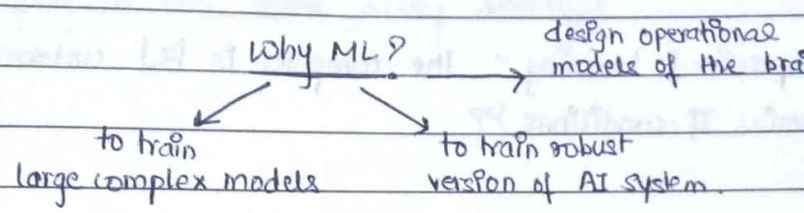
05/08/2024

Vidyalekhan DATE PAGE

* Search - provides a way of solving problems.
* Abstraction - provides a way of seperating important features & variations.
* Use of Knowledge - provides a way of solving complex problems by exploring the structure of objects.



* Features - Independent attributes. ↑ No. of features ≅ Complexity
* Algorithm vs Model - mathematical functions that work with datasets v/s solution providing mechanism.

→ ML model can never have 100% accuracy.

Defn: Machine Learning - is a subset, an application of AI, that offers the ability to the system to learn & improve from experience without being programmed to that level. It uses data to train & find accurate results. ML focuses on the development of a computer program to learn from itself.

Supervised learning	Unsupervised learning	Reinforcement learning
• Image classification	• Customer segmentation	
• Diagnostics	• Target Marketing	
• Predictions.	• Meaningful compression	

Vidyalekhan
DATE
PAGE

Defn: ~~Supervised~~ Semi-Supervised Learning - the computer is fed a mixture of correctly labelled data & unlabelled data, and searches for patterns on its own. The labelled data serves as a 'guidance' from the programmer but do not issue ongoing corrections.

Defn: Unsupervised learning - the computer is fed unlabelled data.
→ Dynamic. If-conditions ??

How? { Input raw data → Interpretation → Algorithm → Processing → Output

Defn: Deep learning - particular kind of machine learning that is inspired by the functionality of our brain cells called neurons which led to the concept of artificial neural network.
→ ML v/s DL - feature extraction, data dependency

* Simplilearn courses - Jupyter Notebook, Anaconda.

06/08/24

Defⁿ : Lists - A list in python is a collection of items which can contain elements of multiple data types which may be either numeric, character logical values etc.

Defⁿ : Array - An array is a vector containing homogeneous elements i.e belonging to the same array data type.

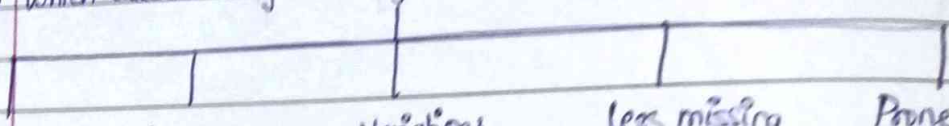
Vidyalekhan

DATE
PAGE

19/08/2024

Guidelines for AI project (CTA2):

1) Pictorial Representation of Data/Results.
2) Data containing outliers. → create your own NLP corpus!
3) Correlation of data (features)
4) Which dataset is good enough for our model?



Ref Book: Mathematics for Machine Learning (Marc Peter)

Defn: * Statistics - collection, organization, analysis, interpretation & representation of data.

* Sample size should be 10% of total data collected.

DPTPK?? Scrapping??

Defn: Crawler - jumping from one webpage to another in order to retrieve data.

* Websites to collect data for ML models - surveymonkey, mfork.

Defn: Population - All objects or measurements whose properties are being observed.

DATE
PAGE

Defn: Parameter – A metric that is used to represent a population characteri

Defn: Sample – A subset of the population studied.

Defn: Variable – A metric for interest for each item in the population.

Types of Sampling

Probabilistic Approach
Selecting samples from a larger pop. using a method based on theory of probab.
Ex: Random, Systematic, Stratified.

Non-Probablistic Appr
Selecting samples based on subjecti judgement of the researcher.
Ex: Convinience, Quota, Snowba

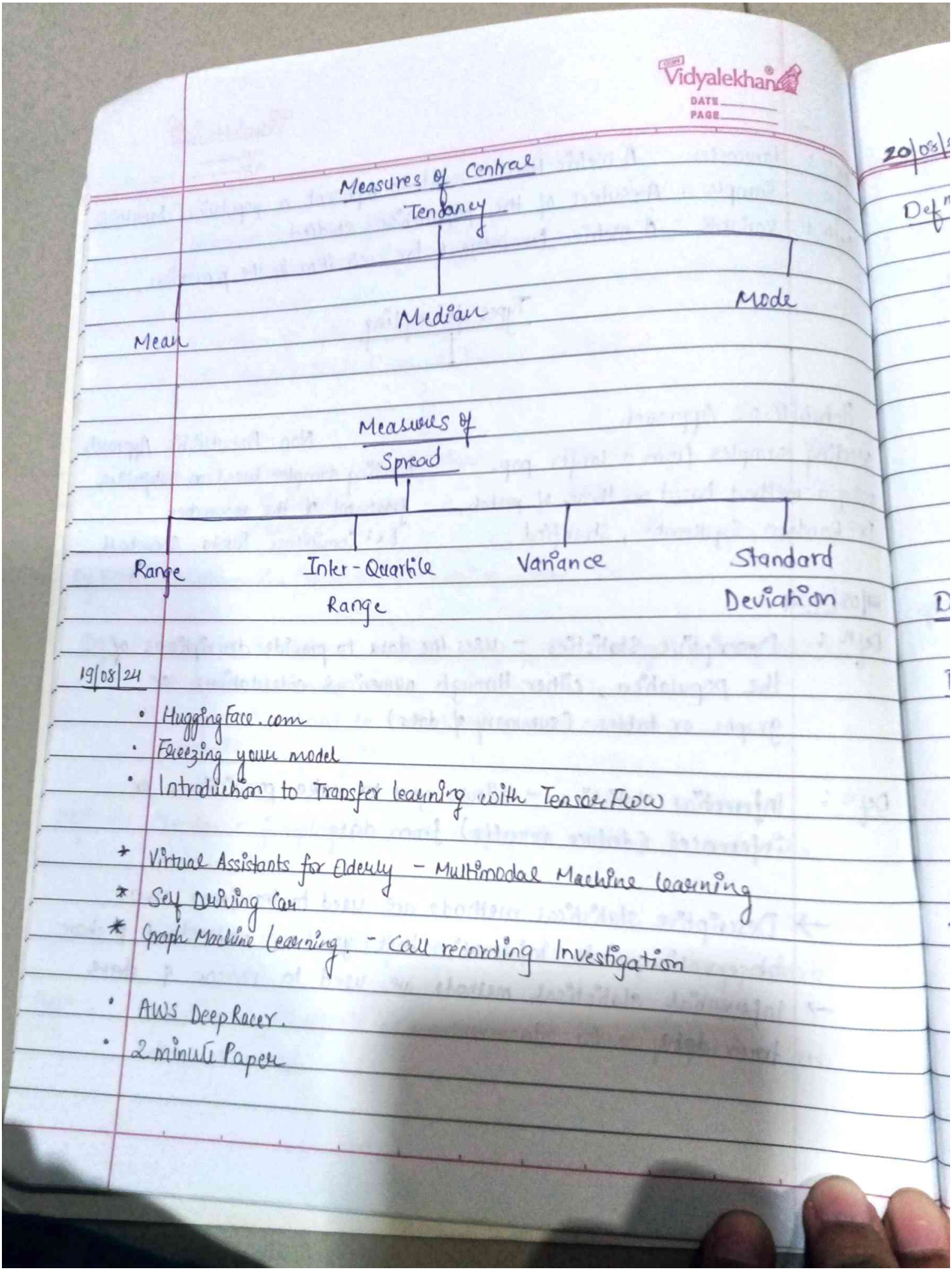
14/08/2024

Defn: Descriptive Statistics – uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables. (summary of data).

Defn: Inferential statistics – allows you to make predictions or inferences (deduce results) from data.

→ Descriptive statistical methods are used to transform raw observations into information that you can understand & s

→ Inferential statistical methods are used to reason & shar from data.


Vidyalekhan
DATE
PAGE
Measures of Central Tendancy
Mean
Median
Mode
Measures of Spread
Range
Inter-Quartile Range
Variance
Standard Deviation
19/08/24
• HuggingFace.com
• Freezing your model
• Introduction to Transfer learning with TensorFlow
* Virtual Assistants for Elderly - Multimodal Machine learning
* Self Driving Car
* Graph Machine Learning. - Call recording Investigation
• AWS DeepRacer.
• 2 minute Paper

20/03/24

Vidyalekhan
DATE
PAGE

Defⁿ: Inferential Statistics – uses statistical techniques to extrapolate information from a smaller sample to make predictions & raw conclusions about a larger population.
It uses probability theory & statistical models to estimate population parameters & test population hypotheses based on sample data.

* Null Hypotheses v/s Secondary Hypotheses.

The main goal of inferential statistics is to ~~find~~ provide information about the whole population hypotheses based on sample data.

Defⁿ: Regression Analysis – to define the relationships b/w variables.

Defⁿ: Regression Function – a function is a set of ordered pairs of numbers (x, y) such that each value of the first variable (x) there corresponds a unique value of the second variable (y).
$y = f(x)$: value of y depends on the value of x.
↳ a function that describes the relationship

Terms: Observed Value, Predicted Value } Random Error
Intercept

21/08/24

- How are covariance & correlation relevant to data analytics
→ Understanding the relationship b/w continuous variables
→ Covariance indicates whether two variables fluctuate in the same (+ve) or opposite (-ve) ~~covariance~~ direction.
→ The numerical value of covariance is not important, only sign is
→ Correlation describes how a change in one variable leads to a change in the percentage of second variable [-1, 1]
→ if correlation value is 0, it suggests that there is no linear link b/w the variables, but another functional relationship may exist.

- What is Covariance
→ Covariance is a statistical term that refers to a systematic relationship b/w two random variables in which a change in the others reflects a change in one variable.
→ The covariance value can range from $-\infty$ to ∞
→ The greater this number, the more reliant the relationship.
→ Covariance is great for defining the type of relationship but is terrible for interpreting the magnitude.

$$\text{Covariance}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$E(x)$, $E(y)$ - expected values of variables.
x_i - data value of x y_i - data value of y.
$\bar{x}$ - mean of x $\bar{y}$ = mean of y
N = no. of data values.

- Application of Covariance.

→ To reduce the dimensions of large data sets, PCA is used.

→ To perform PCA, eigen decomposition is applied to covariance matri

- Types of Correlations.

→ Simple correlation – single no. represents the relationship b/w 2 variab

→ Partial correlation – when one variable's effects are removed, th correlation b/w 2 variables is revealed in partial correlation

→ Multiple correlation – uses 2 or more variables to predict the of one variable.

- Applications of Correlation.

→ The goal when dealing with large amounts of data is to find patterns

→ For use in other analyses when excluding missing values pairwi

- Difference b/w Correlation & Covariance.

22/08/24

$x = 4, 8, 12, 16$

$y = 5, 10, 15, 20$

x	x^2	y	y^2	xy
4	16	5	25	20
8	64	10	100	80
12	144	15	225	180
16	256	20	400	320
40	480	50	750	600

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

$$r = \frac{4 \times 600 - 40 \times 50}{\sqrt{4 \times 480 \times 1600 - 4 \times 750 \times 2500}}$$

$$r = \frac{2400 - 2000}{\sqrt{3072000 - 7500000}} = \frac{400}{\sqrt{-4428000}}$$

$$r = \frac{400}{-2,104} = -0.190$$

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

$$r = \frac{5 \times 14,670 - 89 \times 820}{\sqrt{5(1605 - 7921) \times 5(134986 - 672400)}}$$

$$r = \frac{73,350 - 72,980}{\sqrt{-31580 \times -2687070}}$$

$$r = \frac{370}{\sqrt{-84857670600}} = \frac{370}{-291303}$$

$$= 0.0012$$

Defn: Matrix - a matrix refers to a rectangular representation of an array of numbers arranged in columns & rows.

29/08/24

* Numpy Library *

1-D Array :

1	2	9	10

shape : (4,)

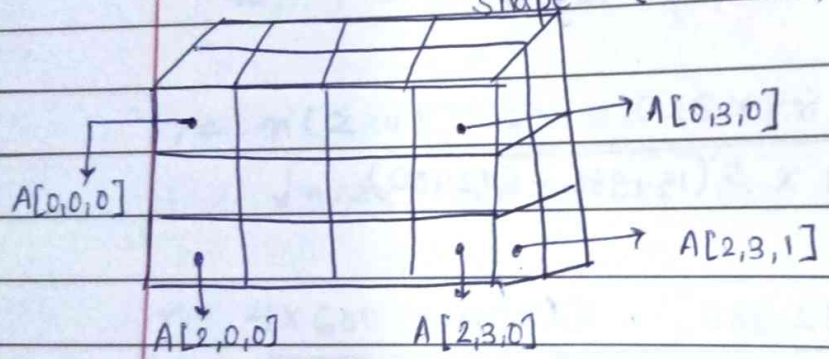
2-D Array :

5	0	1
9	2	4

shape : (2, 3)

3-D Array : shape : (4, 3, 2)

shape : (axis 0 ↓, axis 1 →, axis 2 ↗)



Defn: Slicing - You can select a range of elements in an array given the following syntax : [start : stop : step].

Vidyalekhan
DATE
PAGE

TA1 - Notes

Introduction to Python

→ Brief history of python

- Invented in Netherlands, early 90s, by Guido Van Rossum & named after Monty Python. (British comedy group)
- Open-sourced, scalable, object-oriented, functional, scripting language

→ Basic Python Commands

#	Single-line comments
" "	Multi-line comments
type()	To check the data type (int, string, char etc.) of the variable
range(s:s:s)	To create a list of integers
=	Assignment operator
==	Test for Quality (Conditional statements)
break	Exit out the loop when particular output is reached
continue	Continue with the next iteration of the loop
len()	Length of a data structure

→ Conditional Statements in Python.

```
a = 200   } Assignment
b = 33    }
if (b>a): } Condition (1) check
    print("B is greater than a")
elif (a==b): } if (1) was not true then check this
    print("A & B are equal")
else: } if none of the aforementioned conditions were true
    print("A is greater than B")
```

→ Python For loops.

```
fruits = ["apple", "banana", "cherry"] iterating over a sequence
for x in fruits :
    print(x)
    if x == "banana" :
        break    stop the loop abruptly.
```

Defn : Lists - Lists are used to store multiple items (of diff data types) in a single variable, enclosed in square brackets.

Defn : Dictionaries - Dictionaries are used to store data in key : value pairs. It is an ordered collection of objects & does not permit duplicate key values, enclosed in curly brackets.

→ Functions in python

```
(Keyword) def function_name (arg1, arg2, ..., argn) :    function initiation.
              Statement 1
              Statement 2 ....
              return          (optional)
```

→ NumPy.

- Python library used for working with arrays, linear algebra, Fourier transform, matrices.
- Open source project created by Travis Oliphant in 2005.
- NumPy - Numerical Python.

→ MatPlotLib.

- A low-level graph plotting library in python that serves as a visualization utility.
- Open-source project created by John D. Hunter.
- Pyplot - submodule plt - alias.

Mathematics for Machine Learning.

Defn: Scalar - A quantity that has magnitude but no direction.

Defn: Vector - A quantity that has both magnitude & direction.

$\|v\| = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}$ unit vector $(\hat{v}) = v/\|v\|$

$c = a \cdot b = (a1 \times b1 + a2 \times b2 + a3 \times b3) = \|a\| \|b\| \cos\theta$

Defn: Linear Independence - If 2 vectors point in different directions then the vectors are said to be linearly independent.

Defn: Orthogonal vectors - 2 perpendicular vectors.

Defn: Vector Space - The space in which vectors live.

Defn: Basis - linearly independent minimal set of vectors that, when used in linear combination can represent every vector in a given vector space.

18/9/24

Name (x_1)	Gender (x_2)	YOE (x_3)	Salary (y)
Rajesh	M	2	10,000
Ramesh	M	6	15,000
Suresh	M	4	20,000
Sumit	M	3	25,000
Shubham	M	1	11,30,000

x_1; x_2; x_3 → Features / Attributes.

Defn: Feature selection is a way of selecting the subset of the most relevant features from the original feature set by removing redundant, irrelevant or noisy features.

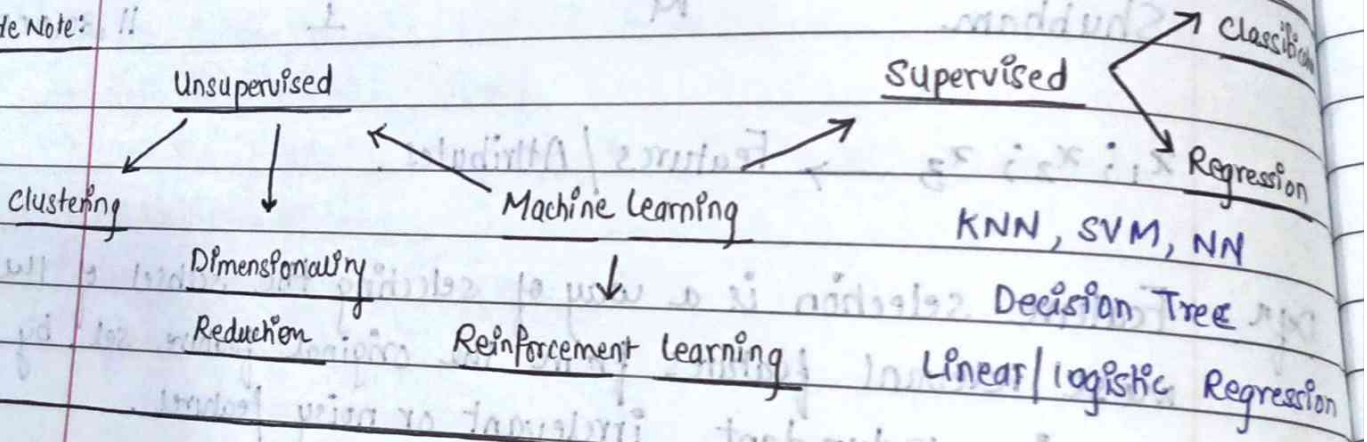
- Reduce the dimensionality (Principal Component Analysis)
- Simplify the model
- Reduce overfitting.

Algorithms & Numericals

Decision Tree:

Definition: A decision tree is a supervised learning algorithm used for classification & regression of new data points. [mainly used for classification]

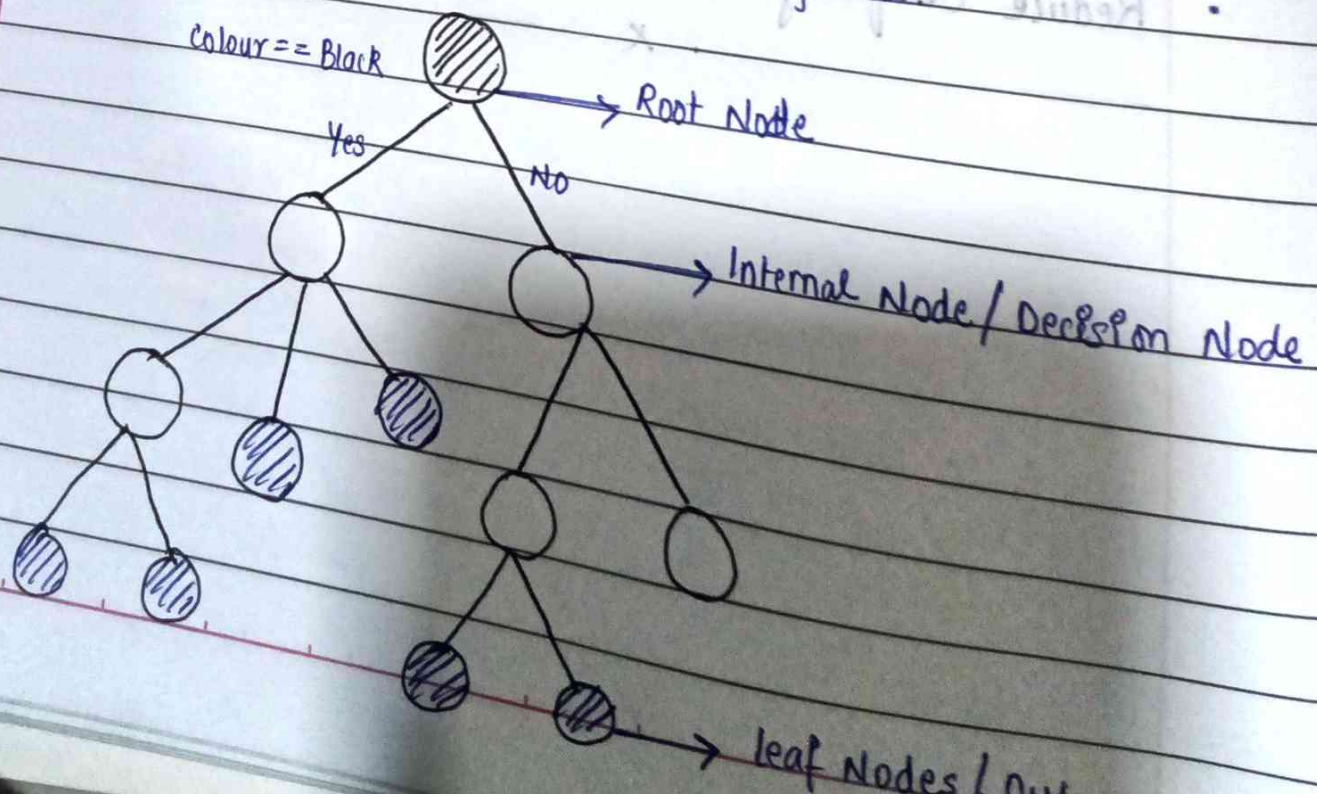
Side Note:



Types:

1) Regression Tree - dependent variable is continuous
2) Classification Tree - dependent variable is categorical

Structure:



Vidyalekhan

DATE
PAGE

Splitting: How will you split your data? What is a good split?

Measures of split: Information Gain, Entropy, Gini Index.

1. Information Gain – how much a particular feature contributes to making accurate classifications.
2. Entropy – measure of uncertainty in a decision split classification.
3. Gini Index – measure of how mixed or impure a dataset is.

Example:

Feature X	Feature Y	Feature Z	Class Labels
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

Assume: 3 features (X, Y, Z) ; 2 output classes (A, B)

Splitting on feature X :

Entropy $= \sum_{i=1}^{2} -p_i * \log_2(p_i)$

$\sum_{i=1}^{2} -p_{A/B} * \log_2(p_{A/B}) \Rightarrow -p_A * \log_2(p_A) + [-p_B * \log_2(p_B)]$

Formulae:

$$\text{Entropy} = \sum_{i=1}^{c} -p_i * \log_2(p_i)$$

c = no. of classes ; i = class labels.

$$\text{Information Gain} = \text{Entropy(Sample)} - \sum_{v} \frac{|S_v|}{|S|} E(S_v)$$

v = values in an attribute ; $E(S_v)$ = Entropy of sample attribute values.

$$\text{Gini Index} = 1 - \sum_{i=1}^{c} (p_i)^2$$

c = no. of classes ; i = class label.

Vidyalekhan

DATE

PAGE

Example:

Day	Weather	Temperature	Wind	Play Football?
1	Sunny	Hot	weak	No
2	Sunny	Hot	Strong	No
3	Cloudy	Hot	weak	Yes
4	Rain	Mild	Weak	Yes
5	Rain	Cool	Weak	Yes

Step 1: Entropy of the entire dataset ($E(S)$):

$S\{+3, -2\} = \frac{-3}{5} \times \log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times \log_2\left(\frac{2}{5}\right)$

$= -3/5 \times -0.737 - 2/5 \times -1.322$

$= 0.4422 + 0.5288$

$= 0.971$

Step 2: Entropy of all attributes ($E(S_v)$):

Weather → Sunny $S\{0, -2\}$

→ Cloudy $S\{1, 0\}$

→ Rainy $S\{2, 0\}$

Temperature → Hot $S\{0, -2\}$

Mild $S\{1, 0\}$

Cool $S\{1, 0\}$

Step 3: Calculate Info. Gain.

IG = Entropy (whole data) $- \frac{2}{5}$ Entropy (Sunny) $- \frac{1}{5}$ Entropy (Cloudy)

$- \frac{2}{5}$ Entropy (Rainy)

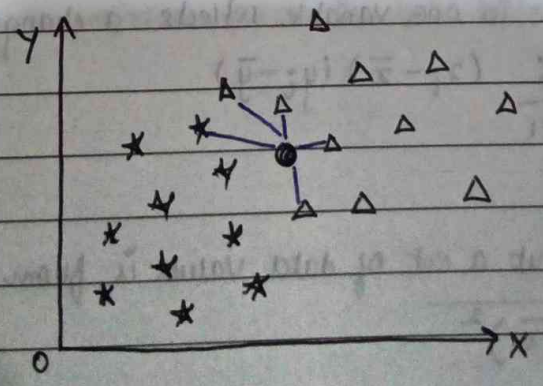
Algorithm:

(1) Begin the tree with the root node, say S, which contains the complete set.
(2) Find the best attribute using Attribute Selection Measure (ASM)
(4) ~~(3)~~ Generate the decision tree node which contains the best attribute.
(3) Divide the S into possible subsets, splitting
(5) Recursively repeat steps (2), (3), (4).

KNN (K Nearest Neighbours)

Algorithm:

(1) Select the number of K neighbours.
(2) Calculate the euclidean distance of K neighbours.
(3) Take K nearest neighbours (min. dist).
(4) Majority Voting
(5) Assign category
(6) Model complete.



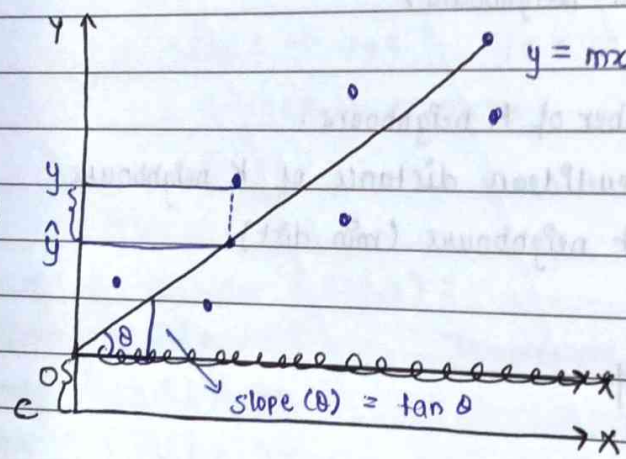
K=5

Linear Regression.

$y = mx + c$ → y intercept (constant)

- y → dependent variable
- x → Independent variable
- m → Slope $\frac{(y_2 - y_1)}{(x_2 - x_1)}$

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \dots + c$$



Covariance = change in one variable reflects a change in another.

$$\frac{1}{N} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

SD = how spread out a set of data values is from mean.

$$\sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}}$$

Correlation = dependency of one variable on another.

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \times \sum(y_i-\bar{y})^2}} \qquad \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

Variance = variability from average or mean.

$$\text{Var} = \sigma^2 = \frac{\sum(x_i-\bar{x})^2}{N}$$

Feature Selection – way of selecting a subset of most relevant features from the original feature set by removing redundant, irrelevant or noisy features (dimensionality reduction, PCA).

Validation Sets – prevent the model from overfitting, subset of the training & test sets.

Overfitting – performs well on training set but poorly on test set.

Underfitting – performs poorly on both training & test sets.

Q1]

Actual \ predicted	Positive	Negative
Positive	1880	51
Negative	40	1003

True positive = 1880 False positive = 40

True negative = 1003 False negative = 51

$$\text{Specificity / Precision} = \frac{tp}{tp + fp} = \frac{1880}{1880 + 40} = \frac{1880}{1920} = \frac{94}{96}$$

$$= 0.9791$$

$$\text{Sensitivity / Recall} = \frac{tp}{tp + fn} = \frac{1880}{1880 + 51} = \frac{1880}{1931}$$

$$= 0.9735$$

$$\text{F1 score} = 2 \times \left[\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right] = 2 \times \left[\frac{0.98 \times 0.97}{0.98 + 0.97}\right]$$

$$= 2 \times \left[\frac{0.95}{1.95}\right] = 2 \times 0.487 = 0.9743$$

Q2]

chills	runny nose	headache	fever	flu(Y/N)
Y	N	Mild	Y	?

$P(flu=Y) = \frac{5}{8}$ $\quad P(flu=N) = \frac{3}{8}$

$P(chills=Y \mid flu=Y) = \frac{3}{5}$ $\quad P(chills=Y \mid flu=N) = \frac{1}{3}$

$P(runnynose=Y \mid flu=Y) = 4/5$ $\quad P(runnynose=Y \mid flu=N) =$

Using Naive Baye's Classifier:

$P(flu=Y) = 0.625$ $\quad P(flu=N) = 0.375$

$P(chills=Y \mid flu=Y) = 0.6$ $\quad P(chills=Y \mid flu=N) = 0.333$

$P(chills=N \mid flu=Y) = 0.4$ $\quad P(chills=N \mid flu=N) = 0.666$

$P(rn=Y \mid flu=Y) = 0.8$ $\quad P(rn=Y \mid flu=N) = 0.333$

$P(rn=N \mid flu=Y) = 0.2$ $\quad P(rn=N \mid flu=N) = 0.666$

$P(headach=M \mid flu=Y) = 0.4$ $\quad P(hach=M \mid flu=N) = 0.333$

$P(hach=N \mid flu=Y) = 0.2$ $\quad P(hach=N \mid flu=N) = 0.3333$

$P(hach=S \mid flu=Y) = 0.4$ $\quad P(hach=S \mid flu=N) = 0.333$

$P(fever=Y \mid flu=Y) = 0.8$ $\quad P(fever=Y \mid flu=N) = 0.333$

$P(fever=N \mid flu=Y) = 0.2$ $\quad P(fever=N \mid flu=N) = 0.666$

→ $P(flu=Y) \times P(chills=Y \mid flu=Y) \times P(rn=N \mid flu=Y) \times P(hach=M \mid f$

$\times P(fever=Y \mid flu=Y) = 0.024$

→ $P(flu=N) \times P(chills=Y \mid flu=N) \times P(rn=N \mid flu=N) \times P(hach=M \mid flu$

$\times P(fever=Y \mid flu=N) = 0.009$

∴ Flu = YES

Q3] Inverse Matrix Approach

$$x + y = 6$$
$$2x + 4y = 20$$

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 20 \end{bmatrix}$$

$$AX = B$$

$$\therefore X = A^{-1}B$$

A X B

Calculating A^{-1}:

$\det(A) = (1 \times 4) - (2 \times 1)$

$= 4 - 2 = 2$

Since $\det(A) \neq 0$, A^{-1} exists.

$$A^{-1} = \begin{bmatrix} 4 & -1 \\ -2 & 1 \end{bmatrix}$$

$$A^{-1} \times B = \begin{bmatrix} 4 & -1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 20 \end{bmatrix} = \begin{bmatrix} 4 \times 6 - 1 \times 20 \\ -2 \times 6 + 1 \times 20 \end{bmatrix}$$

$$= \begin{bmatrix} 24 - 20 \\ -12 + 20 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$$

$$\therefore x = 4$$
$$y = 8$$

Q4]

K-Means algorithm

Dataset = { 2, 3, 4, 10, 11, 12, 20, 25, 30 }

cluster1 = 4 cluster2 = 12

Data Points	Dist. to C1	Dist to C2	Cluster	New Cluster
2	(2) min	10	C_1	
3	(1) min	.	C_1	
4	0	.	.	
10	6	.	.	
11	8	.	.	
12	1	.	.	
20	16	.	.	
25	26	.	.	
30	7	.	.	

new cluster = mean of all centroids in that cluster

Repeat the process with new centroids

Repeat until old cluster assignment & new cluster assignment turn out to be same.

Q5] Perceptron

(0.8) 0.2

(0.4) 0.1

(0.3) -0.3

0.35

$$\sigma = \frac{1}{1+e^{-x}}$$

$$z = 0.2 \times 0.8 + 0.1 \times 0.4 - 0.3 \times 0.3$$

$$= 0.16 + 0.04 - 0.09$$

$$= 0.20 - 0.09$$

$$= 0.11$$

$$\sigma = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-0.11}}$$