Regularization:

$y = 15 + 1.2x_1 + 20x_2 + 39x_3$

x_3 is vimp to determine y / output variable / dependent variable bcz x_3 has highest coefficient.
or x_3 has highest Regression Coefficient.

& x_1 is least imp.

∴ Value of x_3 is too high, working of datasets becomes very much computationaly expensive.
∴ Model becomes V. Complex.
∴ To reduce the complexity of the Model, we use Regularization.

$$\boxed{\therefore y = 0.9 + 0.7x_1 + 2x_2 + 5x_3}$$

∴ this becomes a computationally easy task.

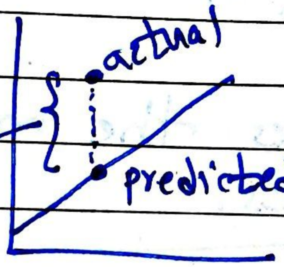
Regularization Types:

1. Ridge
2. Lasso

① Ridge : (It uses L2 Regularization Technique).

Ridge regression $= \text{Loss} + \underbrace{\alpha \|w\|^2}_{\text{penalty}}$

Loss :



Loss/ Residual error
(actual − predicted)

∴ Greater the diff, greater the Loss.

Penalty : It is imposed to reduce the loss. i.e. we try to compensate here for the loss i.e. reduce the loss. Also, it helps to scale down the magnitude of the coefficient.

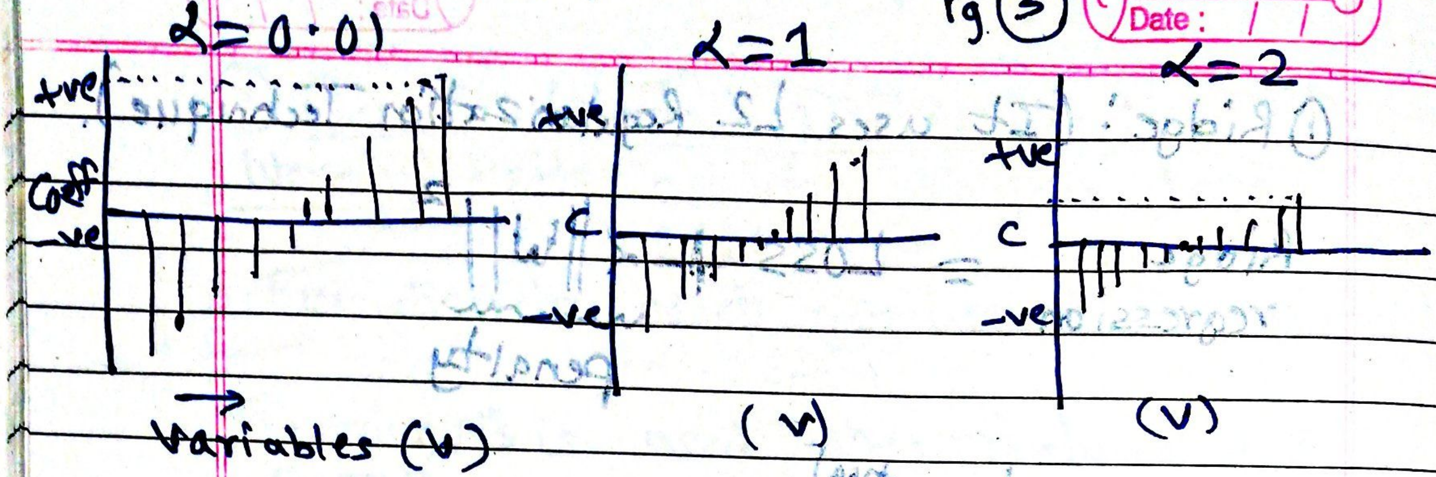
α : It is any constant value.

$\|w\|^2 = w_1^2 + w_2^2 + w_3^2 + \cdots + w_n^2$

w − It is a vector of coefficients

Now, let's understand, how does adding penalty reduces the loss.

Pg 3



α value modulates the magnitude of the coefficients.

$$\boxed{\text{Greater } \alpha \rightarrow \text{Lesser value of coefficient}}$$

Note: ① More computationally efficient ② Penalizes large nos. more ③ More popular ④ Good when encountering colinearity.

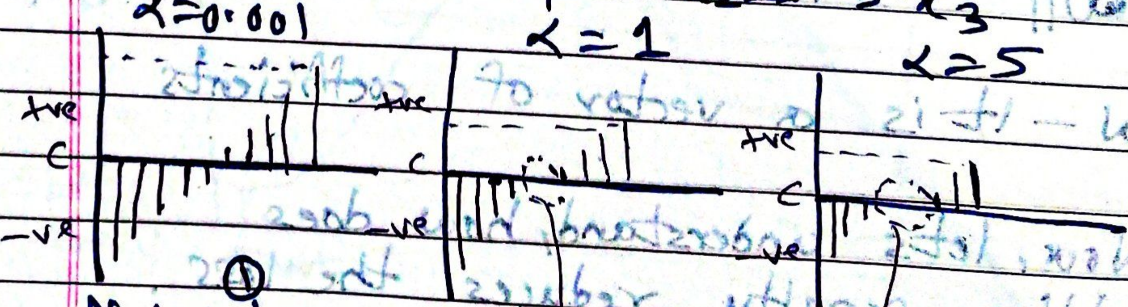
② **Lasso**: (It uses L1 Reg. Technique)

$$\text{Lasso R.} = \text{Loss} + \underbrace{\alpha||W||}_{\text{Penalty}}$$

$$||W|| = |w_1| + |w_2| + |w_3| + \dots + |w_n|)$$

$$y = 15 + 1.2x_1 + 20x_2 + 39x_3$$

$$\Rightarrow y = 0.9 + 0x_1 + 0x_2 + 5x_3$$



① Note: Lasso R. also acts as Feature Selection. Thats hw Lasso is diff. from Ridge R.

② Also, it is more robust to outliers.

③ Elastic Net Regression :

It is taking advantages of both Ridge & Lasso Regression.

Ridge R. = $Loss + \alpha ||W||^2$

Lasso R. = $Loss + \alpha ||W||$

Elastic Net R. = $Loss + \alpha_1 ||W||^2 + \alpha_2 ||W||$

* Ridge R. is used in datasets which have Multicolinearity Problem i.e. when the independent variables are highly corelated with each other.

* Lasso R. is used for Feature Selection.

* Elastic Net. R. inherits advantages of both.
  ∴ It is elastic i.e. flexible.
  ∴ E.N. Reg is a hybrid of RR & LR.

Page :
Date : / /

Naive Bayes Variants

① Bernoulli Distribution:
- Used when feature values is of Binary nature

$P(\text{Success}) = p$
$P(\text{Failure}) = q = (1-p)$

Let X be a random variable.
∴ Random Var.
$X = 1 \rightarrow$ It Signifies Success
$X = 0 \rightarrow$ It Signifies Failure

∴ Random Var. X has Bernoulli Distribution.

$$P(X=x) = p^{x} \cdot (1-p)^{1-x}$$

where $x = 0$ or 1

OR

$$P(X) = \begin{cases} p & \text{if } x=1 \\ q & \text{if } x=0 \end{cases}$$

2) Multinomial NB:

It is usually used in Natural Lang. Processing.

∴ MNB is basically used when we are interested to find the occurence times/frequency of a particular word in a Text Document.

Multinomial Distribution:

MNB is used when we are interested in finding combined probability.

Blood Group	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

So, now we hv a big population among which we are randomly picking "6 Indians" i.e. random sampling & we hv to find the following:

What is the probability that in the given random sample

1 Person is with BloodGroup O

2 Ppl are —"— A

2 —"— B

1 Person is —"— AB

Formula:

$$P(X=x_1, \ldots X_k=x_k) = \frac{n!}{x_1! \ldots x_k!} \cdot p_1^{x_1} \ldots p_k^{x_k}$$

$$\therefore P(X_1=1, X_2=2, X_3=2, X_4=1) = \frac{6!}{1!\,2!\,2!\,1!} \cdot \left(0.44^1 \times 0.42^2 \times 0.10^2 \times 0.04^1\right)$$

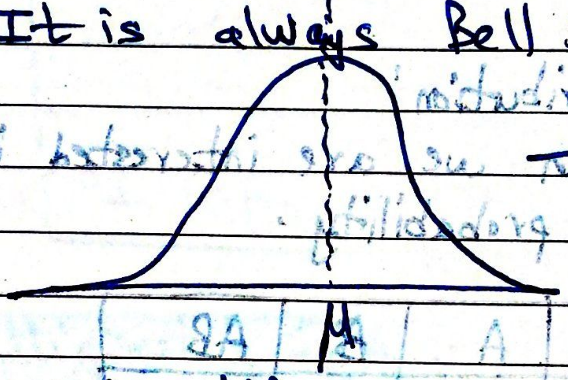
(3) Gaussian/Normal Distribution :

Warning: Don't use it if your data is DISCRETE.

For discrete count, you use Multinomial.

Gaussian NB is used when features are CONTINUOUS.

It is always Bell Shaped.



Gaussian NB can be used on IRIS dataset.

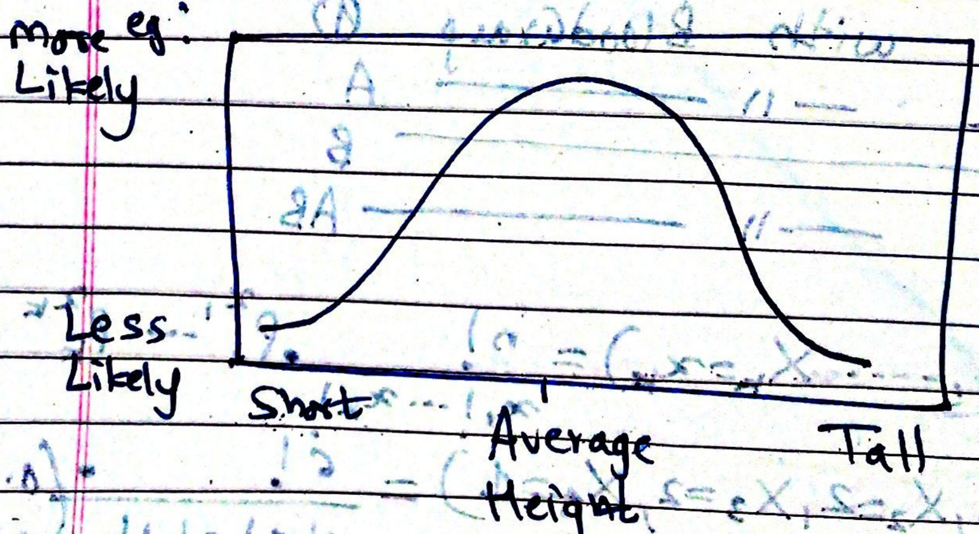
PDF:

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

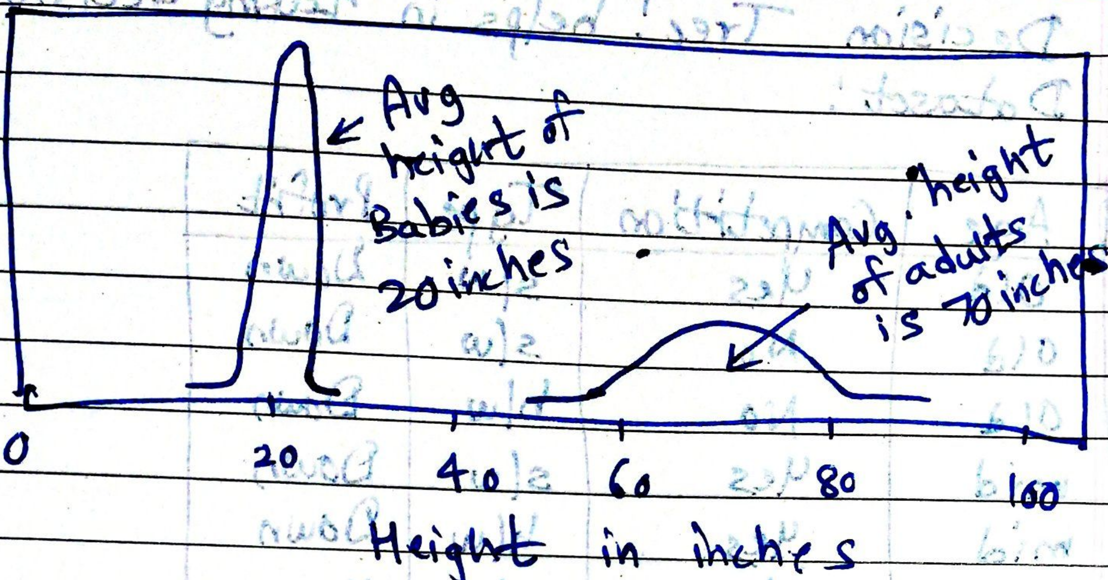
Above formula has 3 imp. Parameters mean, SD, & variance.

What is NORMAL DISTRIBUTION?

eg:



Page:
Date: / /



Normal Distributions are always centered on the average value.
Just by looking at the graph, we can tell that there is a high probability that a new born baby will be between 19 & 21 inches tall.
In contrast, adults are between 60 - 80 inches tall.
Also, we can notice that the curve for babies is way tall compared to curve for adults. This is because there are many more possibilities for adult height than for babies.
∴ Babies hv relatively small SD compared to adults.

$\sigma_{babies} = 0.6$ $\sigma_{adults} = 4$

Normal Curves are always ~~such that 95% of the measurements~~ around +/- 2 around the Mean i.e. 95% of babies fall b/w 20 +/- 1.2 inches & 95% of adults fall b/w 70 +/- 8 inches.