gcr : rmgwvna

Lecture # 1 20 Aug, 24

Ass 1: data scrapping of football

rows ⇒ instances
e.g names of diff countries

columns ⇒ features
e.g genders, age

* read NNI for medical imaging

NLU net ⇒ can take any type of image

Types of AI :

1. Artificial Narrow Intelligence
2. Generative AI ⇒ Chat gpt
3. Super AI ⇒ we haven't reached this stage yet

Data Science

Scanned with CamScanner

22-8-24

Lecture # 2

* analyst has more work in processing, feature related
* data scientist ne poora cycle explore kiya hota hai

descriptive ⇒ houses not selling, air lift is not working

· WHAT HAPPENED OR WHAT IS HAPPENING?

diagnostic ⇒ HOW IT HAPPENED?

Scanned with CamScanner

Date 27 Aug, 24

Lecture #3

* primary method of collecting data is more challenging
* focus groups: teams jinka focused target ho, not so common in pakistan
* static websites: data change nai hota with time and you do not need to change it.
* dynamic websites: info is keep changing like youtube viewers, subscribe
* when you are unable to get data through API, you move towards web scrapping.
* web scrapping can done only in websites that allow you to.
* there are google chrome extentions for scrapping

H.W play with notebook

* In fabt website, every faculty is a card.

H.W we want to add education wala column as well through id tag.

Scanned with CamScanner

29 Aug, 24.

Lecture # 4

scrapping: getting HTML page as a request

* Fast website is Shut. There is no standard in format of education.

strip ⇒ starting ending spaces to remove

we a converting unstructured to structured data through scrapping.

Data types:

qualitative
- ordinal : there is an order but not a number
- nominal : there is not an order neither a number

Lecture # 5

Date 3 Sept, 24.

	very satisfied	satisfied	not satisfied	very unsatisfied
	4	3	2	
we dont want to remember these ←	0100	0011	0010	0001
one hot encoding ←	1	0	0	0

Interval ⇒ grade (not meaningful 0 bcz no one can get an abs 0 not even nobita)

* here grade represents intervals e.g 80 – 85 ⇒ A
* if we were just discussing about grades it would've been ordinal

* we using selenium bcz its best and we dealing with dynamic data, site pe scripts hain data nai, so it extracts data from scripts.
we can use playwright, pupetear as well
* beautifulsoup is used in parsing etc.

* standard deviation is sigma. 3 sigma se bahir ka data will be considered an outlier.

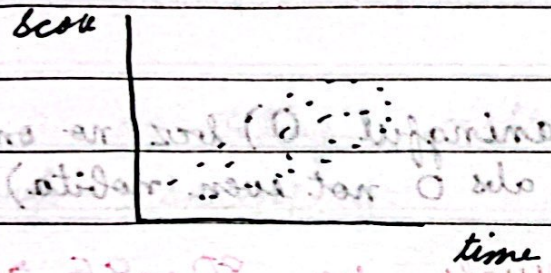
5th Sept, 24.

Lecture #6

nominal ⇒ categorical

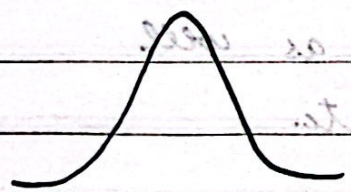
ordinal ⇒ categorical + order

ratio data:

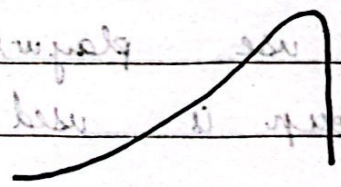


inferential statistical: you're trying to come up with conclusion

percentile: relevant to other points; if 96% is highest score, percentile will be 100%



best criteria ⇒ mean



best criteria ⇒ median

* mean cannot be applied on categorical data
* median doesnt get affected by outlier

Date

* for variance use 'n' if you know the mean, and use n-1 if you know estimated mean.
* for a perfect symmetrical data mean, median & mode are equal

→ Removing outliers:

* 3 sigma k spread se bahir ka data = outlier
this is the privelledge of standard deviation

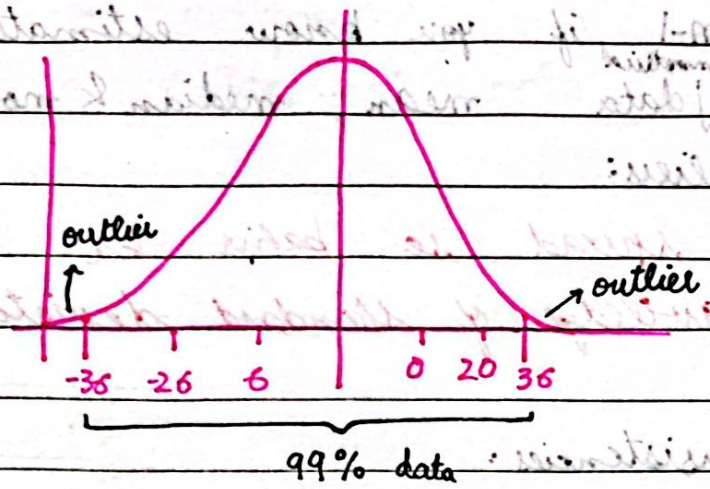
⇒ Removing inconsistencies:
dates k diff formats

10 Sep, 24.

Lecture #7

$$Z = \frac{x - \mu}{\sigma}$$

$$-3 \le Z \le 3$$



Z-score can only detect outlier if we have normal distribution

* linear regression can be used to interpolate missing values
* if missing value is target variable then skip.
* global constant can be used but what favourable would a constant be over a variable one.
* if a feature has 0 correlation with target variable, we remove it.

Data Cleaning :

1. first sort the data : 2, 10, 18, 18, 19, 20, 22, 25, 32

2. • For equi-width

b1 = {2, 10}

b2 = {18, 18, 19, 20}

b3 = {22, 25, 32}

• For equi frequency

b1 = {2, 10, 18}

b2 = {18, 19, 20}

b3 = {22, 25, 32}

has bin mai equal no. of elements

Scanned with CamScanner

Lecture # 8

Date 15 Sept

Smoothing:

Smooth by mean → replace each element by mean

smooth by median → replace each element by median

smooth by boundary → [min, replace with min/max, max acc to least distance]

$$\mu = \frac{20+30+65+80+85+90+110}{7}$$

$$= 68.57 = 68.6$$

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

$$= \frac{1}{7}\left[(20-68.6)^2+(30-68.6)^2+(65-68.6)^2+(80-68.6)^2+(85+68.6)^2+(90-68.6)^2+(110-68.6)^2\right]$$

$$= \frac{1}{7}\left[2362+1440+13+130+269+455+1714\right]$$

$$= \frac{1}{7}\left[6433\right] = 919.4$$

$$\sigma = \sqrt{\text{var}} = 28.6$$

now we can calculate z-score of each point

$$z_1 = \left(\frac{20-68.6}{28.6}\right),\quad z_2 = \left(\frac{30-68.6}{28.6}\right),\quad z_3 = \left(\frac{65-68.6}{28.6}\right),$$

$$z_4 = \left(\frac{80-68.6}{28.6}\right),\quad z_5 = \left(\frac{85-68.6}{28.6}\right),\quad z_6 = \left(\frac{90-68.6}{28.6}\right),$$

$$z_7 = \left(\frac{110-68.6}{28.6}\right)$$

Absent work

15 Sept, 24.

Lecture #8

standard:
$-3 \le z \le 3$

given
$-2 \le z \le 2$

z-score:

data set : [20, 30, 65, 80, 85, 90, 110]

$$\boxed{z = \frac{x-\mu}{\sigma}}$$ μ = mean, σ = standard deviation

$$\mu = \frac{20+30+65+80+85+90+110}{7}$$

$$= 68.57 = 68.6$$

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

$$= \frac{1}{7}\Big[(20-68.6)^2+(30-68.6)^2+(65-68.6)^2+(80-68.6)^2+(85-68.6)^2$$

$$+(90-68.6)^2+(110-68.6)^2\Big]$$

$$= \frac{1}{7}\left[2362+1490+13+130+269+458+1714\right]$$

$$= \frac{1}{7}\left[6436\right] = 919.4$$

$$\sigma = \sqrt{var} = 28.6$$

now we can calculate z-score of each point

$$z_1 = \left(\frac{20-68.6}{28.6}\right), \quad z_2 = \left(\frac{30-68.6}{28.6}\right), \quad z_3 = \left(\frac{65-68.6}{28.6}\right),$$

$$z_4 = \left(\frac{80-68.6}{28.6}\right), \quad z_5 = \left(\frac{85-68.6}{28.6}\right), \quad z_6 = \left(\frac{90-68.6}{28.6}\right),$$

$$z_7 = \left(\frac{110-68.6}{28.6}\right)$$

Scanned with CamScanner

Date

$z_1 = -1.69$, $z_2 = -1.34$, $z_3 = -0.12$, $z_4 = 0.39$,
$z_5 = 0.57$, $z_6 = 0.74$, $z_7 = 1.447$

* only in case of symmetric distribution we can use z-score to remove outliers.

Binning:

1. Equal depth: data should be sorted
bin size = 3
total elements = 7 thus $7/3 = 2.3 = 2$
Bin 1 [20, 30] Bin 2 [65, 80] Bin 3 [85, 90, 110]
smoothing:
Bin 1 ⇒ mean, median same
Bin 2 ⇒ mean, median same
Bin 3 ⇒ mean : [95, 95, 95]
median : [90, 90, 90]
boundary : [85, 85, 110]

2. Equal width: data should be sorted
distance = max value − min value = 110 − 20 = 90
width = $\frac{90}{3}$ = 30
threshold for bin 1 : min + w − 1 = 20 + 30 − 1 = 49 ⇒ 20 − 49
Bin 1 : [20, 30]
threshold for bin 2 : 50 + 30 − 1 = 79 ⇒ 50 − 79
Bin 2 : [65]

Scanned with CamScanner

Bin 3 : [80, 85, 90, 110]

* we are forced to put all remaining elements in Bin 3, since size of bin is only 3.

IQR : data should be sorted

→ first calculate median : $Q_2 = 80$

$$[\underbrace{20, 30, 65}_{Q_1=30}, \underset{Q_2}{80}, \underbrace{85, 90, 110}_{Q_3=90 \text{ (upper quartile ka median)}}]$$

$IQR = Q_3 - Q_1 = 90 - 30 = 60$

min : $Q_1 - [1.5 \times IQR] = 30 - [1.5 \times 60] = -60$

max : $Q_3 + [1.5 \times IQR] = 90 - [1.5 \times 60] = 180$

IQR ⇒ $-60 - 180$

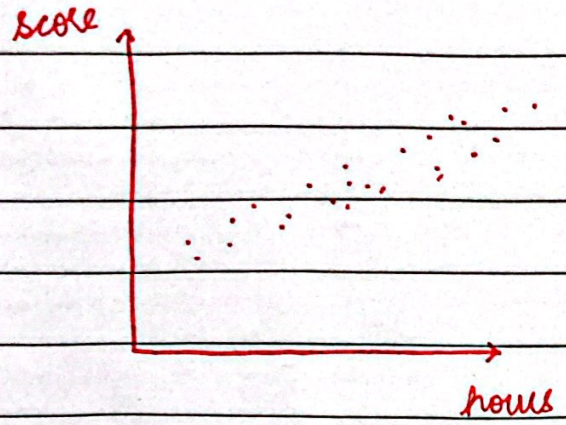
* if any value lies outside this range, it wud be considered an outlier.

Key points :

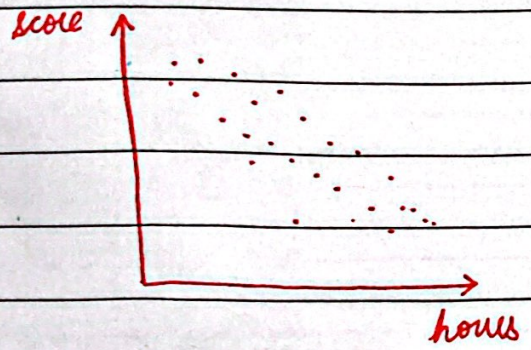
* z-score is only used when we are dealing with data with normal distribution.
* if data is skewed, we do not have an actual mean, in that case, IQR is used bcz it helps us to visualise data on box-plot.

Lecture #9

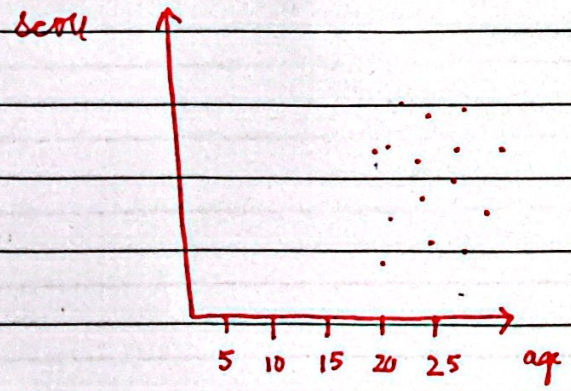
Date 19 Sept, 24



positively correlated



negatively correlated



not co-related

code:

⇒ after every group there is aggregation

H.W

* multiple columns ki merging.

26-11-24

Lecture # 10

x_1	x_2	x_3	Y

* If x_1 and x_2 are strongly correlated and they are correlated with y, we can eliminate one of x_1, x_2.
* If x_1 and x_2 were weakly correlated

Correlation formula:

a_i, b_i ⇒ diff variables

* we are taking their diff with their mean on each point
* $n-1$ for sample data, n for population
* correlation coefficient narrows down our result to $-1 - 1$

-1 ⇒ negatively correlated but strong

1 ⇒ +vely correlated but strong

* 0 ⇒ doesnt mean they are independent and not related instead it just means they are not linearly correlated

e.g $Y = X^2$

Min-Max Normalisation:

* apkay dataset ka min max is your new range
* we could take $-10, 10$ as well

H.W Z-score normalisation (next slides) + multivar

Lecture # 11

Date 1-10-24

Decimal Scaling:

1. $V' = \frac{V}{10^j} = \frac{73000}{10^5} = 0.73$

2. if there was an array, j would be digits of highest value (abs value)

[-10,000, -80, 1230, 110, 2, 37652] $\Rightarrow j = 5$

$\frac{-10,000}{10^5}$ $\frac{-80}{10^5}$ ...

$y = b_0 + b_1x_1 + b_2x_2 + \dots b_nx_n$

target variable (dependent) ← y

x_n → labels (independent)

$SSR = \sum_i (\hat{y}_i - \bar{y})^2$ ⇒ estimated values

$\hat{y}$ → estimated

$\bar{y}$ → real y values ka mean

$0 \le r^2 \le 1$

SSE ⇒ error

$y = mx + c$

$r = (\text{Sign of } b_1)\sqrt{\text{coeff of det.}}$

direction/intercept of slope ⇒ m

Normalisation from z-score:

* we use this as replacement from original point
* +ve ⇒ above the mean
* -ve ⇒ below the mean
* applied on normally distributed

$x_1 = \frac{x_1 - \mu}{\sigma}$

new

* if not normally distributed
⇒ modified z-distribution
• MAD ⇒ median of absolute differences
• mean not suitable here

$$Z_{mod} = \frac{x - median}{MAD}$$

Kurtosis: ⇒ for peakness

gausian shape ⇒ symmetric
kurtosis < 3 ⇒ flattened
kurtosis > 3 ⇒ sharper peak

* sltosa is very unique. it can be classified easily, baki two mai classification would be a bit tough.

* respect is only earned when given.

* you shud have guts to accept your mistakes and own them.

Scanned with CamScanner

Lecture # 12 Date 3rd Oct, 24

Chi square test: categorical against categorical

itni strength nai bata, dependencies bata hai

null hypothesis ko chi square test ki base pe reject ya accept kia jata hai

* Chi square distribution:

jise jese degree of freedom increasing, we are obtaining bell shaped

$$E_{Male, Republican} = 109$$

$$= \frac{(100-109)^2}{109} \qquad \frac{(O-E)^2}{E}$$

$$= 0.743$$

* if table wali critical value (threshold) is less than our value obtained from chi square test, we reject our null hypothesis.

Libraries for this topic

⇒ Mat plot

⇒ C block

Scanned with CamScanner

10-10-24

Lecture #13

Scatter plot:
general trend b/w two entities

Line Chart:
for error versus epochs while determining accuracy

Pie Chart / Donut chart: *
for commulative / propotional analysis

Spider chart:
⇒ Difficult, not very used
⇒ focused on region

Bubble graph:
Bubble size would represent a third variable in our case it is profitability

Makeup:
209 ⇒ h.w run notebook

Scanned with CamScanner

Makeup Lecture

Lecture # 14

Date 11-10-24

Dimensionality Algo:

1. Feature Selection :

⇒ All pixels are in 2D array, and we're picking values from it, might be we're picking alternative values.

⇒ data set is als given, we're choosing values from it

2. Feature Extraction :

⇒ reduction or shrink the data set

⇒ transforming data as in applying an algorithm to our values

⇒ for example in CNN

⇒ Principal Component Analysis for numerical data.

Scanned with CamScanner

15-10-24

Lecture # 15

curse of dimensionality :

⇒ overfitting

⇒ generalisation of the model

Clausi-Constant: They are around 90% same but not 100% so they are not constant and they won't get 0 variance. They have low variance.

Mutual information:

⇒ jiska info gain acc to our classification target zada ayegi, we will pick that feature

⇒ we're judging that among two variables that are correlated with our target, which one to pick

Entropy : checks impurity kitni hai, the more the randomness, the more the entropy, prediction is hard.

⇒ if entropy is kum, we have low info

* we're not happy in either case, we do not want a constant, neither do we want too much randomness.

Scanned with CamScanner

Date:

Labelisation	One-hot encoding
* when we are concerned with rank/order	* when order is not concerned

Example #2 in lecture #13&14 slide:

Formulas to remember:

- $H(X) = -\sum_i (P(x_i) \log_2 (P(x_i)))$
- $H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2 (P(x_i|y_j))$

information gain ←
- $IG(X|Y) = H(X) - H(X|Y)$

H(X):

⇒ H(parent):

$P_1 = 3/8$

$P_0 = 5/8$

using formula 1

$$H(\text{parent}) = -[P_1 \log_2 P_1 + P_0 \log_2 P_0]$$

$$= -\left[\frac{3}{8} \frac{\log(3/8)}{\log 2} + \frac{5}{8} \frac{\log(5/8)}{\log 2}\right]$$

$H(\text{parent}) = 0.954$

⇒ $H(X) = 0.954$

-0.53

H(X|Y) :

⇒ H(parent | hair) :

Hair

1 | 2 | 3

$w_1 = 4/8$

Hair=1	Result
1	1
1	0
1	1
1	0

$P_1 = 1/2,\ P_0 = 1/2$

$H(Hair=1) = -\left[\frac{1}{2}\frac{\log 1/2}{\log 2} + \frac{1}{2}\frac{\log 1/2}{\log 2}\right]$

$= 1$

$w_2 = 3/8$

Hair=2	Result
2	0
2	0
2	0

$P_1 = 0,\ P_0 = 3/3 = 1$

$H(Hair=2) = -\left[0\frac{\log 0}{\log 2} + \frac{1\log 1}{\log 2}\right]$

$= 0$

$w_3 = 1/8$

Hair=3	Result
3	1

$P_1 = 1,\ P_0 = 0$

$H(Hair=3) = 0$

$H(parent|Hair) = w_1(Hair=1) + w_2(Hair=2) + w_3(Hair=3)$

$= \frac{4}{8} \times 1 + \frac{3}{8} \times 0 + \frac{1}{8} \times 0$

$= \frac{4}{8} = 0.5$

IG (X|Y) :

⇒ IG (parent | hair) :

$IG(parent|hair) = H(parent) - H(parent|hair)$

$= 0.954 - 0.5 = 0.454$

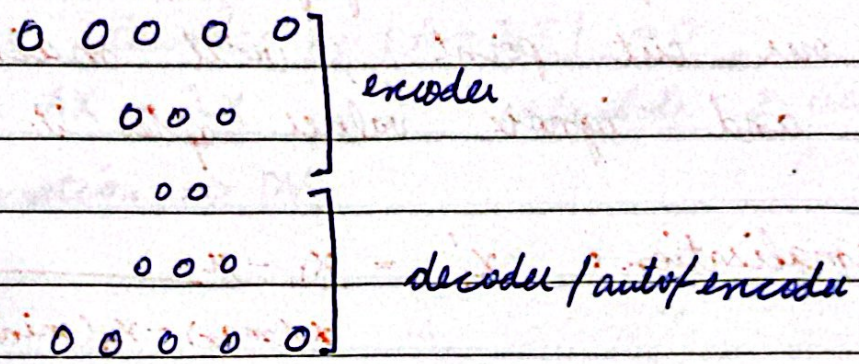
Lecture # 16

Date 17-10-24

- categorical vs categorical

⇒ chi-sq test

⇒ info gain



- Neural Network

$$VIF = \frac{1}{1-R^2}$$ → correlation coefficient ka sq.

PCA :

jitne features utne components

22-10-24

Lecture # 17

PCA:

⇒ Muli colinearity

- $cov(x,x)$ ⇒ variance of x

* scree plot ⇒ jahan se agay graph flat ho, that will be our cut point, we'll maintain data before it and ignore values after it

* mean normalisation: $x_1' = \dfrac{x_1 - \mu}{x_1(max) - x_1(min)}$

* but we using:

$$x_1' = x_1 - \mu_1$$
$$x_2' = x_2 - \mu_2$$

* $cov(x,y) = cov(y,x)$

$$cov(\overset{SBP}{x}, \overset{DBP}{y}) = \frac{(-3 \times -4) + (-1 \times -2) + (-1 \times 0) + (1 \times 0) + (1 \times 2) + (3 \times 4)}{5}$$

* we need to find values of λ, and for each λ, vector will be made

* lemdas represent components

Scanned with CamScanner

Date

* V ⇒ eigen vector matrix D ⇒ Data
* vectors will always be perpendicular
* DV ⇒ transformation

Steps of PCA:

1. center the data ⇒ D
2. calculate co-variances and variances
3. calculate lemdas by $|A - \lambda I| = 0$
4. apply $AX = \lambda X$ and calculate 2 vectors and form V
5. transformation ⇒ DV

* variance of PC1 is basically lemda1.

* either select eigen values greater than 1 or select according to commulative variance

* open cv opens files in BGR
* open cv doesn't have crop function

Scanned with CamScanner

Lecture # 18

Date 24-10-24

No Notes Today

just to give u a flavour

29-10-24

Lecture # 19

corpus ⇒ textual data

accent: not pure english but available in ~~english~~ word

if TF is higher ⇒ the term is very frequently used in the document

if IDF is higher ⇒ the term is very unique across documents

if TF×IDF ⇒ term is important bcz it is frequently used in a document but rarely used accross other documents

base of Bert was sentimental analysis

* drawbacks of BOW:

⇒ still sparcity ⇒ too much space

⇒ if a word has great counts, doesn't mean it's important, it can be a stop word.

* word embedding is mostly used in textual data

Scanned with CamScanner

31-10-24

Lecture # 20

linear regression :

hypothesis

$y_i' = y_{pred} = h(\theta)$

slope input

$y = mx + c$

$\Rightarrow y = \theta_1 x + \theta_0$

* error : $y - y_{pred}$

* gradient descent is an optimizer of cost function
  $\Rightarrow$ parameters kis tha optimize kere hain

* linear regression mai minima ki hee baat hoti hogi since we need to minimize cost function

* if two variables

$\Rightarrow y = \theta_1 x_1 + \theta_2 x_2 + \theta_0$

$\alpha \Rightarrow$ learning rate
mostly
0.01 – 0.1

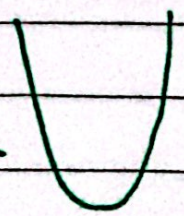
- $\theta_0' = \theta_0 - \alpha \dfrac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_0}$

$\theta_1' = \theta_1 - \alpha \dfrac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_1}$

$\theta_2' = \theta_2 = \alpha \dfrac{\partial J(\theta_0, \theta_1, \theta_2)}{\partial \theta_2}$

* learning rate should be adaptive

convex cost function $\rightarrow$



Scanned with CamScanner

Date

⇒ when cost function is in convex form, start mai α greater lein gy, then ahista ahista low values

Types of gradient descents

① batch training: ⇒ slow

* by default gradient descent is batch gradient descent means sab pe error niklay ga, then updation hogi. here epoch = iteration

poora training set dikhae ga ... weight updation kitni baar hui hai

⇒ sara training set model ko kitni baar dikhaya hai

② stochastic batch:

⇒ fast

⇒ har example pe cost and sath hee updation

⇒ determination of actual minima difficult

③ mini batch:

⇒ chotay batches

⇒ cost nikala

⇒ updation

Example

	x_1	x_2	y
1	1	6	1
2	0	5	0
3	1	4	1
4	2	0	1
5	1	1	2
6	3	1	2

$\theta_0 = \theta_1 = \theta_2 = 0.1$

$\alpha = 0.01$

Scanned with CamScanner

ale

1. Batch gradient descent : too slow, too much computational cost

$y^1_{pred} = 0.1 + 0.1 \times 1 + 0.1 \times 6$

$y^2_{pred} = 0.1 + 0.1 \times 0 + 0.1 \times 5$

$\vdots$

$y^6_{pred} = 0.1 + 0.1 \times 3 + 0.1 \times 1$

when epoch = 1
iteration = 1

$$\theta_0' = 0.1 + \frac{0.01}{6}\begin{bmatrix}(1-y^1_{pred}) + \\ (0-y^2_{pred}) + \\ (1-y^3_{pred}) + \\ (2-y^4_{pred}) + \\ \dots\end{bmatrix}$$

$$\theta_1' = 0.1 + \frac{0.01}{6}\begin{bmatrix}(1-y^1_{pred})1 + \\ (0-y^2_{pred})0 + \\ \dots\end{bmatrix}$$

$$\theta_2' = 0.1 + \frac{0.01}{6}\begin{bmatrix}(1-y^1_{pred})6 + \\ (0-y^2_{pred})5 + \\ \dots\end{bmatrix}$$

Here we gave the whole data set and calculated predicted values against it. Since whole data set is used once, so epoch = 1. Also we updated the weights once, so iteration = 1

2. stochastic gradient descent : too fast, not efficiently learning

$y^1_{pred} = 0.1 + 0.1 \times 1 + 0.1 \times 6$

$E = (y - y_{pred})^2$

when epoch = 1
iterations = 6

$$\theta_0' = 0.1 + \frac{0.01}{6}\left[(1-y^1_{pred})\right]$$

$$\theta_1' = 0.1 + \frac{0.01}{6}\left[(1-y^1_{pred})1\right]$$

$$\theta_2' = 0.1 + \frac{0.01}{6}\left[(1-y^1_{pred})6\right]$$

Scanned with CamScanner

Date: 2-11-2

here, we just provided first value in the data set so epoch = 0. but we updated all weights according to that value so iteration = 1. Likewise we will do this procedure on all 6 values and then epoch = 1 but iterations = 6.

3. Mini batches gradient descent: best

Ex $\begin{cases} y^1_{pred} = 0.1 + 0.1 \times 1 + 0.1 \times 6 \\ y^2_{pred} = 0.1 + 0.1 \times 0 + 5 \times 0.1 \end{cases}$

$\boxed{\text{when epoch} = 1 \quad \text{iterations} = 3}$

$$\theta_0' = 0.1 + \frac{0.01}{6}\begin{bmatrix}(1-y^1\ pred) + \\ (0-y^2\ pred)\end{bmatrix}$$

$$\theta_1' = 0.1 + \frac{0.01}{6}\begin{bmatrix}(1-y^1 pred)1 + \\ (0-y^2 pred)0\end{bmatrix}$$

$$\theta_2' = 0.1 + \frac{0.01}{6}\begin{bmatrix}(1-y^1 pred)6 + \\ (0-y^2 pred)5\end{bmatrix}$$

Here we are picking up data values in pair. In stochastic we were taking values 1 by 1 so it wasn't that efficient but here we are taking 2 values at a time and then updating array weights so after three iterations 1 epoch will get completed.

Scanned with CamScanner

7-11-24

Lecture 21.

linear regression problem itne hrs padhne pe itne number

e.g. 1. 30 hrs → 5 marks

2. 60 hrs → 15 marks

classification problem if 30 hrs study how many chances k he will pass? probability involved.

you're not interested in number. you're interested in the threshold jisse upar you're passing and neche you're failing.

⇒ different threshold values

⇒ then check area

more area → unbiased classification

sigmoid : uses linear function so features should be linearly separable

$\theta^T x$ ⇒ x ⇒ array for features

θ ⇒ array of vectors

hypothesis : prediction

Lecture # 22

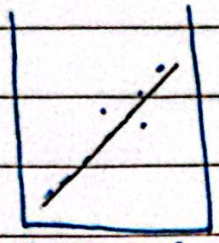
$0.1 \times \frac{1}{4}\begin{bmatrix}(0-0.5) + \\ (1-0.5) + \\ (1-0.5) + \\ (0-0.5)\end{bmatrix}$

Date 12-11-24

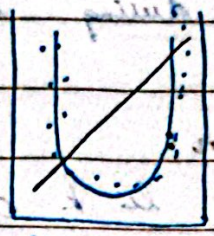
If for two variables: $b_0 + x_1 b_1 + x_2 b_2 + c_0 + c_1 z_1 + c_2 z_2$

if for one variable: $b_0 + x_1 b_1 + x_2$

- Poly regression ⇒ when linear regression is not suitable

- 

linear regression ✓



linear regression ✗ ⇒ underfitting

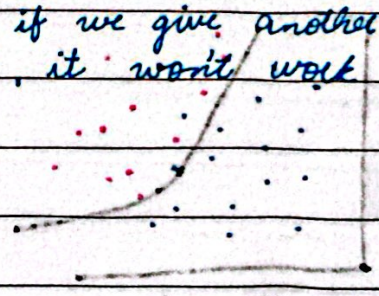
poly ✓✓

-

fig 1

overfitting ⇒ model is going through each and every point

* here the model is overtrained in such a complex data. if we give another data to the model, it won't work



- bias and variance:

⇒ high variance low bias : over fitting

⇒ low variance high bias : under fitting

- if overfitted data, how to introduce regularity

⇒ $\boxed{\text{lasso regression}}$

$\lambda \sum_{j=1}^{p} |b_j|$ ⇒ func of regularity

b_j ↓ weights

⇒ when we take gradient descent of this, this will also contribute

Scanned with CamScanner

advantage:
it is selecting good features ⇒ shrinking

disadvantage:
⇒ highly dependent on lamda
⇒ if we give low value weight to important data, it will diminish it during learning process

Ridge regression:
- sab features ko li k chalay ga
- sabse chotay feature ko bhi weight dega
- in lasso, low valued weights will diminish in just starting iterations (0 out quickly kideta hai)
- ridge gradually does this

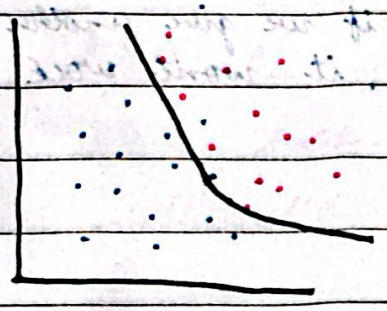
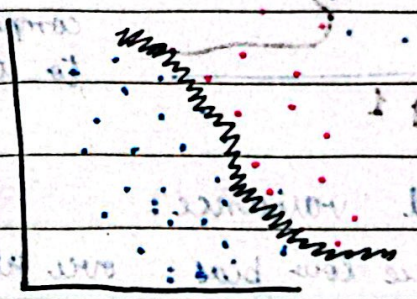


figure 1
now becomes



stochastic

vanishing gradient: learning rukti hai
exploding gradient: too much learning → drastic change in weights

Lecture # 23

Date 14-11-24

moving average: time series data, to judge overall behaviour.

weighted average: how the optimizer smooth out previous paths to determine new one – Model term / Machine learning term

* both are same in formula

Adam: a mixture/combo gradient descent with momentum and RMSprop.

Momentum-based gradient descent:

$\beta \Rightarrow$ decay rate $\Rightarrow$ how much past you want to remember

$V_{t-1} \Rightarrow V_{t-1}$ is previous rate

$\nabla_\theta J(\theta) \Rightarrow$ gradient

Example:

$J(\theta) = \frac{1}{2}\theta^2$ $\Rightarrow$ cost function $\Rightarrow$ derivative

$\alpha = 0.1$

$\beta = 0.9$ $\qquad J(\theta) = \frac{1}{2}2^2 = 2$

$\theta = 2.0$

initial velocity $V_0 = 0$

$V_1 = V_t = 0.9(0) + (0.1)(2)$

$V_1 = 0.2$

$\theta_1 = 2 - 0.1 \times (0.2)$

AdaGrad:

$\Rightarrow$ small gradients $\Rightarrow$ greater alpha $\Rightarrow$ take large steps in learning

$\Rightarrow$ and vice versa

Scanned with CamScanner

1) weighted updation (smoothening)

2) Adaptive learning

Adam combined these two

Lecture # 24

Date 21-11-24

high variance ⇒ overfit ⇒ rattu tota ⇒ no generalisation

high bias ⇒ data nai samaj pa rahay

* dono high hona likely impossible.

* ideal situation ⇒ both should be low

cross validation: generally we used 20% test and 80% train on data set (or 70%, 30%)
but here we further divide training into 80% train and 20% validation (or 70%), 30%)
in validation we do not update weights, we just do forward pass

k fold: just to judge model's performance locally, upar se
k ⇒ no of folds, generally 5 or 10

leave one out hai example ko exclude ke k training ⇒ too much costly

* sigmoid and hyperbolic activation functions were good but there was problem of vanishing gradient, that is why we use relu.
* drawback of 0 is that in deep networks, it converts 0 into negative values to 0 and too many 0s can cause dying relu problem.

Scanned with CamScanner

21-11-24

Lecture # 25

CNN :

* if image is $6 \times 6 \times 16$, 16 is no of channels here

* CNN has been performed on all types of datasets, it worked best on images.

Example :

$$I = \begin{bmatrix} 1 & 1 & 2 & 4 & 3 \\ 4 & 2 & 1 & 1 & 1 \\ 3 & 4 & 2 & 3 & 2 \\ 2 & 1 & 4 & 2 & 1 \\ 3 & 3 & 2 & 4 & 4 \end{bmatrix} \qquad K = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

→ padding = NO

→ stride = 1

$$\text{New weight} = \frac{(\text{old weight} + 2p - F)}{\text{stride}} + 1$$

$$= \frac{5 + (2 \times 0) - 3}{1} + 1 = 3$$

So dimensions of new matrix would be 3×3

Lecture # 26

Date 26-11-24

* most used pooling are max and avg in LeNet-5 AlexNet etc.

* Convolutional layer mai mostly stride is 1.
* Pooling layer mai mostly stride is 2.

* Convolutional layer is mostly for reading while
* pooling layer is for shrinking.

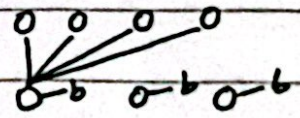
* Pooling doesn't disturb no of maps / channels.

* jitni output classes utne neurons hongy output mai

100 > 100 × 50 weights + 50 biases ← total no of parameters

hidden layers
- 50
- 20
- 10
- 5

1



* valid means no padding.

CNN Dimension Calculation: (from slides)

Conv1: $\frac{64-5+2\times0}{1}+1 = 60$

$(\underbrace{5\times5}_{\text{kernel size}} \times \underbrace{3}_{\text{no of channels}}) \times 8$ → bcz we needed 8 features maps $+ 8$ → bias for each feature map

just to give it a flavour

Pool 1 :

$$\frac{60-2}{2} + 1 \Rightarrow 30$$

Conv 2 :

$$\frac{30-3+2\times 0}{1} + 1$$

* parameter learning is in convolutional layer.

Lect 5 Dimension Calculation (from slides)

POOL2 ↓ we flattening so $400 \times 120 + 120$
FC3

EXTRA

Date

* GAN is mostly used in images and data augmentation.

* Researches are being done on GAN on texts.

LEE-NET → digit classification

BERT → sentiment analysis

Chatgpt 4 by open AI → broke the record of BIRT

Scanned with CamScanner