04/01/22

Data Warehouse and datamining

Data Warehouse:

collecting or gathering data from different sources and organized in centralized place is called data warehouse.

Olap → online analytical process

06/01/22

Mod-2 Data Preprocessing

→ The process of making the data more suitable for data mining.

→ The tasks employed in this process are informed by the process of data understanding.

(Washing the resetables → removing unwanted things & considering only essential things.)

Extracting the meaningful insights from the data which desires decision-driven making → Datamining

increase the profit & minimize the loss

To extract the insights from data → Preprocessing is must.

What is data?

- Collection of data objects and their attributes
- (or) Collection of raw facts & information.

Each instance (row) → record/object/

column → Attributes/features.

- An attribute is a property or characteristic of an object.
  - Eg: eye color of a person, temp etc.
  - Attr is also known as variable, field, characteristic, feature.
- A collection of attributes describe an object:
  - object is also known as record/point/case/sample/ entity/instance
- Absence of information in record → harmful consequence to the validity of the subsequent analysis. → missing value
- No missing values → Complete cases in records
- Missing value in record → Incomplete cases

Presence of data + correctness of data → Acc cannot be -ve

Attribute values

↓

are numbers or symbols assigned to an attribute.

Distinction bw attributes and attribute values:

Same attribute can be mapped to diff. attr. values.

Eg. height can be measured in both feet or meter.

Properties of Attribute values:

1) Distinctness = ≠
2) Order < >
3) Addition + −
4) Multiplication * /

X. eyecolor = black
Y. eyecolor = brown
X = Y
X ≠ Y
X > Y
X < Y

Types of Attributes

(=, ≠)

- Nominal (fixed set of values)
  ID numbers, eye color, zip codes, gender (M/F) [distinctness]

- Ordinal: (in order of ranks) [distinctness & order]
  Rankings (taste of potato. chips on a scale from 1-10), grades, height in {tall, medium, short}. → ordering

- Interval: [distinctness, order & addition]
  Calendar dates, temperature in celsius & Fahrenheit

- Ratio [all 4 properties]
  temp in kelvin, length, time, counts

⊗ ask in exam: Specify type of attribute & properties of attribute given dataset

10/01/2022

Discrete Attribute:

→ Countable set of values.

zip codes, counts

→ Integer values.

→ binary attr → Apl com.

Continuous Attribute.

- real no.
- temp, height, weight
- It is Measured

Record

- data matrix
- Tables
- document data

↓

Each doc → term

Each term → component

Each comp → no of times

word term occur.

No of occurences of words in documents.

- Transaction data
- Graph data

Node → no of users using social net.

Vertices → friend links.

- Chemical data. C_6H_6

- Ordered data

Sequence of transaction items/events

(AB) (D) (CE)

Stream of data.

- Genomic sequence data.

A G C T

Adenine Guanine Cytosine thymine

- Spatio-temporal data

Avg. monthly temp.

Data quality problems

- Noise & Outliers
- Missing values
- duplicate data

Outliers

data objects with characteristics considerably diff from most of the other data objects in the data set.

Missing values (Absence of data)

- Replaced by possible values
- Eliminate data objects
- Estimate missing values
- Ignore it.

Reason

- Attr. may not be applicable to all cases (salary for children x)
- Info not collected.

Duplicate data

Redundant data record

Sample dataset

Instance :

features :

Training i/p :

Testing o/p

Data Preprocessing Tasks

- Data cleaning (removing unwanted / noisy data)
- Data Transformation (changing data format)
- Data reduction (reducing dimensions)
- Data Discretization

  ↓ (Numerical to nominal datatype conversion)

  dividing the data

  discretize into intervals

Data smoothing

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 34

Equal frequency (equi-depth) bins :

Bin 1 : 4, 8, 9, 15

2 : 21, 21, 24, 25

3 : 26, 28, 29, 34

Smoothing by bin means : (mean value)

Bin 1 : 9, 9, 9, 9

2 : 23, 23, 23, 23

3 : 29, 29, 29, 29

Smoothing by bin boundaries (exact min, max value ; all are replaced with closer

Bin 1 : 4, 4, 4, 15

2 : 21, 21, 25, 25

3 : 26, 26, 26, 34.

11/01/22 - lab

i/p : @ RELATION <name>

@ ATTRIBUTE <attribute name> <ATTR. NU

!!

o/p : @ ATTRIBUTE class {< >,< >

class → Nominal

13/01/22

Mean with outlier (m_1) — difference — Mean without outlier (m_2)

$m_1 - m_2 \neq$ large ↓ remove outlier

Outlier → datapoint not close to other data.

Technique: IQR | boxplot used to represent

Missing values. → equivalent zero.

3	10
6	x
9	30
Y	40

Soln

$$\frac{x}{6} = \frac{10}{3}$$

$$x = \frac{10}{3} \times 6 = \frac{10}{3} \times \frac{6}{1} = 20$$

$$\frac{y}{40} = \frac{9}{30}$$

$$y = \frac{3}{10} \times \frac{40}{1} = 12$$

(2)

Buses	Drivers
63	28
9	x

$$\frac{9}{x} = \frac{63}{28}$$

$$\frac{9}{x} = \frac{9}{4}$$

$x = 4$

Feature selection

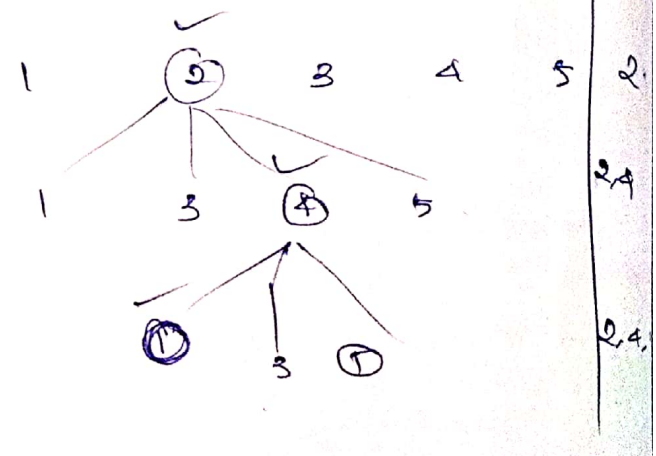
· Sequential Forward Selection.

Find the feature f_1 that gives the best performance.

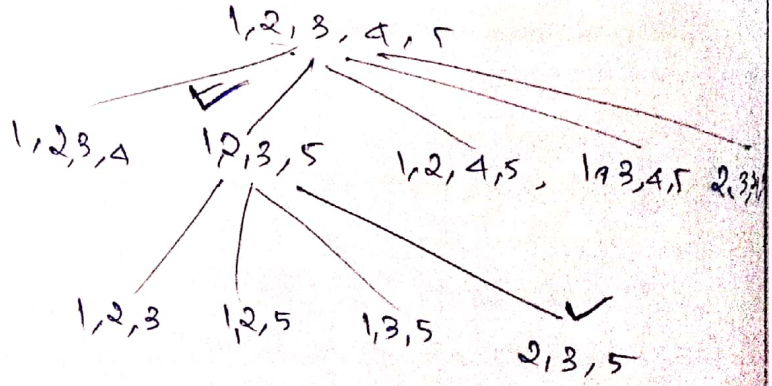
Find the feature f_2 such that (f_1, f_2) give the best performance

Repeat for as many feature as desired

Q5 find the best 3 / 5 features.



· Sequential Backward Elimination.



Data reduction of

Age from 47 individuals

17/01/22

Data discretization in data mining

→ converts a large no. of data values into smaller ones, so that data evaluation & data management becomes easy.

Age → 10, 11, 13, 14, 17, 19, 30, 32,

Young | Mature | old.

Binarization

Mapping a categorical attribute to a set of attribute that are binary.

Ordinal attribute:

{small, Medium, large} →

00, 01, 10, 11

could take any 3/4 to represent.

Qn. How

Ans: 3

$\overline{\log_2(x)}$ = ceiling of $\log_2(3) = \overline{\log_2(3)} = 3$

Issues

x_2 & x_3 are now correlated bcz 'good' is encoded only both

Actual	0	0	0	0
poor	1	0	0	1
ok	2	0	1	0
good	3	0	1	1
Great	4	1	0	0

5 categorical values → 3 binary vars.

Data reduction:-

Age of 47 individuals,

Range = 69 − 20 = 49

N = 5 (No. intervals)

width of class = 49/5 = 9.8 = 10

class intervals = 20-29, 30-39, 40-49, 50-59, 60-69

data reduction ↓

Continuous 20, 21, 22, .. 69

↓

Interval 20-29, 30-39

↓

Ordinal twenties, thirties

↓

Nominal Young's, Mature, old

not data reduction ↑

Data transformation: Normalization:

- Min max
- Z score
- by decimal scaling

Min max (new_min, new_max)

$$v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$$

Let Income range.

Eg $12,000 to $98,000 normalized to [0.0 - 1.0].

$73,600 is mapped to

$$\frac{73,600 - 12,000}{98,000 - 12,000}(1 - 0) + 0 = 0.716$$

68,000

$$\frac{68000 - 12000}{98000 - 12000} * (1-0) + 0$$

$$= \frac{56000}{86000} * 1$$

$$= 0.6511.$$

Z-score (μ: mean, σ: std)

$$v' = \frac{v - \mu}{\sigma}$$

Eg $\mu = 54,000$

$\sigma = 16000$

then $\frac{73,600 - 54,000}{16,000} = 1.225.$

Normalization by decimal scaling:

$$v' = \frac{v}{10^j}$$

j → smallest integer such that $Max(|v'|) < 1$

22/01/25

Covariance

	Hours(H)	Mark (M)
data	9	39
	15	56
	25	93
	14	61
	10	50
	18	75
	0	32
	16	85
	5	42
	19	70
	16	66
	20	80
total	167	749
Avg:	13.92	62.42

$$Var(x): \sum_{i=1}^{n} \frac{(x_i-\bar{x})(x_i-\bar{x})}{(n-1)}$$

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

positive :- $x\uparrow y\uparrow$, $x\downarrow, y\downarrow$

Negative :- $x\uparrow y\downarrow$, $x\downarrow, y\uparrow$

H	M	$(H_i-\bar{H})$	$(M_i-\bar{M})$	$(H_i-\bar{H})(M_i-\bar{M})$
9	39	-4.92	-23.42	115.23
15	56	1.08	-6.42	-6.93
25	93	11.08	30.58	338.83
14	61	0.08	-1.42	-0.11
10	50	-3.92	-12.42	48.69
18	75	4.08	12.58	51.33
0	32	-13.92	-30.42	423.45
16	85	2.08	22.58	46.97
5	42	-8.92	-20.42	182.15
19	70	5.08	7.58	38.51
16	66	2.08	3.58	7.45
20	80	6.08	17.58	106.89
tot.				1149.93
Avg:				104.54

$$Cov_{xy} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{N}}{N-1}$$

$$r = \frac{Cov_{xy}}{S_x S_y}$$

$r \Rightarrow$ [~~0 to 1~~] -1 to $+1$

~~$r = n\Sigma(xy) - \Sigma(x)\Sigma(y)$~~

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

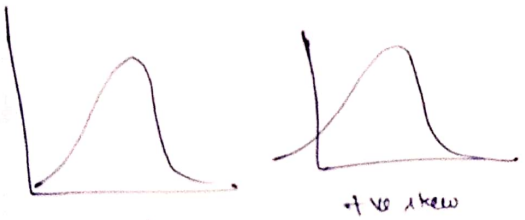
Forward features selection
its implementation

Variable transformation

In data modelling, transformation refers to the replacement of a variable by a function.

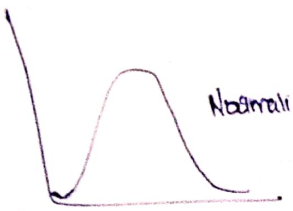
for instance, replacing a var x by square / cube root or log x is a transformation

"Transformation is a process that changes the distribution or relationship of a variable with others."



-ve skew +ve skew

transformation



Normalized

(10,12) (2,2)

$Sim(A,B)$

$= \frac{10\times2 + 12\times2}{\sqrt{100+144}\times\sqrt{4+4}} = \frac{20+24}{\sqrt{244}\times8}$

$= \frac{44}{244.18} = 0.995 \simeq 1$

Similarity & Dissimilarity

Similarity

- Numerical Measure of how alike two data obj are.
- Higher when obj are more alike ~~often 1~~
- often falls in range [0,1]

$Sim(x_1, x_2) = \frac{0.9}{0.3}$

based on similarity consider any one attr. if both are similar.

Dissimilarity:

- Numerical measure of how diff. how are 2 data obj.
- Lower when objects are more alike
- Minimum dissimilarity often 0.
- Upper limit varies.

Proximity refers to a similarity or dissimilarity

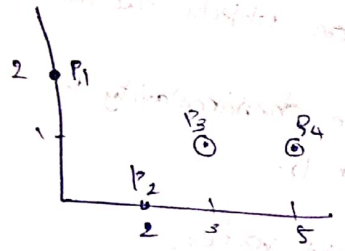
dissimilarity = 1 - similarity

$\text{Similarity}(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum A_i^2} \times \sqrt{\sum B_i^2}}$

$\text{Similarity}(x,y) = \cos\theta = \frac{x \cdot y}{\|x\|\|y\|}$

Attribute type	dissimilarity	similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p - q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, n = no of values)	$s = 1 - \frac{\lvert p - q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$, or $s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$

Euclidean distance



$$dist = \sqrt{\sum (p_k - q_k)^2}$$

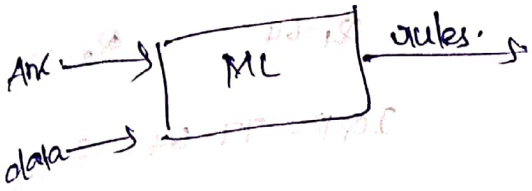
Point	x	y
p_1	0	2
p_2	2	0
p_3	3	1
p_4	5	1

	p_1	p_2	p_3	p_4
p_1	0	2.828	3.162	5.099
p_2	2.828	0	1.414	3.162
p_3	3.162	1.414	0	2
p_4	5.099	3.162	2	0

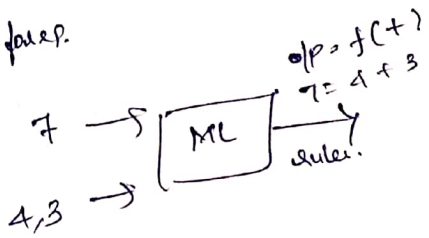
distance matrix

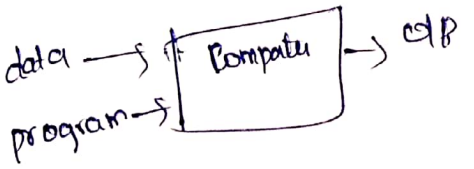
Regular programming

Machine Learning

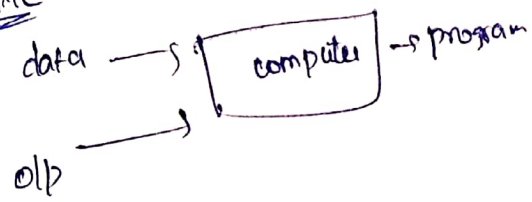


for eg.





ML



ML - Machine Learning :-

Learning is any process by which a system improves performance from experience (data).

— Herbert Simon

ML is the study of algorithms that improve their perfo P at some task T.

Field of study that gives computer ability to learn without being explicitly programmed.

— Arthur Samuel (1959)

Types of learning :-

Supervised (inductive) learning
training data + desired o/p (labels)

Unsupervised learning
training data (without desired o/p)

Semi-supervised learning
training data + a few desired o/p.

Reinforcement learning.
Rewards from sequence of actions.

Lab

No. of animal characters in each kid's book.

(1) 4, 4, 10, 11, 15, 7, 14, 12, 6 IQR = ?

Sorting

4, 4, 6, 7, 10, 11, 12, 14, 15,

Median, $Q_2 = \frac{N+1}{2}^{th}$ $N = 9$

$= 9\frac{1}{2} = 5^{th}$ element.

$Q_2 = 10$

4, 4, 6, 7

Q_1 = Median of this set = $\frac{4+6}{2} = \frac{10}{2}$

$Q_1 = 5$

11, 12, 14, 15

Q_3 = Median = $\frac{12+14}{2} = \frac{26}{2} = 13$

$Q_3 = 13$

IQR = $Q_3 - Q_1$

$= 13 - 5 = 8$

IQR = 8

Upper fence = $Q_3 + 1.5 \times IQR$

$= 13 + 1.5 \times 8$

$= 13 + 12 = 25$

lower =

$Q_1 - 1.5 \times IQR$

$5 - 1.5 \times 8$

$5 - 12 = -7$

lower = -7

(2)

71

62 63 64 64 70 72

$Q_1 = 64$ $Q_3 = 7$

IQR = 77 - 64 = 13

07/01/2022

ML

i/p x o/p y

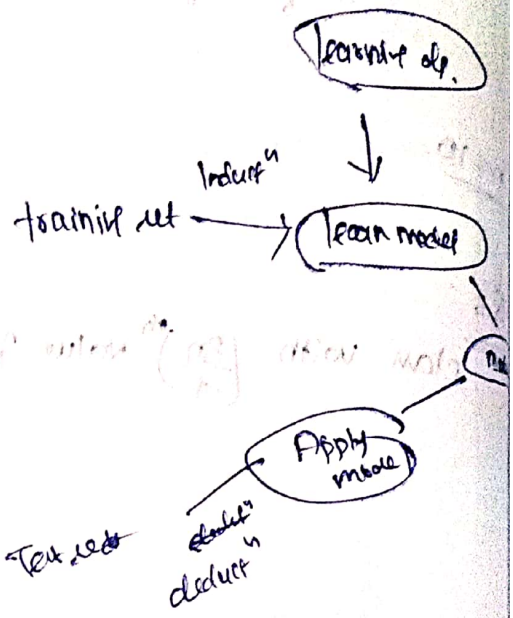
bike label 'motorcycle' (Supervised)

Classification:-

Classification is a process of categorizing a given set into classes, it can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

Ex. Mail spam/not spam.

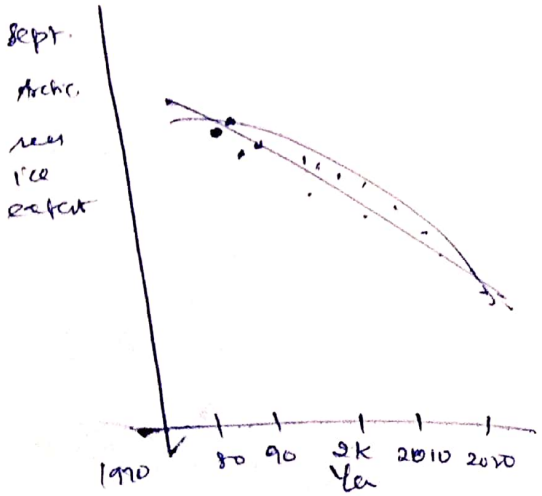
human decision making - classification



Supervised learning: Regression

f(x) is predict y given x

y is real valued = regression



y = mx + c

Evaluation of classification models

Counts of test records that correctly predicted by the classification model, Confusion matrix

Predicted class

Actual class	class = 1	class = 0
class = 1	f_{11}	f_{10}
class = 0	f_{01}	f_{00}

Actual

Predicted	Positive 1	Negative 0
Positive 1	TP	FP
Negative 0	FN	TN

$$\text{Accuracy} = \frac{\#\text{ correct predictions}}{\text{total \# of predictions}}$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Accuracy} = \frac{(TN + TP)}{TN + FP + FN + TP}$$

$$\text{Error rate} = \frac{\#\text{ wrong predictions}}{\text{total \# of predictions}}$$

$$= \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Precision} = \frac{TP}{FP + TP}$$

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

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

Measure	Formula
Recall	$\frac{TP}{P}$
F measure	$\frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{Recall})}$

Accuracy

Recall (sensitivity)

(Sensitivity)

True positive recall

$$\frac{TP}{TP+FN}$$

Predicted \ Actual	TP	FP
	FN	TN

False positive recall

$$\frac{FP}{FP+TN}$$

Same

Precision

Actual

Predicted	TP	FP
	FN	TN

Precision $\rightarrow \frac{TP}{TP+FP}$

True Negative Rate

(Specificity)

$$\frac{TN}{TN+FP}$$

Actual

Precision	TP	FP
	FN	TN

False Positive rate

(1 - specificity)

$$\frac{FP}{FP+TN}$$

01/02/22

Decision Tree classifier:-

→ A classification scheme which generates a tree and a set of rules from given data set.

→ the set of records available for developing classification methods is divided into 2 disjoint subsets - a training set
- testing set.

—

Inner node → Attribute

edge → test on father node attr.

leaf → class.

— Top down strategy

— Based on training data.

—

DT- ID3 algorithm
- Entropy [large]
- Gain (high value)

DT- CART
- Gini (small).

Naive Bayes Model

$$P(C_k|x) = \frac{P(C_k) * P(x|C_k)}{P(x)}$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$P(C_k|x)$ is the posterior probability of class k, given predictor x.

$P(C_k)$ → prior prob. of class k.

$P(x)$ → prior prob. of predictor.

$P(x|C_k)$ → current likelihood of predictor given class.

Eg

Outlook = Sunny, Temp = Cool, Humidity = high, Wind = Strong.

$P(\text{Playtennis} = \text{Yes}) = 9/14 = 0.64$

$P(\text{Play} = \text{No}) = 5/14 = 0.36$

Outlook	Y	N
Sunny	2/9	3/5
Overcast	4/9	0
rain	3/9	2/5

Temp	Y	N
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
High	3/9	4/5
Normal	6/9	1/5

Windy	Y	N
Strong	3/9	3/5
Weak	6/9	2/5

$$V_{NB} = \text{argmax}_{P(v_j)} \prod_i P(a_i|v_j)$$

$$= \text{argmax}\, P(v_j)$$

$v_j \in [\text{Yes}, \text{No}]$

$P(\text{Outlook} = \text{Sunny}|v_j)\, P(\text{Temp} = \text{cool}|v_j)$

$P(\text{Humidity} = \text{high}|v_j)\, P(\text{Wind} = \text{Strong}|v_j)$

$V_{NB}(\text{Yes}) = P(\text{Yes})\, P(\text{Sunny}|\text{Yes})$

$P(\text{cool}|\text{Yes})\, P(\text{high}|\text{Yes})$

$P(\text{Strong}|\text{Yes})$

$= 0.0053$

$V_{NB}(\text{No}) = P(\text{no})\, P(\text{Sunny}|\text{no})$

$P(\text{cool}|\text{no})\, P(\text{high}|\text{no})$

$P(\text{Strong}|\text{no})$

$= 0.0206$

$V_{NB}(\text{No})$ is higher.

$$V_{NB}(\text{Yes}) = \frac{V_{NB}(\text{Yes})}{V_{NB}(\text{Yes}) + V_{NB}(\text{No})}$$

$= 0.205$

$$V_{NB}(\text{No}) = \frac{V_{NB}(\text{No})}{V_{NB}(\text{Yes}) + V_{NB}(\text{no})}$$

$= 0.795$ ✓

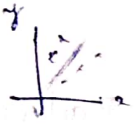
Prob. of $V_{NB}(\text{No})$ is high, so given 90 can be No.

Regression Analysis

Study of 2 variables in an attempt to find a relationship, a correlation.

Linear Regression $Y = \beta_0 + \beta_1 x$

estimate β_0, β_1



x	y
-2	-1
1	1
3	2

$Y = ax + b$

slope, intercept

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{1}^{n} x_i \sum_{1}^{n} y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{1}{n}\left[\sum y_i - a\sum x_i\right]$$

X	Y	XY	X^2
-2	-1	2	4
1	1	1	1
3	2	6	9

$\Sigma X = 2 \quad \Sigma y = 2 \quad \Sigma xy = 9 \quad \Sigma x^2 = 14$

Calculate value of a

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

$$= \frac{3(9) - 2\times 2}{3\times 14 - 4} = \frac{24-4}{42-4}$$

$$a = \frac{23}{38}$$

$$b = \frac{1}{n}\left[\Sigma y - a\Sigma x\right]$$

$$= \frac{1}{3}\left[2 - \frac{23}{38}\times 2\right]$$

$$= \frac{2}{3}\left[\frac{38-23}{38}\right]$$

$$= \frac{2}{3}\times\frac{15}{38}$$

$$b = \frac{5}{19}$$

$Y = ax + b$

$$Y = \frac{23}{38}x + \frac{5}{19}$$