Too high learning rate may cause the loss function to diverge to infinity. Instead of finding the minimum.

Loss functions are Higher for worse results and lower for better Results.

Loss functions

- Also called as
- Lower the loss
  Higher the
- We normally e
when dealing with

Notations

$L(y, t)$ - Loss function
$C(y, t)$ - Cost function
$E(y, t)$ - Error function.

Notations are different. But they carry same meaning.

Common Loss

Regression
outputs are Numbers

L2-Norm

$$L(y,t) = \sum_i (y_i - t_i)^2$$

y_i - predicted value
t_i - target value.

Lower the sum of N2-Norm, lower the error of prediction.
∴ Lower the cost function.

Lower the sum of Cross Entropy, lower the error of prediction.

Too high learning rate may cause the loss function to diverge to infinity. Instead of finding the minimum.

Loss functions are higher for worse results and lower for better Results.

Reward functions

. Higher the reward function
Higher the level of accuracy of the model.

. Usually reward functions are used in reinforcement learning where the goal is to maximise a specific result.

→ * Optimization Algorithm. (295 and 296)

. It consists of the mechanics through which we vary the parameters of the model to optimize the objective.

. Simplest and most fundamental optimization algorithm is Gradient Decent.

1 - Using Gradient Decent we can find the minimum value of a function through a trial and error method.

2 - There is an update rule that allows us to cherry pick the trials so we can reach the minimum faster. Each consequent trial is better than the previous one with a nice update rule.

3 - Learning rate should be high enough so we dont iterate forever, but low enough so we don't oscillate forever.

Goal is to find minimum using the Gradient descent methodology.

Eg: Considering a non machine learning example to understand the logic behind the gradient descent. (1 dimensional Gradient Decent)

$$f(x) = 5x^2 + 3x + 4$$

Step 1: Find derivate

$$f'(x) = 10x + 3$$

Step 2: Select any arbitary number x_0

eg: $x_0 = 4$

Then calculate x_1 (update operation)

$$x_{i+1} = x_i - \eta f'(x_i)$$

$$\therefore x_1 = x_0 - \eta f'(x_0)$$

$$x_1 = 4 - \eta (10(4) + 3)$$

$$= 4 - \eta(43)$$

Here η is the learning rate.

Learning rate is the rate at which the machine learning algorithm forgets old beliefs for new one's.

Continue to calculate $x_2, x_3, x_4, x_5 \ldots$ by performing update operation.

After conducting the update operation long enough, the values will eventually stop updating.

That is the point at which we know we have reached the minimum of the function. This is because the first derivative of the function is zero when we have reached the minimum.

$$x_{i+1} = x_i + \eta \underbrace{f'(x_i)}_{=0}$$

$\therefore x_{i+1} = x_i$

no longer updates.

N-Dimensional Gradient Decent. (Refer 296)