$$w_1^* = w_1 - \alpha\left(\frac{\partial F}{\partial w_1}\right)$$

$$= 0.1 - 0.02\left(1.00563 \times 10^{-3}\right)$$

$$\boxed{w_1^* = 0.0999}$$

6/25/22

K-NN.

common thing among all classifiers is that they all operate on two-steps.

i, Inductive step: for eg., in LDA, deriving vectors to transform data,

eager learning
- → for neural network training weights (learning weights)
- → for SVM, reading data and adjusting hyperplane parameters '$\bar{w}$' and 'b'

this is called model building.

ii, Apply the model — deductive step;

lazy learning: when data is available, you don't grab it to an model, you keep the data as is until testing phase arrives, the teste samples are compared with the stored data.

we don't have a gigantic model spending too much time in learning, for all LDA, svm and Neural networks the stored data is utilized in the inductive step for learning,

→ but in K-NN, during this inductive step, you only simply store the data and keep it ready that's it

Eg:- Rote classifier

Rote learning means, simply compare test with existing and finds an exact match, (memorizes training set).

Potential drawback:- rote learner will struggle

K-NN:-

Find training examples which are relatively similar, 'nearest neighbours' can be used to determine the class label, when given a test sample its proximity to rest of data points in the training set.

↓ depends on kind of distance used.

Euclidion Vs Manhattan distance.,

Euclidion ↓ direct distance

Manhattan ↓ to move from one points to another. It must be computed from previous point.

→ Choose an odd 'K' value for a 2 class problem. for deciding majority.

→ k must not be a multiple of the number of classes.

→ drawback is complexity involved in searching the nearest neighbours,

Numerical:- A person cannot like both games.

Name	Age	Gender	Sport
Ajay	32	M	Football
Mark	40	M	neither
Sara	16	F	cricket
Zonia	34	F	cricket
Sachin	55	M	neither
Rahul	40	M	cricket
Pooja	20	F	neither
Smith	15	M	cricket
Laxmi	55	F	Football
Michael	15	M.	Football.

Predict what class of sport does a person with name 'Angelina', age of '5', Female likes?

does she belongs to football / cricket / neither

assume that $k = 3$, finding 3 neighbours.,

(i) step: convert Discrete data to numeric.,

let male = 0;

Femal = 1;

transformed data.

Name	Age	Gender	Sport
Ajay	32	0	Football
Mark	40	0	neither
Sara	16	1	Cricket
Sachin	55	0	Neither
Rahul	40	0	cricket
pooja	20	1	neither
smith	15	0	cricket
Laxmi	55	1	football
Michael	15	0	football
Zaira	34	1	Cricket
Angelina	5	1	?

to find out distance, let's use Euclidean distance

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

Calculate distance from Angelina to all other reference persons

here in eg., x = age

y = gender.,

if there are multiple features in the dataset

then euclidean distance becomes

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \ldots + (n_1 - n_2)^2}$$

(i) distance b/w Ajay & Angelina

ajay $x_2 = 32$; $y_2 = 0$

Angelina ⇒ $x_1 = 5$; $y_1 = 1$

$= \sqrt{(5-32)^2 + (1-0)^2}$

$= \sqrt{729+1}$

$= 27.02$

distance of Angelina to ajay is 27.02. Ag. ajaj.

~~ajay~~

→ w.r.t Mark $= \sqrt{(5-40)^2+(1-0)^2} = 35.01$

→ w.r.t Sora $= \sqrt{(5-16)^2 + (1-1)^2} = 11.00$

→ w.r.t sachin $= \sqrt{(5-55)^2+(1-0)^2} = 50$

→ w.r.t Rahul $= \sqrt{(5-40)^2+(1-0)^2} = 35.01$

→ w.r.t Pooja $= \sqrt{(5-20)^2 + (1-1)^2} = 15$

→ w.r.t smith $= \sqrt{(5-15)^2 + (1-0)^2} = 10$

→ w.r.t Laxmi $= \sqrt{(5-55)^2 + (1-1)^2} = 50$

→ w.r.t Michael $= \sqrt{(5-15)^2 + (1-0)^2} = 10.05$

→ w.r.t Zaira $= \sqrt{(34-5)^2 + (1-0)^2} = 9.00$

⇒ '3' closest records to angelina, {find first '3' min distance

↳ Zaira (9) → cricket ✓ } 2 votes.
↳ Smith (10) → Cricket ✓
↳ michael (10.05) → Football.

so angelina likes 'cricket'

⇒ How to compute / use Manhattan or minkowski ?

minkowski

$2 = p = 1$

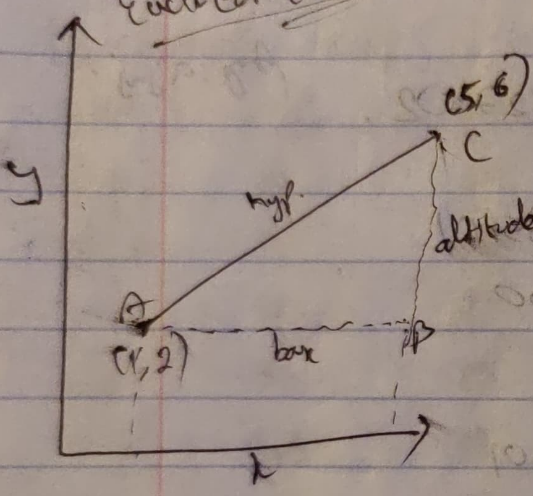
$$\left[\sum_{i=1}^{N} |x_i - y_i|^p \right]^{1/p}$$

manhattan Distance :- {mean absolute distance}

$$\sum_{i=1}^{n} |x_i - y_i|$$

$= (x_2 - x_1)$.

Euclidean distance :-



$AC^2 = \sqrt{AB^2 + BC^2}$

i.e. $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

$\sqrt{\underbrace{(x_2 - x_1)^2}_{AB} + \underbrace{(6-2)^2}_{BC}}$

$= \sqrt{4^2 + 4^2}$

$= \sqrt{32}$

↓

→ 2 or 3 dimensions only, (small, preferably < 5 dimens).

more than 10 → we cannot use Euclidean,

→ Can be used when there are no outliers.

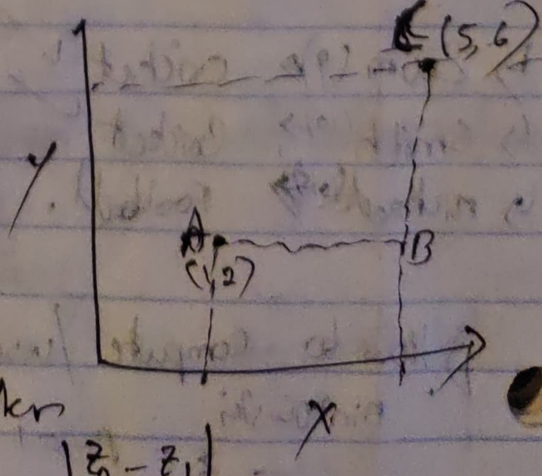
Manhattan Distance.

$$\sum_{i=1}^{n} |x_i - y_i|$$

$|x_2 - x_1| + |y_2 - y_1|$

if there are z-dimension, then

$|x_2 - x_1| + |y_2 - y_1| + \ldots |z_2 - z_1|$



Eg :- Chess movements.,

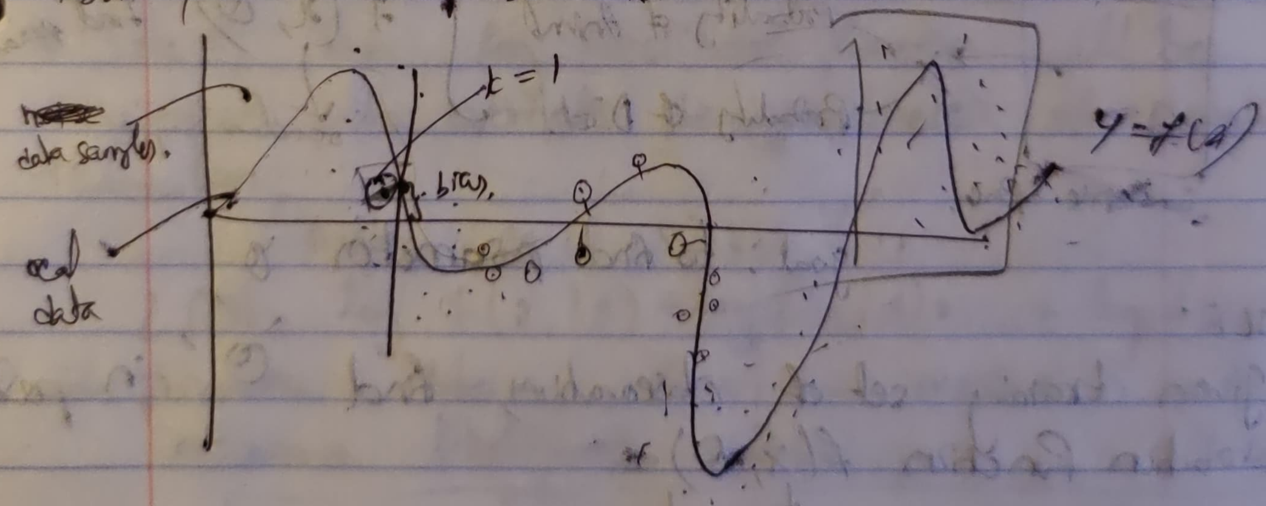
Used
→ where there is attack of outliers

it's less influenced by outliers.,

→ when data is in multidimensional (>5)

Issues with Small k : low bias → high variance (complex)
if there is an outlier, it is effecting the distance
so there is a chance of miss classification.

with large k :- high bias → low variance. (simple)
say - has less examples / samples and if k = 19,
there is a chance that all '6' could get into the
first / initial positions.



bias = distance b/w data point & curve.

↳ use cross-validation :- break data into train, validation and test subsets.. 60-20-20 split.

↳ try different values of 'k' and use the one that gives minimum error on the validation set.

Maximum Likelihood Estimation.

about estimating parameters, using data samples, probability distribution function.

3R 2W 4G.

$x \to R, W, G$

$$P(R) = 3/9 \quad \frac{\text{no. of red balls}}{\text{total balls}}$$

$$P(W) = 2/9$$

$$P(G) = 4/9$$

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

$$P(x, \theta)$$

x: random sample

θ: unknown parameter that describes this function.

In classification problem

$x \to$ Animal

$x \to$ Bird.

instance space $\to x$:

- Probability of Animal
- Probability of Bird.

probabilities, function to describes this.

$f(x, \theta)$ find parameter θ.

x: random sample

goal: to find parameter θ.

MLE:

given training set of observations, find θ in probability distribution function $f(x, \theta)$

$$P(\theta | \text{data}) = \frac{P(\text{data} | \theta) \cdot P(\theta)}{P(\text{data})}$$

get from base theorem

$P(\text{data} | \theta)$: observations, likelihood function, $l(\theta)$

$$\boxed{P(x|\theta)} = l(\theta)$$

data is random sample 'x'

where x can be defined as

$$x = \underbrace{\{x_1, x_2, \ldots, x_n\}}_{\text{data features}}$$

so $l(\theta)$ can be written as,

$l(\theta) = P(X|\theta)$

$= P((x_1, x_2 \dots x_n)|\theta)$

$l(\theta) = P(x_1|\theta) \cdot P(x_2|\theta) \dots P(x_n|\theta)$

Applying log on both sids,

$\log l(\theta) = \log [P(x_1|\theta) \cdot P(x_2|\theta) \dots P(x_n|\theta)]$

Log likelihood function

$L(\theta) = \log [P(x_1|\theta) \cdot P(x_2|\theta) \dots P(x_n|\theta)]$

we know

$\log(a \cdot b) = \log a + \log b$

$$\boxed{L(\theta) = \log P(x_1|\theta) + \log P(x_2|\theta) + \dots + \log P(x_n|\theta)} \quad (1)$$

In MLE, this log likelihood function will be maximized to estimate the parameter of probability distribution,

writing eq (1) w.r.t f,

$L(\theta) = \log f(x_1|\theta) + \log f(x_2|\theta) + \dots + \log f(x_n|\theta)$

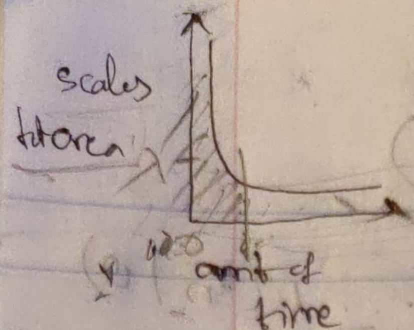
the paramtr obtained from this method is called maximum likelihood estimate,

"Exponential distribution"

statistical distribution that models the time b/w events.

→ How long will you wait before you get another text sms

→ How much time will pass before the next person uses the video?



if we are interested in the probability of an event like someone viewing the video happening with 0 to 5 sec, we solve for area under curve from $x=0$ to $x=5$ seconds

equation for exponential distribution,

$$\boxed{y = \lambda e^{-\lambda x}}$$ → plug some value for x

o/p is value for y

rate parameter which is proportional to how things quickly happen,

for the above graph, $\lambda = 1$
means that someone watching this video on an avg every for avg second,

'λ' models the happening of event

→ if $\lambda = 2$, models someone watching this video, on avg twice every second

→ $\lambda = 0.5$, and this models someone watching the video on avg, once for every 2 seconds

"The goal of maximum likelihood is given a set of measurements find the optimal value of λ

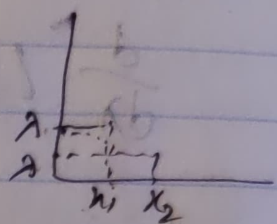
let's say collection of dataset about time passed b/w views of the video,

x_1 = the amt of time that passed b/w 1st & 2nd views
x_2 = the amt of time that passed b/w 2nd & 3rd view
x_3 = " " " " " " " 3rd & 4th view

n measurements for $x_1, x_2, \ldots x_n$.

- assume that we already have good value for λ, what is the likelihood ~~of~~ 'λ' given first measurement x_1?

$$L(\lambda \mid x_1) = \lambda e^{-\lambda x_1}$$
$$L(\lambda \mid x_2) = \lambda e^{-\lambda x_2}$$



what's likelihood of λ given both x_1, x_2

$$L(\lambda \mid x_1) = \lambda e^{-\lambda x_1} \qquad L(\lambda \mid x_2) = \lambda e^{-\lambda x_2}$$

$$L(\lambda \mid x_1 \text{ and } x_2) = L(\lambda \mid x_1) \cdot L(\lambda \mid x_2)$$
$$= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2}$$
$$= \lambda^2 \left[e^{-\lambda x_1} \cdot e^{-\lambda x_2} \right]$$
$$= \lambda^2 \left[e^{-\lambda (x_1 + x_2)} \right]$$

what is likelihood ~~best~~ of λ given all of the data $x_1, x_2, \ldots x_n$?

$$L(\lambda \mid x_1, x_2 \ldots x_n) = L(\lambda \mid x_1) \cdot L(\lambda \mid x_2) \ldots L(\lambda \mid x_n)$$
$$= \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \ldots \lambda e^{-\lambda x_n}$$
$$= \lambda^n \left[e^{-\lambda x_1} e^{-\lambda x_2} \ldots e^{-\lambda x_n} \right]$$
$$= \lambda^n \left[e^{-\lambda (x_1 + x_2 + \ldots x_n)} \right].$$

Computing MLE

- what if we don't have good val for 'λ'?

A). To find the maximum likelihood
→ take derivative of $L(\lambda \mid x_1, x_2, \ldots x_n)$

→ solve for λ when derivative set to equal to '0'

$$\frac{dL}{d\lambda} = 0$$

so, step 1,
derivative of the likelihood function,

$$\frac{d}{d\lambda} L(\lambda \mid x_1, x_2, \dots x_n) = \frac{d}{d\lambda} \lambda^n \left[e^{-\lambda(x_1 + x_2 \dots + x_n)} \right]$$

any time if there is function having 'e' in it, its easier to take derivative of the log of that function.

so apply log on that function,

$$= \frac{d}{d\lambda} \log\left(\lambda^n \left[e^{-\lambda \cdot (x_1 + x_2 + \dots x_n)} \right] \right)$$

"why we have to do this?"
→ derivative of a function and the derivative of the log of a function equal's or becomes zero at the same place,
so, for the purpose of finding where the derivative equals '0', the original function and its log are interchangable,

$$L(\lambda) = \frac{d}{d\lambda} \log(\lambda^n) + \log e^{-\lambda(x_1 + x_2 + x_3 + \dots + x_n)}$$

$$= \frac{d}{d\lambda} n.\log\lambda + \left(-\lambda(x_1 + x_2 + x_3 + \dots + x_n)\right)$$

$$\left\{ \because \begin{array}{l} \log a^b = b \log a, \\ \log e^x = x \end{array} \right\}$$

$$= \frac{d}{d\lambda}(n \log \lambda) - \frac{d}{d\lambda}\left[\lambda(x_1 + x_2 + \dots + x_n)\right]$$

$$= n \cdot \frac{1}{\lambda} - (x_1 + x_2 + \dots x_n)$$

$$\frac{d}{d\lambda} L(\lambda | x_1, x_2 \dots x_n) = 0$$

$$0 = n \cdot \frac{1}{\lambda} - (x_1 + x_2 + \dots x_n)$$

$$(x_1 + x_2 + \dots + x_n) = n \cdot \frac{1}{\lambda}$$

$$\lambda(x_1 + x_2 + \dots x_n) = n$$

$$\boxed{\lambda = \frac{n}{x_1 + x_2 + \dots x_n}}$$

is maximum likelihood estimate for λ.

example:-

	Data point	value	
$n = 3$ here (total observation)	x_1	2	diff sample events during video watch
	x_2	2.5	
	x_3	1.5	

$$\lambda = \frac{3}{2 + 2.5 + 1.5}$$

$\boxed{\lambda = 0.5}$ is maximum likelihood estimate

MLE for Gaussian distribution, {Normal / univariant}

we have data D & model θ,

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{\underbrace{P(D|\theta)}_{\text{likelihood}} \cdot \underbrace{P(\theta)}_{\text{Prior}}}{P(D)}$$

argmax $p(\theta/D)$ is hard for computation as we don't know 'θ'

rather $p(D/\theta)$ is more easier finding 'θ' by given data D.

The gaussian function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

density for univarient gaussian is $p(x|\mu,\sigma) = f(x)$

so, for $p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

for $x = x_1, x_2, x_3, \ldots x_n$

$\rightarrow \ell(\theta) \Rightarrow \ell(\mu,\sigma)$ here in this case

$$\ell(\mu,\sigma) = p(x|\mu,\sigma)$$
$$= P((x_1, x_2, x_3, \ldots x_n)|\mu,\sigma)$$
$$= p(x_1|\mu,\sigma) \cdot p(x_2|\mu,\sigma) \ldots p(x_n|\mu,\sigma)$$

$$\therefore l(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}}$$

$$\cdots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Applying log, on both sides,

$$\log(l(\mu,\sigma)) = \log\left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$L(\mu,\sigma) = \log\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right)$$

$$L(\mu,\sigma) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$\frac{\partial}{\partial\mu}$ diff w.r.t μ

$$\frac{\partial L}{\partial\mu} = 0 - \frac{1}{2\sigma^2}\sum 2(x_i-\mu)\cdot(-1)$$

$$= \frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{n}\cancel{2}(x_i-\mu)$$

$$= \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n} x_i - \mu$$

$$\frac{\partial L}{\partial\mu} = \frac{1}{\sigma^2}(x_i - n\mu)$$

$$\frac{\partial L}{\partial\mu} = 0$$

$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

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

$$\boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i} \text{ — MLE}$$

second derivative,

$$\frac{\partial^2 l}{\partial \mu^2} = \frac{\partial}{\partial \mu}\left(-\frac{n\mu}{\sigma^2} + \frac{\sum x_i}{\sigma^2}\right)$$

$$= -\frac{\partial}{\partial \mu}\frac{n\mu}{\sigma^2} + 0$$

$$= \frac{-n(1)}{\sigma^2} < 0$$

which is minimum.

we write $\mu = \bar{x}$

so, $$\boxed{\bar{x} = \frac{1}{n}\sum_{i} x_i}$$

Bias:-

$$\underbrace{bias(\theta) = E(\theta)}_{\text{expected val of estimator}} - \theta \quad \text{true value of estimator}$$

Ideally $bias(\theta) = 0$, then estimator is unbiased.

If $bias(\theta) \neq 0$ it's biased,

$\lim_{m \to \infty} bias(\theta) = 0 \rightarrow$ asymptotically unbiased,

Estimator for Gaussian Distribution.

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

estimator for mean (μ),

we already know $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i$

To prove that estimator is unbiased,

as we have,, $bias(\theta) = E(\theta) - \theta$

$$\Rightarrow bias(\hat{\mu}) = E\left(\frac{1}{m}\sum_{i=1}^{m} x_i\right) - \mu$$

$$= \frac{1}{m}\sum_{i=1}^{m} E[x_i] - \mu$$

$$= \frac{1}{m}\sum_{i=1}^{m} \underbrace{E[x_i]}_{\text{expected val of random variable}} - \mu$$

by definition $E[x_i]$ is μ,

$$\therefore bias(\hat{\mu}) = \frac{1}{m}\left[\sum_{i=1}^{m} \mu - \mu\right]$$

$$= \frac{1}{m}\left[m\mu - \mu\right]$$

$= \frac{1}{n} n\left[\mu - \frac{\mu}{}\right]$ ~~=~~

$\therefore$ bias $(\hat{\mu}) = 0$

hence prove the estimator is unbiased.

Additional estimators:-

(i) sample mean $\Rightarrow \hat{\mu}$

(ii) sample variance $\Rightarrow \hat{\sigma}$

i.e, different $L(\mu, \sigma)$ w.r.t σ,

we get $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

bias $(\theta) = E(\theta) - \theta$

bias $(\sigma^2) \Rightarrow E(\sigma^2) = E\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2\right)$

$= \frac{1}{n} E\left[\sum_{i=1}^{n} x_i^2 - 2 \cdot x_i \cdot \mu + \mu^2\right]$

$= \frac{1}{n} E\left[\sum x_i^2 - 2 \cdot \sum x_i \cdot \mu + \sum \mu^2\right]$

here $\mu = \frac{1}{n} \sum x_i$ as we already know

$n\mu = \sum x_i$

$= \frac{1}{n} E\left(\sum x_i^2 - 2 \cdot n \cdot \mu \mu + n\mu^2\right]$

$= \frac{1}{n} E\left(\sum x_i^2 - 2n\mu^2 + n\mu^2\right]$.

$= \frac{1}{n} E\left(\sum x_i - n\mu^2 \right)$

$= \frac{1}{n} \left(\sum_{i=1}^{n} E(x^2) - \cancel{n} E(\mu^2) \right)$

$= \frac{1}{n} \sum_{i=1}^{n} E(x^2) - E(\mu^2)$

$= \frac{1}{\cancel{n}} \cdot \cancel{n} \cdot E(x^2) - E(\mu^2)$

$$\boxed{E(\sigma^2) = E(x^2) - E(\mu^2)} \quad \text{—} \ (1).$$

we know $E(\mu^2) = \mu^2$

$E(\sigma^2) = \sigma^2$

from (1).,

$\sigma^2 = E(x^2) - \mu^2$

i.e., $E(x^2) = \sigma^2 + \mu^2$ — (2)

μ is a random variable,

$\sigma_\mu^2 = E(\mu^2) - [E(\mu)]^2$ — (3).

we also know that

if we have 'n' random variable $x_1, x_2, \ldots x_n$

their variance

$var(x_1 + x_2 + \ldots + x_n) = n \cdot \sigma^2$

$\sigma_\mu^2 \approx var\left(\frac{x_1 + x_2 + \ldots + x_n}{n} \right) = \frac{n\sigma^2}{n^2}$

dividing n on val is equvalent to diving n^2 on variance side

$\Rightarrow \quad \sigma_\mu^2 = \frac{n \cdot \sigma^2}{n^2}.$

Substituting in (3),

$$\sigma_{\mu}^2 = E(\mu^2) - [E(\mu)]^2$$

$$\frac{n\sigma^2}{n^2} = E(\mu^2) - [E(\mu)]^2$$

here $E(\mu) = \mu$ only

$$\Rightarrow \frac{\sigma^2}{n} + [E(\mu)]^2 = E(\mu^2)$$

$$\boxed{\frac{\sigma^2}{n} + \mu^2 = E(\mu^2)} \quad \text{—— (4)}$$

Substitute $E(\mu^2)$ from (4) & $E(x^2)$ from (2) in eq (1).

$$E(\sigma^2) = E(x^2) - E(\mu^2)$$

$$= \sigma^2 + \mu^2 - \left[\frac{\sigma^2}{n} + \mu^2\right]$$

$$= \sigma^2 \cancel{+\mu^2} - \frac{\sigma^2}{n} \cancel{-\mu^2}$$

$$= \sigma^2\left(1 - \frac{1}{n}\right)$$

$$= \sigma^2\left(\frac{n-1}{n}\right)$$

hence, $E(\sigma^2) \neq \sigma^2$

it's biased,

will become σ^2 if $\frac{n}{(n-1)}$ is multiplied on B.S.

$$\frac{n}{n-1} \cdot (\sigma^2) = \sigma^2$$

redefining $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

$$\frac{n}{n-1} \sigma^2 = \frac{\cancel{n}}{n-1} \cdot \frac{1}{\cancel{n}} \sum_{i=1}^{n} (x_i - \mu)^2$$

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