OSE 303 pmim sir MID01 Dmim sir

Date : Page : Hossain Mim 2021-1-60-07

(P-P1) 60, 70, 80, 75, 65, 70, 80, 70, 65, 65

(Ans) sorted Order :— 60, 65, 65, 65, 70, 70, 70, 75, 80, 80

Here, $n = 10$

$$\text{mean}(\bar{x}) = \frac{\sum x_i}{n}$$

$$= \frac{700}{10}$$

$$= 70 \quad \text{(Ans)}$$

$n = 10$

$$\text{median} : \frac{\left(\frac{n}{2}\right)\text{th value} + \left(\frac{n}{2}+1\right)\text{th value}}{2}$$

$$= \frac{5\text{th value} + 6\text{th value}}{2}$$

$$= \frac{70+70}{2}$$

$$= 70 \quad \text{(Ans)}$$

MID 01 part 02

(P-P2)

60, 70, 80, 75, 65, 70, 100, 70, 65, 65

$$\text{mean}(\bar{x}) : \frac{720}{10}$$

$$= 72 \quad \text{(Ans)}$$

For Trimmed mean :—

sorted the dataset : 60, 65, 65, 65, 70, 70, 70, 75, 80, 100

For P = 1 so, 60 and 100 will be discard.

$$\text{mean}(\bar{x}) = \frac{\sum_{i=p+1}^{n-p} x_i}{(n-2p)}$$

$$= \frac{560}{10-2}$$

$$= \frac{560}{8}$$

$$= 70 \quad \text{(Ans)}$$

weighted mean :— $x_i = 40, 45, 80, 75$ and 10 have

$w_i = 1, 2, 3, 4$ and 5

$$(\bar{x}_w) = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i}$$

$$= \frac{(40\times1)+(45\times2)+(80\times3)+(75\times4)+(10\times5)}{1+2+3+4+5}$$

$$= \frac{720}{15} = 48 \quad \text{(Ans)}$$

MID 01 part 02

Hossain Mim 2021-1-60-07

mean Absolution deviation :—

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$

Dataset 1 : 30, 40, 50, 60, 70

$\bar{x} = 50$

$$MAD :- \frac{|(30-50)| + |40-50| + |50-50| + (60-50) + |70-50|}{5}$$

$$= \frac{20+10+0+10+20}{5}$$

$$= \frac{60}{5}$$

$$= 12 \quad \text{(Ans)}$$

Date :
Page :

median Absolute divation:—

$MAD = \text{median}(|x_1 - m|, |x_2 - m|, \ldots |x_N - m|)$

Dataset :(2) 0 ; 25 , 50, 75 , 100 $n = 5$ (odd number)

Ans: MAD: median $(|0-50|, |25-50|, |50-50|, |75-50|, |100-50|)$ $\left(\frac{n+1}{2}\right)$th

$= \text{median}(50, 25, 0, 25, 50)$ $= \left(\frac{5+1}{2}\right)\left(\frac{6}{2}\right)$ 3th valu

$= \text{median}(0, 25, 25, 50, 50)$

middle একটা

$n = 5$ (odd)

$= \left(\frac{n+1}{2}\right)$th value

$= \left(\frac{5+1}{2}\right)$ th

$= 3$th value

$= 25$ Ans

variance: (6²)

$6^2 = \frac{(x - \bar{x})^2}{n}$

Example:— The Highet : 600 mm, 470 mm, 170 mm, 430 mm, 300 mm

mean $(\bar{x})$: $\frac{1970}{5} = 394$

variance (6^2): $\frac{(600-394)^2 + (470-394)^2 + (170-394)^2 + (430-394)^2 + (300-$

$= 21704$ Ans

middle আরেকটা

standar deviation :(6) $= \sqrt{21704}$

$= 147.32$ Ans

sample variance (s^2) :—

Dataset 1: 30, 40, 50, 60, 70

$n = 5$

$\bar{x} = \frac{250}{n} = \frac{250}{5} = 50$

$s^2 = \frac{(30-50)^2 + (40-50)^2 + (50-50)^2 + (60-50)^2 + (70-50)^2}{5-1}$

$= \frac{1000}{4}$

$= 250$ Ans

Standard deviation : $s = \sqrt{s^2}$

$= \sqrt{250}$

$= 15.82$ Ans

Dataset 2 : 0, 25, 50, 75, 10

$\bar{x} = \frac{250}{n}$

$= 50$

$s^2 = (0-50)^2 + (25-50)^2 + (50-50)^2 + (75-$
$(100-50)^2$

$5-1$

$= \frac{6250}{4}$

$= 1562.5$

Standard deviation:—

$s = \sqrt{1562.5}$

$= 39.53$

Ans

Date :
Page :

Practice Example:—

Dataset:— 8, 9, 10, 10, (10), 11, 11, 11, 12, 13

mean ($\bar{x}$) : $\frac{105}{10}$

$= 10.5$ (Ans)

n = 10 (Even number)

median = $\frac{(\frac{n}{2})^{th} \text{ value} + (\frac{n}{2}+1)^{th} \text{ value}}{2}$

$= \frac{10+11}{2}$

$= 10.5$ (Ans)

mode : (10, 11) Bi-modal dataset (Ans)

Range : (max − min)

$= (13 - 8)$

$= 5$ (Ans)

Hossain Mim
2021-1-60-071

Variance

$\sigma^2 = \frac{(8-10.5)^2 + (9-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (12-10.5)^2 + (13-10.5)^2}{10}$

$= 1.85$

$\sigma = \sqrt{1.85}$

$= 1.36$ (Ans)

MID 01

Sample variance

$s^2 = \frac{(8-10.5)^2 + (9-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (12-10.5)^2 + (13-10.5)^2}{(10-1)}$

$= 2.055$

$s = \sqrt{2.055}$

$= 1.43$ (Ans)

Mean Absolute Deviation:— $MAD = \frac{|8-10.5| + |9-10.5| + |10-10.5| + |10-10.5| + |10-10.5| + |11-10.5| + |11-10.5| + |11-10.5| + |12-10.5| + |13-10.5|}{10}$

$= \frac{5+3+2}{10} = \frac{11}{10} = 1.1$ (Ans)

median Absolute Deviation:—

MAD: median $(|8-10.5|, |9-10.5|, |10-10.5|, |10-10.5|, |10-10.5|, |11-10.5|, |11-10.5|, |11-10.5|, |12-10.5|, |13-10.5|)$

n = 10

$(\frac{n}{2})^{th}$ value $+ (\frac{n}{2}+1)^{th}$ value

$\frac{n}{2}$ $\frac{0.5+0.5}{2}$

= median {2.5, 1.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 2.5}

= median (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 2.5, 2.5)

(Ans)

Date :

Page :

Finding percentile:— 8, 9, 10, 10, 10, 11, 11, 11, 12, 13, 15, 16, 16, 17, 20

Total data point $n = 15$

25th percentile $= 25\% \times 15$
$= 3.75$ th value
$= 10 + (10-10) \times 0.75$
$= 10$ Ans
Q_1

50 percentile $= 50\% \times 15$
$= 7.5$ th value
$= 11 + (11-11) \times 0.5$
$= 11$ Ans
Q_2

75th percentile $= 75\% \times 15$
$= 11.25$ th value
$= 15 + (16-15) \times 0.25$
$= 15.25$ Ans
Q_3

IQR $= (Q_3 - Q_1)$
$= (15.25 - 10)$
$= 5.25$ Ans

(Hossain min)

95th percentile $= 95\% \times 15$
$= 14.25$ th value
$= 17 + (20-17) \times 0.25$
$= 17.75$ Ans

BOX Plot

64, 64, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85

Min = 64
max = 85

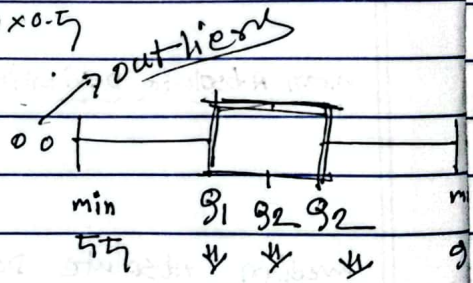
Q_1 = 25th percentile $= 25\% \times 14$
$= 3.5$ th value
$= 68 + (69-68) \times 0.5$
$= 68.5$

Q_3 = 75th percentile $= 75\% \times 14$
$= 10.5$ th value
$= 75 + (80-75) \times 0.5$
$= 77.5$

IQR: $(77.5 - 68.5)$
$= 9$

upper Extreme: $Q_3 + IQR \times 1.5$
$= (77.5 + 9 \times 1.5)$
$= 91$

Lower Extreme: $Q_1 - IQR \times 1.5$
$= 55$



Any values greater the upper Extreme and smaller th Lower Extreme would be called outliers

Data: 0, 5.6, 8.7, 14.1, 14.1, 15, 17.2, 19.2, 19.3, 24.1, 24.7

max: 24.7
min: 0

IQR: $Q_3 - Q_1 \Rightarrow (19.225 - 7.925) \Rightarrow 11.3$

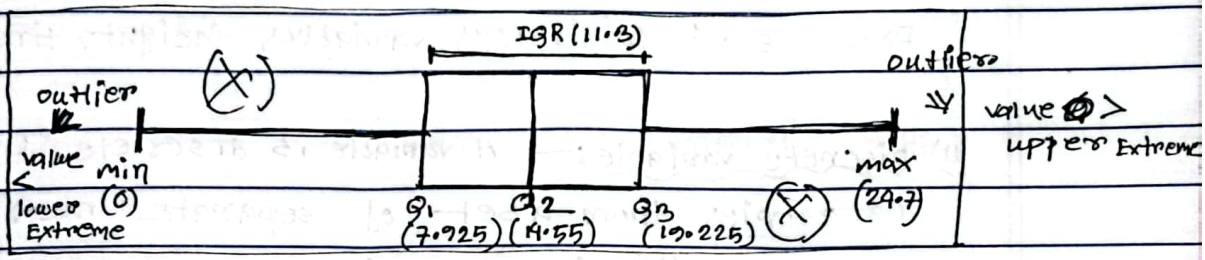
$Q_1 = 25\% \times 11$
$= 2.75$ th value
$= 5.6 + (8.7 - 5.6) \times 0.75$
$= 7.925$

$Q_3 = 75\% \times 11$
$= 8.25$ th value
$= 19.2 + (19.3 - 19.2) \times 0.25$
$= 19.225$

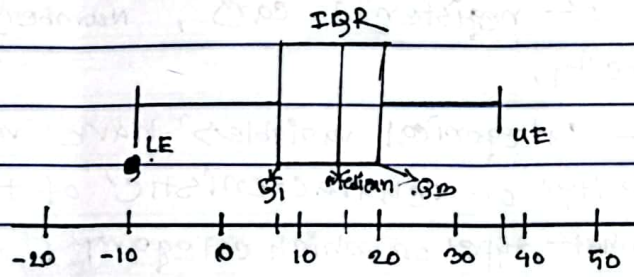
$Q_2 = 50\% \times 11$
$= 5.5$ th value
$= 14.1 + (15 - 14.1) \times 0.5$
$= 14.55$

Upper Extreme: $Q_3 + IQR \times 1.5$
$= 19.225 + 11.3 \times 1.5$
$= 36.175$

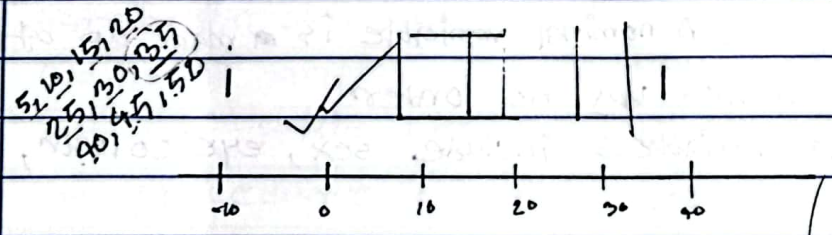
Lower Extreme: $Q_1 - IQR \times 1.5$
$= 7.925 - 11.3 \times 1.5$
$= -9.025$

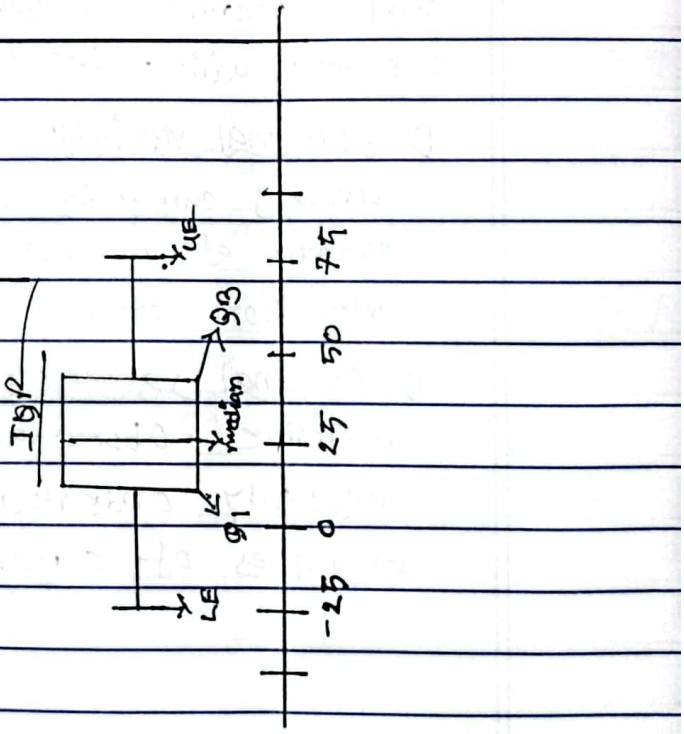


Box Plot



Hossain Mim
2021-1-60-071





$Q_1 = 25\% \times 10$
$= 2.5$ th value
$= 10 + (15 - 10) \times 0.5$
$= 12.5$

$Q_2 = 50\% \times 10$
$= 5$th value
$= 25$

$Q_3 = 75\% \times 10$
$= 7.5$ th value
$= 35 + (40 - 35) \times 0.5$
$= 37.5$

$UE = (Q_3 + 1.5 \times IQR)$
$= 75$

$LE = (Q_1 - 1.5 \times IQR)$
$= -25$

$IQR = (37.5 - 12.5)$
$= 25$

Date :
Page :

variable: A variable is any characteristic, number or quantity that can be measured. A variable may also be called a data item. Age, sex, eye colour, country of birth etc are examples of variable.

Important

Numerical variables:— Numerical variable have values that describe a measurable quantity as a number, like 'how many' or 'How much'.

i) **Continuous variable:—** A continuous variable is a numeric variable. Observations can take any value between a ~~cert~~ certain set of real numbers.

Examples of continuous variables height, time, age etc.

ii) **Discrete variable:—** A variable is discrete if it possible categories from a set of separate numbers

Examples of ~~does~~ discrete variables, numbers of children in a family, Number of registered cars, Numbers of boy in the ~~schoole~~ university.

categorical variables:— categorical variables have values that describe a 'quality' or 'characteristic' of the the data unit like 'what type' or which category'.

i) **Nominal variable:—** A nominal variable is made up of various categories which has no order

Example of nominal variables include sex, eye colour, religion etc.

ii) **Ordinal variable:—** An ordinal variable is a categorical variable. Observation can take a value that can be logically ordered or rank.

Examples of ordinal variables include

Hassnain
2021-1-60-071

Date :
Page :

There are mainly seven steps involved into it :—

I) Decide on the objectives :— These objectives may usually require significant data collection and analysis.

II) what to measure and how to measure :— measurement generally refers to the assigning of the numbers to indicate different values of variables

Important

III) Data collection :— once you know what type of data you need for your statistical study then you can determine whether your data can be gathered.

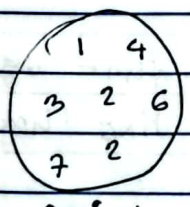
IV) Data cleaning :— This is another crucial setp in data analysis pipeline is to improved data quality for your existing data.

V) summarizing :— Exploratory data analysis helps to understand the data better

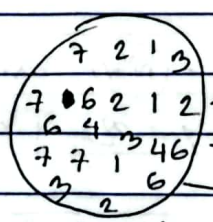
VI) Data modeling :—

VII) Optimize and Repeat :— The data analiysis is a repeatable process and sometime leads to continuous improvements, both to the business.

The Bootstrap :— one easy and effective way to eastimate the sampling distribution of a statistic of a model parameters, is to draw additional samples, with replacement from the sample itself and recalculate the statistic or model each resample. This procedure is called bootstrap and it dose not necessarily involve any assumption about the data or sample statistic ~~beign~~ being normally distributed.



Orginal sample



sample relicated a huge number of times

Draw lots of resample

MIDOL analysis

Date :
Page :

The algorithm for a bootstrap resampling of the mean is as follows for a sample of size n :—

1) Draw a sample value, record, replace it
2) Repeat n times
3) Record the mean of the n resampled values
4) Repeat steps (1-3) R times
5) use the R results to :—
   i) calculate their standar deviation
   ii) produce a histogram or boxplot
   iii) Find a confidence interval.

Confidence Interval :— A (CI) is statistics is a rage of values thats used to estimate an unknown population parameter, such as the mean or proportion. It provides a range of plausible values for the parameter and is calculate from sample data.

For example 95% (CI) for mean weight of a population might be from 60kg to 70kg, indicating that we're 95% confident that the true population mean weight falls with in this rage.

The width of the interval is influenced by the sample size and the desired level of confidence.

variable :— A variable is any characteristic, number, quantity that can be mesured. A variable also be called a data item. Age, sex, eye colour, country of birth etc are Example of variable.

1) Numerical variables :— (Nv) have values that describe a measurable quantity as a numbe like 'how many' or 'How much'

   i) continuous variable :— A (CN) is a numerical variable. observations can take any value between a certain set of real number. Example of continuous variable include weight, time, age.

7.3

Standard Error:— Dataset : 85, 90, 88, 92, 87

Mean ($\bar{x}$) $= \frac{442}{5}$

$= 88.4$

$s^2 = \frac{(85-88.4)^2 + (90-88.4)^2 + (88-88.4)^2 + (92-88.4)^2 + (87-88.4)^2}{5-1}$

$= 7.3$

$s = \sqrt{7.3}$

$= 2.073$

$SE = \frac{s}{\sqrt{n}}$

$= \frac{2.73}{\sqrt{5}}$

$= 1.220$ Ans

ordinal:— observation can take a value that can be logically ordered and rank. Example ordinal variable include Health quality (good, adequate, bad)

Example:— Random sampling:— dataset {10, 8, 3, 7, 1, 2, 5, 6, 9, 4}

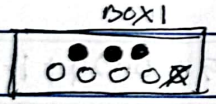
sample size = 4

sample 1 (with replacement) 2, 2, 4, 6 $= \frac{2+2+4+6}{4}$ with out

$= 3.5$ Ans

sample 2 (with replacement) 9, 10, 10, 8 $= \frac{9+9+10+8}{4}$

$= 9$ Ans



Find the probability of the picking two Black ball from Box1 when you draw the samples without replacement?

P(1st ball is Black $\frac{3}{7}$) × P(2nd ball is black $\frac{2}{6}$)

$\Rightarrow \frac{3}{7} \times \frac{2}{6}$

$= \frac{6}{42}$ Ans

You draw the samples with replacement

$\frac{3}{7} \times \frac{3}{7}$

$= \frac{9}{49}$ Ans

$H_0 \rightarrow$
$H_1 \rightarrow$

$\frac{RT \times CT}{TOB}$

Date :
Page :

(Ans)

H_0 : There is no link between gender and political party

H_1 : There is link between gender and political party

Now,

The Expected value for male republican $= \frac{290 \times 200}{420}$

$= 114$

The Expected value for female republican $= \frac{290 \times 220}{420}$

$= 125.72$

The Expected value for male demorat $= \frac{200 \times 130}{420}$

$= 61.91$

The Expected value for female demorat $= \frac{220 \times 130}{420}$

$= 68.01$

The Expected value for male Independant $= \frac{200 \times 50}{420}$

$= 23.81$

The Expected value for female Indepent $= \frac{220 \times 50}{420}$

$= 26.19$

- $x_c^2 = \frac{(O_i - E_i)^2}{E_i}$

x^2 is the sum of the all value : $(1.72 + 1.63 + 1.05 + 8.95 + 1.61 + 1.78)$

$= 8.74$

we have

$\{(r-1) \times (c-1)\}$

$\Rightarrow \{(2-1) \times (3-1)\}$

$= (1 \times 2)$

$= 2$

Hossain mim
2021-1-60-027

As we see that for alpha level of 0.05 and two degrees of freedom the critical statistic is 5.991 which is less then our obtain statistic of 8.74

obtain statistic > critical statistic

H_0 can be rejected.

(Ans)

যদি obtain statistic > α value
তবে > হবে
H_0 reject

যদি obtain statistic < α
তবে
H_0 do not rejected

A confidence interval is satisfitsite is a range of values that used to estimate an unknown population parameter, such as true mean or proportion. Date:

It provides a range of plausible values for the parameters and is calculated from sample data.

95% ⇒ It means that should we repeat an experiment or survey over and over again, 95% of the time our results will match the results, we got from a ~~prop~~ population.

self-selection bais in ~~stat~~ statistic occurs when individuals or entities self-selection into a sample being studied, leading to a non-random or biased representation of the population.

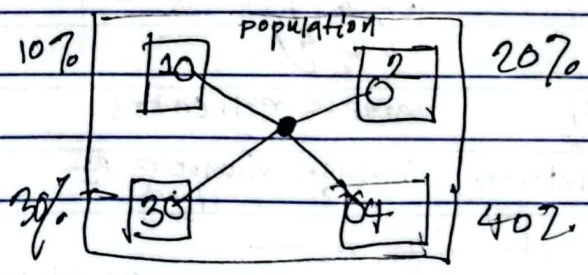
An Example self-selection bias in political polls.

consider a situation where a poll about a certain candidate anonline plat from people who actively seek out and participate in such poll might have stronger opinions or be more engaged it the topics. compared to general population.

Important

stratified sampling is a sampling technige in statistic where the population is divided into distinct subgroups, called strata, based on certain characteristics that are relevant to the study.

The strata are homogenous whit in themselves but different from ~~eet~~ each other in some key aspects



1st

Date :

Page :

10, 15, 18, 22, (25) 28, 33, 43, 50

$n = 9$ It is odd number.

median = $\frac{n+1}{2}$ th value

= $\frac{10}{2}$ th value

= 5 th value

= 25 (Ans)

10, 15, 18, 22, (25) (28), 30, 33, 50, 55

$n = 10$ It is the even number

median = $\frac{(\frac{n}{2}) \text{ th value} + (\frac{n}{2}+1) \text{ th value}}{2}$

= $\frac{5 \text{ th value} + 6 \text{ th value}}{2}$

= $\frac{25+28}{2}$

= 26.5 (Ans)

(5), 6, (5), 7, (5), 8, 9, (5)

mode: 5 [here, 5 is most frequent item (Ans)

(5), (6), (5), (6), (5), 8, (6), (5), (6)

mode: 5 and 6 [Here, 5 and 5 both are most frequent item] (Ans)

(60), (70), (80), (75), (65), (70), (100), (70), (65), (65)

sorted data: 60, 65, 65, 65, 70, 70, 70, 75, 80, 100

mean = $\frac{720}{10}$

= 72

trimmed mean = $\frac{560}{10-(2\times1)}$

= $\frac{560}{8}$ [P=1]

= 70 Ans (Ans)

(60), (70), (80), (75), (65), (70), (10), (70), (65), (65)

sorted data: 10, 60, 65, 65, 65, 70, 70, 70, 75, 80

mean = $\frac{630}{10}$

= 63

trimmed mean = $\frac{540}{10-(2\times1)}$ [P=1]

= 67.5 (Ans)

$x_i = 40, 45, 80, 75, 10$

$w_i = 1, 2, 3, 4, 5$

weighted mean = $\frac{\Sigma x_i w_i}{\Sigma w_i}$

= $\frac{(40\times1)+(45\times2)+(80\times3)+(75\times4)+(10\times5)}{1+2+3+4+5}$

= 48 (Ans)

Data: 600, 470, 170, 430, 300

mean: $\frac{600+470+170+430+300}{5}$

= 394

variance (σ^2) = $\frac{(600-394)^2+(470-394)^2+(170-394)^2+(430-394)^2+(300-394)^2}{5}$

= 21704

standard deviation (σ) = $\sqrt{21704}$

= 147.32 (Ans)

MID 01 Ans 2(b)

Hossain min 2021-1-60-071

sorted data:— 64, 65, 68, 69, 70, 71, (72), 72, 75, 75, 80, 81, 83, 85

25th percentile (Q1) = 25% × 14
= 3.5 th value
= 68 + (69 − 68) × 0.5
= 68.5

50th percentile (median) = 50% × 14
= 7th value
= 72

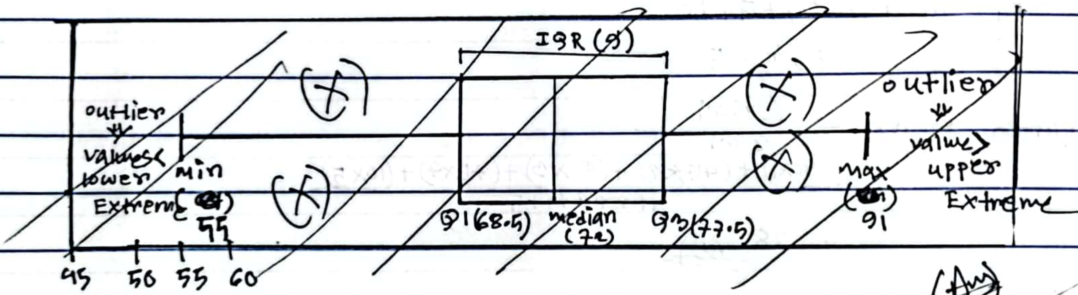
75th percentile (Q3) = 75% × 14
= 10.5 th value
= 75 + (80 − 75) × 0.5
= 77.5

IQR = (Q3 − Q1)
= (77.5 − 68.5)
= 9

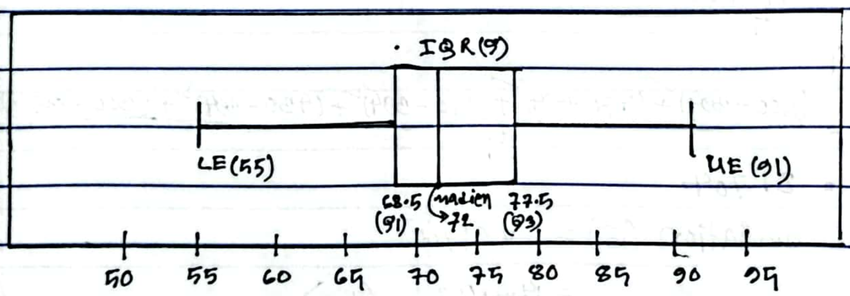
max = 85
min = 64

upper Extreme = {Q3 + (IQR × 1.5)}
= {77.5 + (9 × 1.5)}
= 91

Lower Extreme: Q1 − IQR × 1.5
= {68.5 − (9 × 1.5)}
= 55



(Ans)



we can see
no outlier

Ans

Q:

	S	NS	
M	50	70	120
F	30	50	80
T	80	120	200

Date :
Page :

Ho : There is no significant association between gender and somking

H1 : There is a significant association between gender and somking Habits

Expected value for male somoker $= \frac{80 \times 120}{200}$

$= 48$

Expected value for Female somoker $= \frac{80 \times 80}{200}$

$= 32$

Expected value for meal Non-somker $= \frac{120 \times 120}{200}$

$= 72$

Expected value for Female Non-somker $= \frac{120 \times 80}{200}$

$= 48$

Now,

$$x_c^2 = \frac{(Oi - Ei)^2}{Ei}$$

x_c^2 all values are sum $= (0.084 + 0.125 + 0.056 + 0.084)$

$= 0.349$

Degrees of freedom $= (2-1) \times (2-1)$

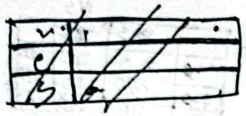
$= 1$

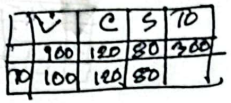
As we can see for alpha level of 0.05 and 1 degree of freedom.
The critical statistic is 3.841.
And obtain statistic is 0.349
So, critical statistic > obtain statistic
Ho do not rejected

4.30

(Ans)

$(2-1)\times(3-1)$
$2-1$





Date :
Page :

H_0: There is no significant difference in the ice-cream flavory among kids

H_1: There is a significant difference in the ice-cream flavory among kids

Now,

Expected value for vanilla flavory $= \frac{100 \times 300}{300}$

$= 100$

Expected value for chocolate flavory $= \frac{120 \times 300}{300}$

$= 120$

Expected value for strawberry flavory $\frac{80 \times 300}{300}$

$= 80$

Now,

$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$

χ^2 all values are sum $(0+0+0)$

$= 0$

Degrees of freedom $= (3-1)$

$= 2$

As, we can see that alpha level of 0.05 and 2 Degrees of freedom.

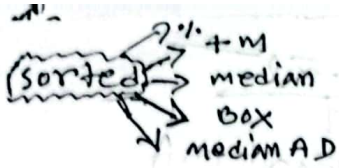
The critical statistic is 5.991

The obtain statistic is 0

So, Critical statistic > obtain statistic

H_0 do not rejected.

(Ans)



Date :
Page : 10

60, 70, 80, 75, 65, 70, 80, 70, 65, 65

Sorted order: 60, 65, 65, 65, 70, 70, 70, 75, 80, 80

- mean = 70 (Ans)
- median = $\frac{(\frac{n}{2})\text{th value} + (\frac{n}{2}+1)\text{th value}}{2}$

$= \frac{70+70}{2}$

$= \frac{140}{2} = 70$ (Ans)

- trimmed mean = $\frac{560}{10-(2\times1)}$ [For p=1]

$= 70$ (Ans)

60, 70, 80, 75, 65, 70, 70, 65, 65

- trimmed mean = $\frac{540}{10-(2\times1)}$

$= 67.5$ (Ans)

5, 6, 5, 6, 5, 8, 6, 5, 6

- mode: 5 and 6 [here, 5 and 6 both are most frequent item]

mean absolution deviation:

0, 25, 50, 75, 100 mean = $\frac{250}{5}$

$= 50$

$MAD = \frac{\sum|x-\bar{x}|}{n}$

$= \frac{|0-50|+|25-50|+|50-50|+|75-50|+|100-50|}{5}$

$= \frac{50+25+0+25+50}{5}$

$= \frac{140}{5}$

$= 30$ (Ans)

30, 40, 50, 60, 70 mean = $\frac{250}{5}$

$= 50$

$MAD = \frac{|30-50|+|40-50|+|50-50|+|60-50|+|70-50|}{5}$

$= \frac{20+10+0+10+20}{5}$

$= \frac{60}{5}$

$= 12$ (Ans)

median absolution deviation:

30, 40, 50, 60, 70, 30, 50, 40, 60

sorted order: 30, 40, 50, 60, 70

$n = 5$ odd

$\text{median} = \frac{5+1}{2}$

$= 3$ the value

$= 50$

Dataset: 50, 75, 100, 0, 25

sorted: 0, 25, 50, 75, 100

$n = \frac{5+1}{2}$ the value

$= 3$ th value

$= 50$

median $(|0-50|, |25-50|, |50-50|, |75-50|, |100-50|)$

median $(50, 25, 0, 25, 50)$

median $(0, 25, 25, 50, 50)$

median $= 25$ (Ans)

median $(|30-50|, |40-50|, |50-50|, |60-50|, |70-50|)$

median $(20, 10, 0, 10, 20)$

median $(0, 10, 10, 20, 20)$

median $= 10$ (Ans)

Date :
Page :

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

data: 600, 470, 170, 430, 300

mean: $\frac{600+470+170+430+300}{5}$

$= \frac{1970}{5}$

$= 394$

$\sigma^2 = \frac{(600-394)^2 + (470-394)^2 + (170-394)^2 + (430-394)^2 + (300-394)^2}{5}$

$= 21704$

$\sigma = \sqrt{21704}$

$= 147.32$ (Ans)

30, 40, 50, 60, 70

mean $= \frac{250}{5}$

$= 50$

$s^2 = \frac{(30-50)^2 + (40-50)^2 + (50-50)^2 + (60-50)^2 + (70-50)^2}{5-1}$

$= 250$

$s = \sqrt{250}$

$= 15.82$ (Ans)

Hossain, mim

0, 25, 50, 75, 100

mean $= 50$

$s^2 = \frac{(0-50)^2 + (25-50)^2 + (50-50)^2 + (75-50)^2 + (100-50)^2}{5-1}$

$= 1562.5$

$s = \sqrt{1562.5}$

$= 39.53$ (Ans)

sorted order:— 8, 9, 10, 10, 10, 11, 11, 11, 12, 13

mean $= \frac{105}{10}$

$= 10.5$ (Ans)

n = 10 (even no)

$= \frac{n/2 \text{ th value} + (n/2+1) \text{ th value}}{2}$

$= \frac{5\text{th value} + 6\text{th value}}{2}$

$= \frac{10+11}{2}$

$= 10.5$ (Ans)

mode = 10 and 11 [here, 10 and 11 both are most frequent item]

variance $(s^2) = \frac{(8-10.5)^2 + (9-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (11-10.5)^2 + (12-10.5)^2 + (13-10.5)^2}{10-1}$

$= \frac{18.5}{9}$

$= 2.056$ (Ans)

$s = \sqrt{2.056}$

$= 1.43$

MAD: $\frac{|8-10.5| + |9-10.5| + \cdots}{10}$

$= \frac{11}{10}$

$= 1.1$

median $(|8-10.5|, |9-10.5|, |10-10.5|, |10-10.5|, |11-10.5|, |11-10.5|, |11-10.5|, |12-10.5|, |13-10.5|)$

median $(2.5, 1.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 2.5)$

$= \frac{0.5 + 0.5}{2} = 0.5$ AM

H_0: There is no linked between gender and political party preferrance
H_1: There is linked between gender and political party preference

Expected value for male republican $= \frac{240 \times 200}{420}$

$= 114.28$

Expected value for female republican $= \frac{240 \times 220}{420}$

$= 125.72$

Expected value for male democrat $= \frac{130 \times 200}{420}$

$= 61.91$

Expected value for female democrat $= \frac{130 \times 220}{420}$

$= 68.09$

Expected value for male Independant $= \frac{50 \times 200}{420}$

$= 23.81$

Expected value for female Indepent $= \frac{50 \times 220}{420}$

$= 26.19$

Now,

$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$

χ^2 are all sum of values $= (1.79 + 1.63 + 1.05 + 0.97 + 1.61 + 1.46)$

$= 8.51$

degrees of freedom $(2-1) \times (3-1)$

$= (1 \times 2)$

$= 2$

we can see that alpha level of 0.05 and 2 degrees of freedom

the critical statistic is 5.991

the obtain statistic is 8.51

cr < ob so, Ho rejected (Ans)

Chi-Square test

Standard Error:— The SE is a single metric that sum up the variability the sampling distribution for a statistic

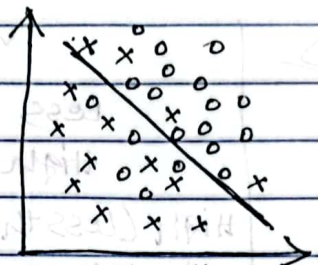
$$SE = \frac{s}{\sqrt{n}} \Rightarrow \text{(sample)}$$

Dmim STR CE 303 Final

Date :
Page :

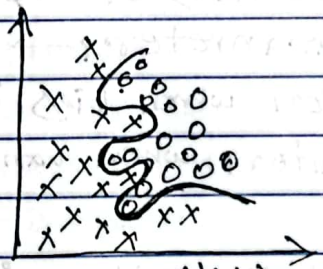
Under-fitting :- when a model has not learned learned the patterns in the training data well and is unable to generalization well on the new data, it is known as under-fitting. An under-fitting model has poor performance will result in unreliable. under-fitting occurs due to high bias and low variance. It too simple model and too much regulazation in the model.

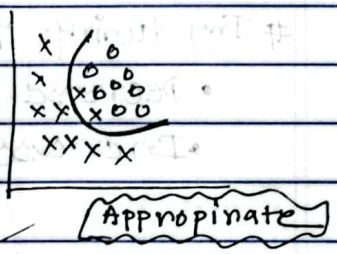
Expmoty



hossain mim
2021-1-60-071

over-fitting :- over-fitting is a phenomenon that occurs when a machine learning model is constria constraint to training set and not able to perform well on unseen data. over-fitting model is very complex and too littel regutaz regularization over-fitting model occurs due to low bias and high variance





Bias :- we can ~~can~~ define bias the error between average model prediction and ground truth.

- A model which High-Bias would not match the data set closely
- A model Low-Bias will closely match the training data set

Date :
Page :

variance refers to the changes in the model when using different portions of the traning data set. It is the variability in the model prediction - how much the machine learning function can adjust ~~develops~~ depending on the given data set

Important table

Algorith	Bias	Variance
• Linear ~~or~~ regression	High	Less
• Decision tree	Low	High
• Bagging	Low	High (Less than Decision tree)
• Random forest	Low	High (Less than Dtree & B

Get more then data :—
- when we are have high variance

Try different features :—
- Adding features helps fix High ~~Bai~~ Bias
- using smaller set of features ~~fix~~ fix high variance.

Important

Try tuning your hyperparameter :—
- Decrease regularization when bias is high
- Increase regularization when variance is high.

when a model has not learned the patterns in the training data well and is unable to generalization well on the new data, it is known as under-fitting.

when a model has not learned the patterns in the training data well and is unable to generalization well on the new data.

over-fitting is a phenomenon that occurs when a machine learning mode is ~~constraint~~ constraint to training set and not able to perform well on unseen data.

* over-fitting is a phenmenon that occurs when a ML model is constraint to training ~~data~~ set and not able to perform well on unseen data.

confusion matrix

Actual class \ Predicted class	Crocodile	Alligator
Crocodile	440 (TP)	80 (FN)
Alligator	60 (FP)	420 (TN)

- Accuracy $= \left(\frac{TP + TN}{TP + TN + FP + FN}\right)$
  $= \left(\frac{440 + 420}{440 + 420 + 80 + 60}\right)$
  $= 0.86$
  $= 86\%$

- Recall $= \left(\frac{TP}{TP + FN}\right)$
  $= \left(\frac{440}{440 + 80}\right)$
  $= 0.85$
  $= 85\%$

- Recall of crocodile (TPR) $= \frac{TP}{TP + FN}$
  $= \frac{440}{440 + 80}$
  $= 0.85$
  $= 85\%$

- Recall of Alligator (TNR) $= \frac{TN}{TN + FP}$
  $= \frac{420}{420 + 60}$
  $= 0.875$
  $= 87.5\%$

- precision for crocodile (PPV) $= \frac{TP}{TP + FP}$
  $= \frac{440}{440 + 60}$
  $= 0.88$
  $= 88\%$

- precision for Alligator (NPV) $= \frac{TN}{TN + FN}$
  $= \frac{420}{420 + 80}$
  $= \frac{420}{500}$
  $= 0.84$
  $= 84\%$

The accuracy is at 86% meaning it correctly predicts classes about 78% of the time.

This suggests that the classifier performs better in identifying crocodile instances than Alligator instances, especially in terms of precision and recall.

Ans

F1-score crocodile $= 2 \times \frac{0.85 \times 0.88}{0.85 + 0.88}$
$= 0.864$
$= 86.9\%$

F1-score Alligator $= 2 \times \frac{0.875 \times 0.84}{0.875 + 0.84}$
$= 0.857$
$= 85.7\%$

Inverse matrix:—

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 3 & 5 & -7 \end{bmatrix} \quad x = \begin{bmatrix} 8 \\ 6 \\ 14 \end{bmatrix}$$

$Ax = B$

$x = A^{-1}B$

$|A| = \begin{vmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 3 & 5 & -7 \end{vmatrix} \Rightarrow 16$

Actual \ Predicted	Crocodile	Alligator
Crocodile (P)	440 (TP)	60 (FN)
Alligator (N)	80 (FP)	420 (TN)

For corodile:—

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

$= \frac{440 + 420}{440 + 420 + 80 + 60}$

$= 0.86$

$= 86\%$ (Ans)

2) Recall $= \frac{TP}{TP + FN}$

$= \frac{440}{440 + 60}$

$= 0.88$

$= 88\%$ (Ans)

3) Precision $= \frac{TP}{TP + FP}$

$= \frac{440}{440 + 80}$

$= 0.846$

$= 84.6\%$ (Ans)

4) F1-score $= 2 \times \frac{0.846 \times 0.88}{0.846 + 0.88}$

$= 0.862$

$= 86.2\%$ (Ans)

For Alligator:—

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

$= \frac{420 + 80}{420 + 80 + 440 + 60}$

$= 0.5$

$= 50\%$ (Ans)

2) Recall $= \frac{TN}{TN + FP}$

$= \frac{420}{420 + 80}$

$= 0.84$

$= 84\%$ (Ans)

3) Precision $= \frac{TN}{TN + FN}$

$= \frac{420}{420 + 60}$

$= 0.875$

$= 87.5\%$ (Ans)

4) F1-score $= 2 \times \frac{0.875 \times 0.84}{0.875 + 0.84}$

$= 0.857$

$= 85.7\%$

This suggest that the classifier perform better in identifying "x" instances than "y" instances, especially in terms of pres and recall.

• In multiplication is associative justify :-

Date :
Page :

$$x = \begin{bmatrix} A & B \\ C & D \end{bmatrix}_{2}^{1} \qquad y = \begin{bmatrix} E & F \\ G & H \end{bmatrix}_{1}^{2} \qquad z = \begin{bmatrix} I & J \\ K & L \end{bmatrix}$$

$$\{(x \times y) z\} = \{x \times (y \times z)\}$$

Now,

$$(x \times y) = \begin{bmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{bmatrix}$$

$$\{(x \times y) z\} = \begin{bmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{bmatrix} \times \begin{bmatrix} I & J \\ K & L \end{bmatrix}$$

$$= \begin{bmatrix} AEI + BGI + AFK + BHK & AEJ + BGJ + AFL + BHL \\ CEI + DGI + CFK + DHK & CEJ + DGJ + CFL + DHL \end{bmatrix}$$

$$(y \times z) = \begin{bmatrix} EI + FK & EJ + FL \\ GI + HK & GJ + HL \end{bmatrix}$$

$$\{x \times (yz)\} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} EI + FK & EJ + FL \\ GI + HK & GJ + HL \end{bmatrix}$$

$$= \begin{bmatrix} AEI + AFK + BGI + BHK & AEJ + AFL + BGJ + BHL \\ CEI + CFK + DGI + DHK & CEJ + CFL + DGJ + DHL \end{bmatrix}$$

So, $\{x \times (yz)\} = \{(x \times y) \times z\}$ Ans.

Matrix multiplication is associative.

Final answer

Date :
Page :

In not commutative justify

$$x = \begin{bmatrix} a & b & p \\ c & d & q \end{bmatrix} \qquad y = \begin{bmatrix} e & f \\ g & h \\ i & j \end{bmatrix}$$

$(xy) \neq (yx)$

$$(xy) = \begin{bmatrix} a & b & p \\ c & d & q \end{bmatrix} \times \begin{bmatrix} e & f \\ g & h \\ i & j \end{bmatrix}$$

$$\begin{bmatrix} ae+bg+pi & af+bh+pj \\ ce+dg+qi & cf+dh+qj \end{bmatrix}$$

$$(yx) = \begin{bmatrix} e & f \\ g & h \\ i & j \end{bmatrix} \times \begin{bmatrix} a & b & p \\ c & d & q \end{bmatrix}$$

$$\begin{bmatrix} ea+fc & eb+fd & ep+eq \\ ga+hc & gb+hd & gp+hq \\ ia+jc & ib+jd & ip+jq \end{bmatrix}$$

so, $(xy) \neq yx$

we can say that the martix multipuiction is not commutative (Ans)

Final

• Addition
• Inverse
• deduction

Date :
Page :

a_{11} a_{12} a_{13}

$$A = \begin{bmatrix} 1 & 2 & -3 \\ -6 & -8 & 10 \\ -1 & 9 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 0 & -9 & 6 \\ 3 & 4 & -2 \\ 0 & 8 & -1 \end{bmatrix}$$

a_{21}

যোগ বা বিয়োগ এর জন্য order equal হতে হবে।

$$(A+B) = \begin{bmatrix} 1 & -7 & 3 \\ -3 & -4 & 8 \\ -1 & 17 & 3 \end{bmatrix} \text{ Ans} \qquad (A-B) = \begin{bmatrix} 1 & -11 & -9 \\ -9 & -12 & 12 \\ -1 & 1 & 5 \end{bmatrix} \text{ Ans}$$

$$(A+B) = \begin{bmatrix} 3 & 1 & -1 \\ 2 & 3 & 4 \\ -4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 3 & 1 \\ 1 & 4 & 2 \\ 1 & 6 & 9 \end{bmatrix} \Rightarrow \begin{bmatrix} 4 & 4 & 6 \\ 3 & 7 & 6 \\ -3 & 11 & 15 \end{bmatrix} \text{ Ans}$$

$$(A-B) = \begin{bmatrix} 2 & -2 & -2 \\ 1 & -1 & 2 \\ -5 & -1 & -3 \end{bmatrix} \text{ Ans}$$

$$6A+3B = \begin{bmatrix} 18 & 6 & -6 \\ 12 & 18 & 24 \\ -24 & 30 & 36 \end{bmatrix} + \begin{bmatrix} 3 & 9 & 3 \\ 3 & 12 & 6 \\ 3 & 18 & 27 \end{bmatrix} = \begin{bmatrix} 21 & 15 & -3 \\ 15 & 30 & 30 \\ -21 & 48 & 63 \end{bmatrix} \text{ Ans}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \begin{bmatrix} 3 & 2 \\ 4 & 4 \\ -1 & 6 \end{bmatrix}$$

1 → 2
3 → 4

* প্রথম ম্যাট্রিক্সের কলাম সংখ্যা = দ্বিতীয় ম্যাট্রিক্স সারি সংখ্যা

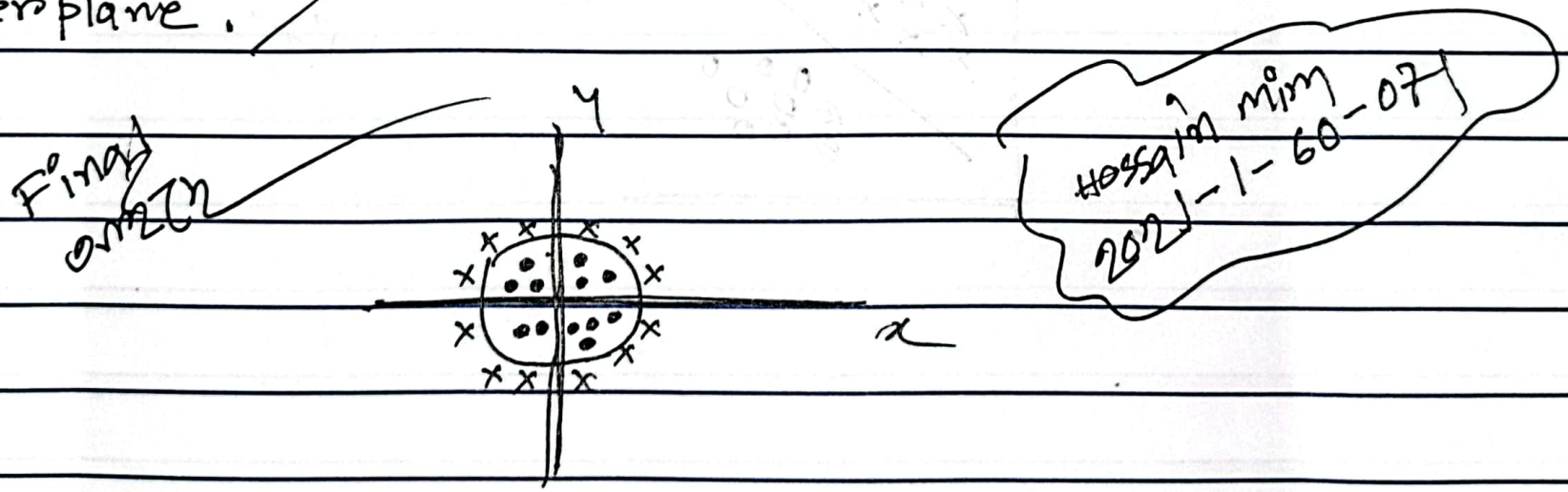
$$\begin{bmatrix} 3+8-3 & 3+8+18 \\ 12+20+(-6) & 8+20+36 \end{bmatrix} = \begin{bmatrix} 8 & 28 \\ 26 & 64 \end{bmatrix} \text{ Ans}$$

$$A = \begin{bmatrix} 1 & 3 & 8 \\ 2 & 4 & 6 \end{bmatrix}$$

$$A^T = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 8 & 6 \end{bmatrix} \text{ Ans}$$

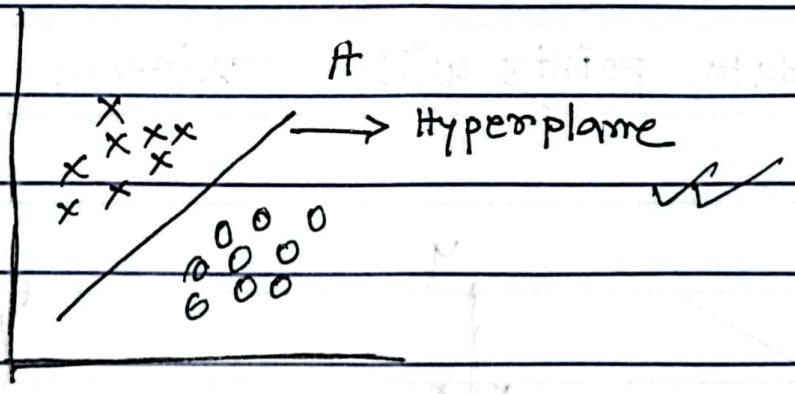
The svm kernal is a function that takes low dimensional input space and transform it into a High dimensional space that means it converts not separable problem to a separable problem. It is mostly useful in non-linear data separation problem. It simply put does some extremely complex data transmation, then it finds out the process the data based on the labels or output.

It works by finding the best possible boundaries that can separate two classes data points with maximum margin, also it known as hyperplane.



Ans

The support vector machine is defined as a machine learning algorithm that uses supervised learning model to solved complex classification, regression and outlier detection problem by performing the optimal data tranymation that determine boundaries between the data points based on the labels or output →



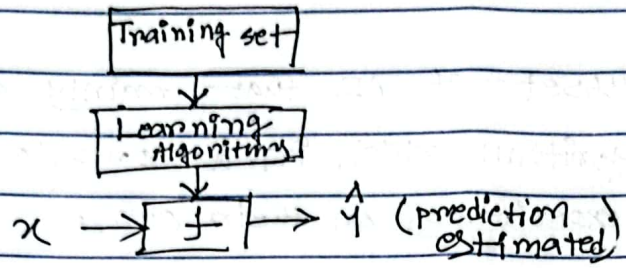
ML is said as a subset of AI that mainly content with development of algorithm which in a computer to learn form the data and past experiences on their own.

ML is said as a subset of AI that mainly contrants with development of algorithms which in a computer to learn from the data and past experiens experiences their own.

7

Date :
Page :

Traing set: Traing set is the major portion of orginal dataset which is used for ~~testing~~ training the ML model.



- model works accuracy $y = \hat{y}$
- gap between y and $\hat{y}$ is known as error

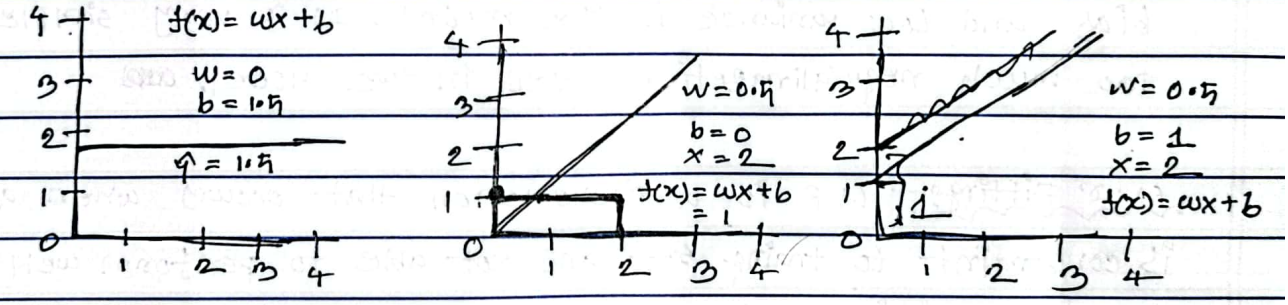
Equation = $f_{w,b}(x) = wx + b$

$= f(x) = wx + b$

$= \hat{y}$

cost function: Cost function quantifies the error between predicted and target values and present that error in the from of a single real number

Imp

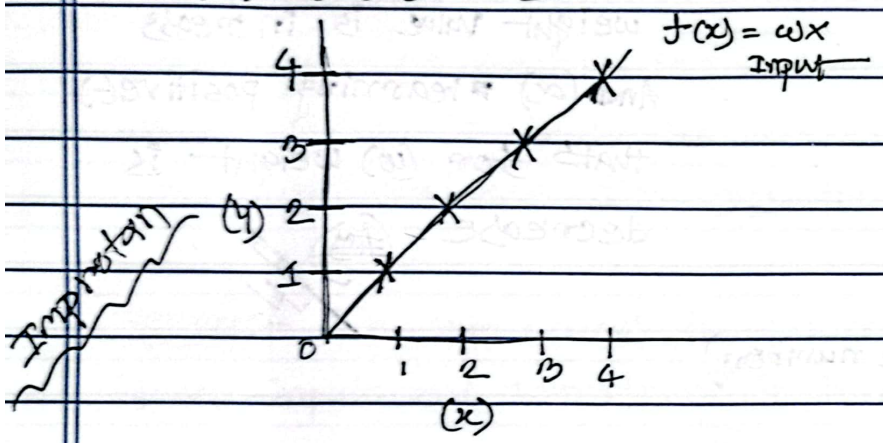


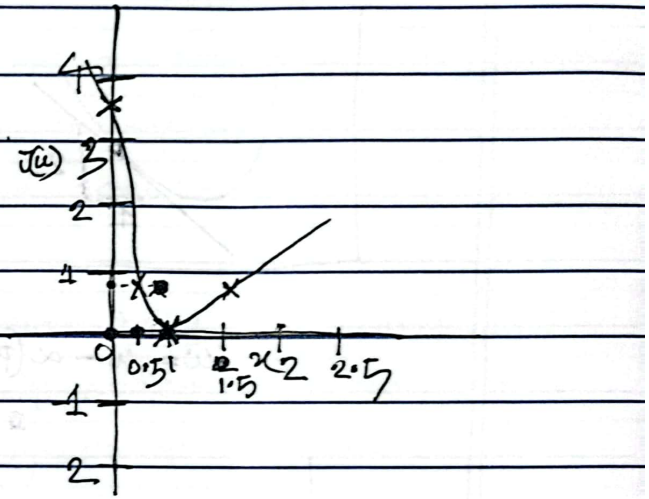
cost function:

$$J(w,b) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})$$

m = number of training example.

Regression problem:—





when, w = 0

$f(x) = wx$

$= (0 \times 1) = (0 \times 2) = (0 \times 3) = (0 \times 4)$
$\Rightarrow 0 \quad \Rightarrow 0 \quad \Rightarrow 0 \quad = 0$

Now, $J(w) = \frac{1}{2 \times 4}\{(0-1)^2 + (0-2)^2 + (0-3)^2 + (0-5)^2\}$

$= 3.75$

when $w = 0.5$

$\Rightarrow (0.5 \times 1) = (0.5 \times 2) \Rightarrow (0.5 \times 3) = (0.5 \times 4)$
$= 95 \quad \Rightarrow 1.0 \quad = 1.5 \quad = 2.0$

Now, $J(w) = \frac{1}{2 \times 4}\{(0.5-1)^2 + (1.0-2)^2 + (1.5-3)^2 + (2-5)^2\}$

$= 0.94$

when, $w = 1$

$\Rightarrow (1 \times 1) \Rightarrow (1 \times 2) \Rightarrow (1 \times 3) \Rightarrow (1 \times 4)$
$= 1 \quad \Rightarrow 2 \quad \Rightarrow 3 \quad = 4$

$\Rightarrow J(w) = \frac{1}{4 \times 2}\{(1-1)^2 + (2-2)^2 + (3-3)^2 + (4-4)^2\}$

$= 0$

when $w = 1.5$

$= 1.5 \times 1 \quad \Rightarrow 1.5 \times 2$
$= 1.5 \quad = 3$

$= 1.5 \times 3 \quad = 1.5 \times 4$
$= 4.5 \quad = 6$

$J(w) = \frac{1}{2 \times 4}\{(1.5-1)^2 + (3-2)^2 + (4.5-3)^2 + (6-5)^2\}$

$\Rightarrow 0.94$

So, we can see lowest w value is

$w = 1$

Best regression line is $w = 1$ (Ans)

GD

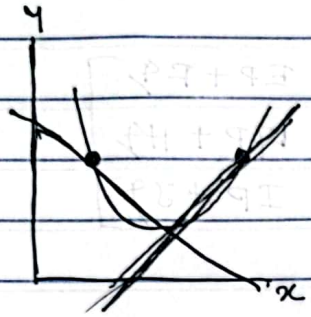
Date :
Page :

GD Algo Problem :—

we know that

$$w = w - \alpha \frac{\partial}{\partial w} J(w,b)$$

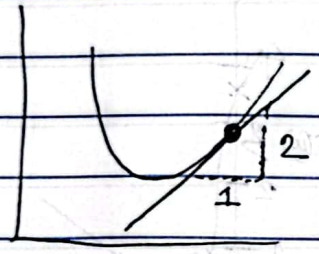
weight — learning rate



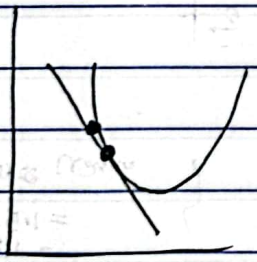
Assum,
$w = 2$
$\alpha = -1$
$w = \{2 - (-1)\}$
$= 2+1$
$= 3$

Assum,
$w = 2$
$\alpha = 1$
$w = (2-1)$
$= 1$

So, we can say that the learning rate (α) is negative thats for weight value is increase And (α) learning positive(+) thats for (w) weight is decrease (Ans)



$w = w - \alpha$ (positive number)

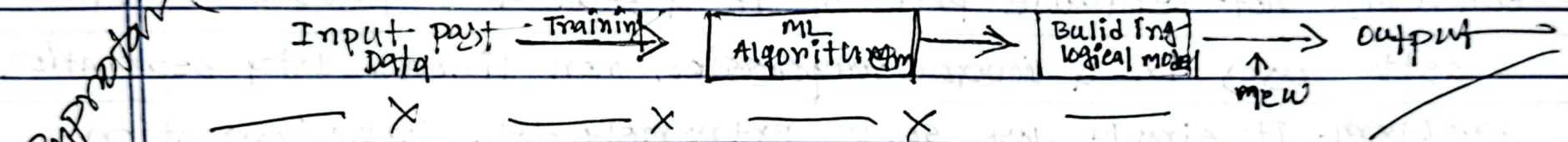


$w = w - \alpha$ (negative number)

Hossain mim
2021-1-60-071

- If (α) learning rate is too small the GD is may be slow
- If (α) learning rate is too large the GD may be overshoot, never minimum and fail to converge

machine learning is said as a subset of AI that mainly corent with development which in a comput to learn from the data and past ~~experiences~~ experiens thier own.



Supervised learning :— supervised learning is a type of machine learning method in which we provied sample labeled data to to the machin learning system in oder to traing it and on that basis, it predicts the output.

The goal of supervised learning is map input data with the out put data. The supervised learning is based on supervision.

supervised learning can be grouped futher in catagories

1) Classification
2) Regression

The example of SL is spam filtering, Image recognition system.

Date :
Page :

unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled and algorithms needs to act on that data without any supervision The goal of USL is to restructure the ip input data into new features with similar patterns. In unsupervised learning we do not have a predetermined result

I) clustering II) Associations.

For example Finding out which customers made similar product purchases,