7/2

UNIT- II

COMPARISION BETWEEN LINEAR REGRESSION AND MACHINE LEARNING

LINEAR REGRESSION	MACHINE LEARNING
* It is try to fit the Model with a best possible hyperplane by Minimizing the errors between the hyperplane & Actual Observation	* The best Model is the problem has been Converted into an optimization problem in which errors are Modeled in squared form to Minimize errors by altering the wages
* the samples are drawn from the population & the Model will be fitted on the sample Data	* even small no. such as 30 Observations would be good enough to update the weights at the end of each iteration, it can be best fit even with 1 Observation.
* Linear Regressions performed to check the validity of the Model with parameters.	* It is non-parameteric in nature. Do not have any parameter, these models learn by themselves based on provided Data.

* Overall Model accuracy on training Data, Check using adjusted R^2.	* here the Accuracy will be Mean square ML.

COMPENSATING FACTORS IN ML MODELS:

13/2

Assumptions of Linear Regression.

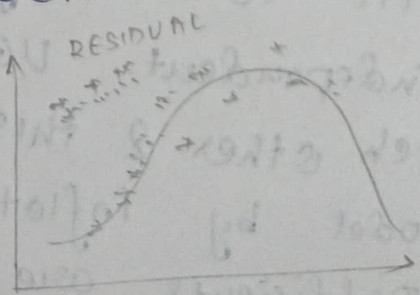
* the Dependent variable should be a linear Combination of Independent variables.
* No auto-correlation in error terms.
* errors should have zero mean and be normally distributed.
* No or little Multi-colinearity
* error terms should be homoscedistic

* Dependent variable (Y) should be a linear Combination of Independent (x) variables.
  - → Y should be the linear combination of x variables.
  - → even though x^2, Y should be Dependent variable.

$$y = a_0 + B_1(x) + B_2(x_2)^2$$

RESIDUAL:

A Measure of how far away a point is vertically from the vertical line.



→ When residual are less, when the patterns are More → non-linearity.

PATTERNS:

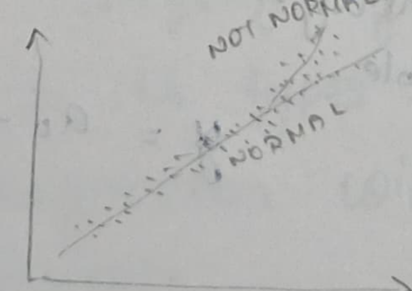
A linear Model provides the fit to the Data.

→ When residual is More: No pattern formed

* Presence of Correlation in error terms penalised Model accuracy.

* error should have Zero Mean for the Model to create an unbiased estimate.

→ normal distribution → too wide or too narrow should not divert anywhere.



* Multi-colinearity is case in which Independent variables are coordinated with each other & this situation creates unstable Model by inflathi, the Magnitude of coefficients are estimates.

→ It is Determined by variance inflation factor (VIF)

$$VIF = \frac{1}{1 - R^2}$$ R → independent variables correlation.

VIF ≤ 4, then it is no Multi-colinearity.

* error should be homoscedistic

→ errors should have a const. variance with respect to the independent variable which leads to impractically wide or narrow confidence intervals for estimates, which the grades the Models performance.

→ the error should be always homoscedistic.

STEPS APPLIED IN LINEAR REGRESSION MODELLING.

* Missing values and outlayer treatment
* Correlation check of Independent variables
* train & test random classification
* fit the Model on train Data.
* evaluate Model on test Data.

24/2

Example of Simple Linear regression from first Principle.

SLR is an approach for predicting the Dependent ~~or~~ response variable Y, given the independent or predicted variable X. it assumes there is a linear relationship between X & Y.

Residual Sum of Squares (RSS)

RESIDUAL:

Differences bet. the ~~ith~~ ith absorbed response value & predicted value from the Model.

RSS:

it is a Measures the level of variance of the error term or residuals.

Residual → error value.

if it is Smaller RSS, → Better/Best fit Model.

RSS higher Value → worst fit model.

Two hypothesis testing:

h_0 → null hypothesis (no relation bet X & Y)

h_1 → there is a relation bet X & Y

$\boxed{Y = \beta_0 + \beta_1 x} \rightarrow$ LR EQUATION

$\boxed{H_0 : \beta_1 = 0}$; $\boxed{H_0 : \beta_1 \neq 0}$

if $\beta_1 = 0$, then the Model Shows no association between both variables. $Y = \beta_0 + \varepsilon$ (null hypothesis assumption)

27/2 Machine Learning Models Lasso & Ridge Regressions.

here, a penatly is applied called as shrinkage penalty on coefficient values to Regularize the coefficients with the tuning parameter Lambda: λ

LASSO REGRESSION:

It is a regression Analysis method that performs on both variable Selection & Regularization in order to enhance the prediction accuracy & inter of the resulting Statistical Model.

Laso tends to make coefficients to absolute '0' when compare to Ridge regression

Laso will set to absolute zero, where as Ridge will not set of absolute zero.

RIDGE:

Model of tuning Method, used to Analyse any Data that suffers from Multi-collinearity

PENALTY:

Cost on the optimization func. to make the optimal solution unique.

When $\lambda = 0$ (tuning Parameter) the penalty has no impact, it will produce the same result of linear regression.

if $\lambda = \infty$, it brings the coefficients to '0'

The objective func. which is the one will minimize the Lasso & Ridge.

OBJECTIVE FUNCTION:

It is specific to the prob. domain.

it may involve pluging the candidate sol. into a model & evaluating it against the position of the training Dataset leads to error score on cost.

To set as Absolute zero:

to eliminate the unnecessary predictors from the model.

LOGISTIC REGRESSION VS RANDOM FOREST

1/3

MAXIMUM LIKELIHOOD ESTIMATION:

it is a method of estimating the parameter of models given observation, by finding the parameter values that maximize the likelihood of making the observation. This means finding parameters that maximize the probability P of event 1 & 1-P of non-even zero.

P(event + non-event) = 1

Eg: sample of (0, 1, 0, 0, 1, 0) is drawn from Binomial distribution.

What is the Max. likelihood estimate of mean.

Sol: for Binomial distribution $P(x=1) = \mu$.

$$P(x) = 0 \Rightarrow P(x=0) = 1-\mu$$

where $\mu \rightarrow$ parameter:

likelihood of $\mu \Rightarrow L(\mu)$.

$$L(\mu) = P(x=0) * P(x=1) * P(x=0) * P(x$$
$$* P(x=0) * P(x=1) * P(x=0)$$
$$= (1-\mu) * \mu * (1-\mu) * (1-\mu) * \mu * (1-\mu)$$
$$= (1-\mu)^4 * \mu^2$$

Apply log:

$$\log(L(\mu)) = \log((1-\mu)^4 \mu^2)$$
$$= 4 * \log(1-\mu) + 2 * \log(\mu)$$

$$\boxed{\frac{\partial}{\partial\mu} \log(L(\mu)) = 0} \leftarrow$$ Determine the Max value of μ, to Derivative ze

$$4 * \frac{1}{1-\mu} * (-1) + 2 * \frac{1}{\mu} = 0 \leftarrow$$ Partially diff

$$-4 * \mu + 2 * (1-\mu) = 0$$
$$\mu = 1/3$$

$$\frac{\partial^2}{\partial\mu^2} \log(L(\mu)) = -4 * \frac{1}{(1-\mu)^2} - \frac{2}{\mu^2}$$

$$\frac{\partial^2}{\partial\mu^2} \log(L(\mu)) = -4 * \frac{1}{(1-\cancel{\mu}\frac{1}{3})^2} - \frac{2}{(\frac{1}{3})^2}$$

$$= -9 - 18 = -27$$

$$L(\mu) = (1-\tfrac{1}{3})^4 \tfrac{1}{3}^2 = 0.021948$$
$$\ln(L(\mu)) = \ln(0.021948)$$

AKAIKE INFORMATION CRITERION

$$= -3.819 \Rightarrow -2\ln(L(\mu)) = -2 * -3.8$$
$$= 7.63$$

$$\boxed{AIC = -2 * \ln(L) + 2 * K}$$

AIC: [Mathematical Model]
for evaluating how well a model fits the Data it was generated from.

6/3 TERMINOLOGIES INVOLVED IN LOGISTIC REGRESSION:

INFORMATION VALUE (IV)

It is useful in the priliminary filtering of variables prior to encoding them in the model it is mainly used by industry for eliminating major variables in the first step prior to fitting the model.

$$IV = \ln\left(\frac{\%\ good}{\%\ bad}\right) * (\%\ good - \%\ bad)$$

$$\text{Weight of Evidence (WOE)} = \ln\left(\frac{\%\ good}{\%\ bad}\right)$$

Informative value	Predictive value.
less than 0.02	useless for prediction
0.02 - 0.1	it is a weak predictor
0.1 - 0.3	medium predictor
0.3 - 0.5	Strong predictor
greater than 0.5	it is suspicious or to good predictor.

it will give the ranking Based on values.

WEIGHT OF EVIDENCE (WOE):

It transform a Continuous independent variable into a set of grp based.

AIC: [AKAIKE INFORMATION CRITERION]

This Measures the relative quality of the Statistical Model for a given set of Data ~~&~~ Variance & it is a trade off between Bias.

$$AIC = -2 * \ln(L) + 2 * k$$

$L \rightarrow$ Max. value of livelihood.

$k \rightarrow$ no. of variables in the Model.

ROC: Receiver operator characteristics

This is a graphical plot that illustrates the performance of the Binary classifier as its Discriminant threshold is varied.

it purily Depends on TPR & FPR.

RANK ORDERING:

After Sorting the observations in desc order by predicted probabilities, Deciles are created.

This is nothing but given a ranking.

CONCORDANCE OR C-STATISTICS:

this is the measure of quality of fit for a Binary outcome in a logistic regression model.

ACTUAL	PREDICTED
0	0.34
0	0.12
0	0.82

Actual	Predicted
1	0.92
0	0.34
0	0.12
1	0.4
1	0.64
0	0.82
1	0.84

ACTUAL	PREDICTED
1	0.92
1	0.4
1	0.64
1	0.84

ACTUAL	PREDICTED	ACTUAL	PREDICTED	CONCORDANT	DISCORDANT
1	0.92	0 c	0.34		
c		0 c	0.12		
		0 D	0.82		
1	0.4	0 c	0.34		
D		0 c	0.12		
		0 D	0.82		
1	0.64	0 c	0.34		
		0 c	0.12		
		0 D	0.82		
1	0.84	0 c	0.34		
c		0 c	0.12		
		0 D	0.82		

Percentage of Concordant pairs

$$\frac{\text{no. of Concordant pairs}}{\text{total pairs.}}$$

% of Discordant pairs

$$\frac{\text{no. of Discordant pairs}}{\text{Total pairs.}}$$

% of tile pairs: no. of tile pairs / total pairs.

C-index / C-statistics:

$$0.5 + \text{\% of conc. pairs} - \frac{\text{\% of discor}}{2}.$$

DIVERGENCE:

the Dist. bet the Avg score of the Default accounts & the Avg score of non-Default accounts

POPULATION STABILITY INDEX: (PSI)

This is a Matric used to check the Credit score of the Model

PSI <= 0.1 ⇒ no change in the current Population w.r.to the Developement Population

$0.1 < PSI <= 0.25 \Rightarrow$ Some changes has to be taken place on the current population & warns for the Attention of the Data set.

$PSI > 0.25 \Rightarrow$ there is large changes on the current population compared with the Development time.

4/5 Applying Steps in Logistic Regression Model:

→ exclusion criteria & good bad Definition finalization

→ initial Data preparation & Analysis of Data

→ Derive or Dummy variable creation.

→ find classing & coarce classing

→ fitting the Logistic Model on the training Data

→ evaluate the Model on test Data.

→ find classing:

It grps the values into no. of bins

→ coarce classing:

After these bins created, it is used to measure the variables predictive power based on IV

K- Nearest Neighbours:

→ KNN is a non-parametric M.L Model in which the Model memorises the training observation for classifying the unseen test Data.

→ It is AKA instance based learning or lazy learning find the value k that must be a +ve integer. for the given Datapoint it needs to classify using the dist. bet. the Datapoints & all other Data of the given Dataset.

→ find the k-nearest neighbour based on Distance.

→ the given Data pt. placed in the majority of the classes respect to neighbours.

PROBLEM:

find the Person x belongs to which class based on the given values weight = 57kg, height = 170cm

x_2 , y_2

WEIGHT	HEIGHT	class	Euclidean Dist.
51 $\sqrt{(6)^2+(3)^2}=\sqrt{45}$	167	underweight	6.7
62	182	Normal	1.3
69	176	Normal	13.41
64	173	Normal	7.615
65	172	Normal	8.24
56	174	underweight	4.12
58	169	Normal	1.41 N1
57	173	Normal	3 N3
55	170	Normal	2 N2

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

$x_1 = 57$, $x_2 = 170$.

the person who's weight is falling at normal.

find the person x fans falling with the given values

age = 20, gender = male. [M = 1, F = 0]

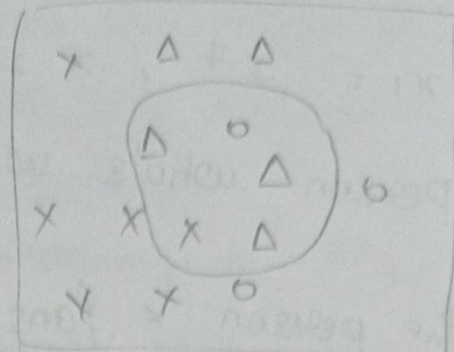
$X_2 = 20$, $y_2 = 1$

Assume K = 3

NAME	AGE	GENDER	class fan	
A	32	M	X1	$\sqrt{(-12)^2 + 0} = \sqrt{144} = 12$
B	40	M	X1	$\sqrt{(-20)^2 + 0} = \sqrt{400} = 20$
C	16	F	N2 $\boxed{X2}$	$\sqrt{(4)^2 + (1)^2} = \sqrt{17} = 4.123$
D	14	F	X1	$\sqrt{(6)^2 + (1)^2} = \sqrt{37} = 6.08$
E	55	M	X2	$\sqrt{(-35)^2 + 0} = \sqrt{1225} = 35$
		M	X1	$\sqrt{} = 20.$
F	40	F	N1 $\boxed{X2}$	$\sqrt{1} = 1$
G	20	M	N3 $\boxed{X2}$	$\sqrt{(5)^2 + 0} = \sqrt{25} = 5$
H	15	F	X1	$\sqrt{1225 + 1} = \sqrt{1226} = 35.01$
I	55	M	N3 $\boxed{X2}$	$\sqrt{(5)^2 + 0} = \sqrt{25} = 5$
J	15			

KNN - Voter Example: REAL TIME EXAMPLE

The given value of $\boxed{K = 5}$



The Majority is Δ from the selected 5 sample.

4. Curse of Dimensionality:

→ KNN Depends on the Distance.

→ Size of the Data Space grows exponentially with a no. of Dimensions.

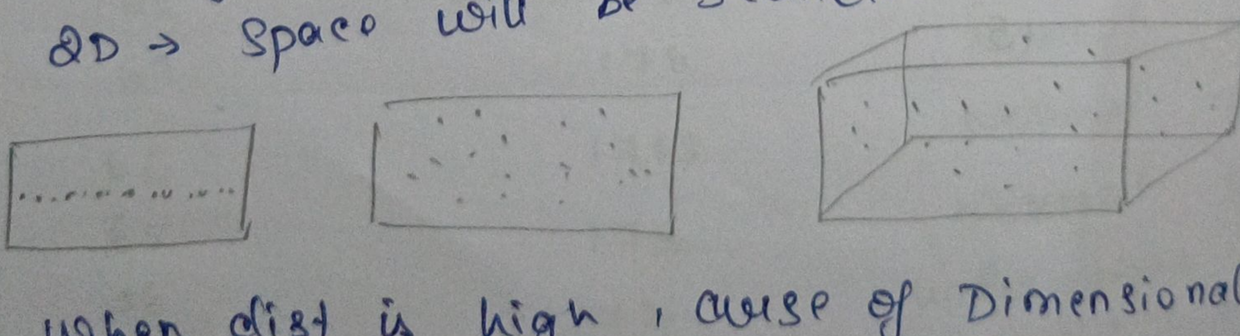
→ Curse of Dimensionality is to understand when KNN Determine's its Predictive power with a increase in no. of variables required for Prediction

→ 1 D, 2 D, 3 Dimensionality.

↓

Dist of the Data points will be less.

2D → Space will be scattered.



When dist is high, curse of Dimensionality will be high & vice-versa.

Posterior probability - Likelihood · Prior prob of proposition

---

Prior probability of evidence.

Ex: Consider the given weather dataset with attributes outlook, temp, windy, Humidity and class label play with values (yes, No). find a person can play a game or not with given weather conditions.

Outlook = Sunny.

Temp = Cool.

windy = strong 2

humidity = high.

S.NO	OUTLOOK	TEMP	HUMIDITY	WINDY	PLAY
1.	Sunny	hot	high	weak	N
2.	Sunny	hot	high	Strong	N
3.	overcast	hot	high	weak	Y
4.	Rainy	mild	high	weak	Y
5	Rainy	Cool	Nominal	weak	Y
6.	Rainy	Cool	Nominal	Strong	N
7.	overcast	Cool	nominal	strong	Y
8.	Sunny	Mild	high	weak	N
9.	Sunny	Cool	nominal	weak	Y
10.	Sunny	Mild	nominal	strong	Y
11.	Sunny	Mild	nominal	Strong	Y
12,	overcast	Mild	high	strong	Y
13,	overcast	hot	nominal	weak	Y
14.	rainy	Mild.	high	Strong	N

Overall probability of yes P(Y) = $\frac{9}{14}$.

Overall probability of NO P(N) = $\frac{5}{14}$

Outlook:-	Y	N	P(Y)	P(N)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0
Rainy	3	2	3/9	2/5

Temperature:	Y	N	P(Y)	P(N)
hot	2	2	2/9	2/5
Mild.	4	2	4/9	2/5
Cool	3	1	3/9	1/5

humidity:	Y	N	P(Y)	P(N)
high	3	4	3/9	4/5
Nominal	6	1	6/9	1/5

windy	Y	N	P(Y)	P(N)
weak	6	2	6/9	2/5
strong	3	3	3/9	3/5

$P(x \mid play = yes) \cdot P(play = yes)$

$$2/9 \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$$

$$= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = \frac{486}{91854} = 5.29 \times 10^{-3}$$

$$= 0.00529 //$$

$P(x \mid play = no) \cdot P(play = no)$

$$\frac{3}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{4}{5} \times \frac{5}{14} = \frac{180}{8750} = 0.0205 //$$

Now we need to find $P(x) = P(yes) + P(No)$

$$= P(x \mid play = yes) + P(x \mid play = No)$$

$$= 0.00529 + 0.0205$$

$$P(x) = 0.0253$$

As per naive bayes we need to find

$$P(play = yes \mid x)$$

$$P(play = yes \mid x) = \frac{P(x \mid play = yes) \cdot P(play = yes)}{P(x)}$$

$$= \frac{0.0053}{0.0253} = 0.209$$

$$P(play = No | x) = \frac{P(x | play = no) \cdot P(play = No)}{P(x)}$$

$$= \frac{0.0205}{0.0253} = 0.810$$

Probability of No is higher.

$\therefore$ the player cannot play.

27/3 find the person can go for play or not when Outlook = rainy, temp = mild, humidity = nominal windy = weak.

Overall probability of yes

$$P(Y) = \frac{9}{14}$$

Overall probability of N

$$P(N) = \frac{5}{14}$$

Outlook	Y	N	P(Y)	P(N)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0
Rainy	3	2	$\boxed{3/9}$	$\boxed{2/5}$

Temperature	Y	N	P(Y)	P(N)
hot	2	2	2/9	2/5
mild	4	2	$\boxed{4/9}$	$\boxed{2/5}$
cool	3	1	3/9	1/5

humidity	Y	N	P(Y)	P(N)
high	3	4	3/9	4/5
nominal	6	1	$\boxed{6/9}$	$\boxed{1/5}$

windy	Y	N	P(Y)	P(N)
weak	6	2	$\boxed{6/9}$	$\boxed{2/5}$
strong	3	3	3/9	3/5

$P(x \mid play = yes) \cdot P(play = yes)$

$$\frac{9}{14} \times \frac{3}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} = \frac{8}{21 \times 9} = 0.042$$

$P(x \mid play = NO) \cdot P(play = NO)$

$$\frac{2}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = \frac{40}{5 \times 5 \times 5 \times 5 \times 14} = 0.0045$$

Now we need to find $P(x)$

$= P(yes) + P(NO)$

$= P(x + play = yes) + P(x + play = NO)$

$= 0.042 + 0.0045 = 0.0465$

As per naive bayes

$$P(play = yes/x) = \frac{0.042}{0.0465} = 0.9032$$

$$P(play = No/x) = \frac{0.0045}{0.0465} = 0.0967$$

$\therefore$ yes > NO

he has probability of playing.

find the patient x having flu or not.

Cold = yes

running nose = No

headache = mild

fever = yes

S.NO	COLD	RUNNING NOSE	HEADACHE	FEVER	FLU
1	Y	N	mild	Y	N
2	Y	Y	N	N	Y
3	Y	N	strong	Y	Y
4	N	Y	mild	Y	Y
5	N	N	N	N	N
6	N	Y	strong	Y	Y
7	N	Y	strong	N	N
8	Y	Y	mild	Y	Y

Overall probability = 8.

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

Cold	Y	N	P(Y)	P(N)
Yes	3	1	3/5	1/3
No	2	2	2/5	3/3

Running nose	Y	N	P(Y)	P(N)
Yes	4	1	4/5	1/3
NO	1	2	1/5	2/3

Headche	Y	N	P(Y)	P(N)
mild	2	1	2/5	1/3
NO	1	1	1/5	1/3
Strong	2	1	2/5	1/3

Fever	Y	N	P(Y)	P(N)
Y	4	1	4/5	1/3
N	1	2	1/5	2/3

$P(x \mid flu = yes) \cdot P(flu = yes)$

$$\frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{5}{8} = \frac{120}{5000} = 0.024$$

$P(x \mid flu = NO) \cdot P(flu = NO)$

$$\frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{3}{8} = \frac{6}{648} = 0.009259$$

According to naive bayes
we need to find P(x)

$$P(x) = 0.024 + 0.009259 = 0.0332$$

$$P(flue = yes \mid x) = \frac{0.024}{0.0332} = 0.7228$$

$$P(flu = No \mid x) = \frac{0.009259}{0.0332} = 0.2969$$

$\therefore$ yes > No He gets flu.