TM-1 ⊛ what are potential consequence of setting Learning rate high or low?

CSE

$$\boxed{w = w - \alpha \frac{d}{dw} J(w)}$$

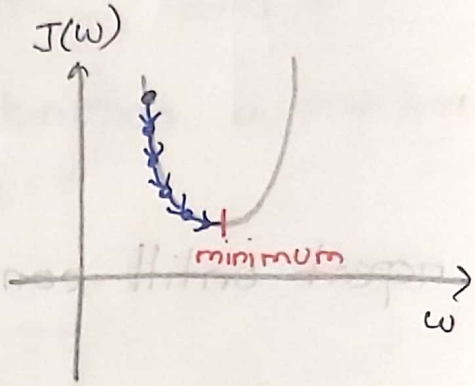
If α is too small

Gradient descent is too small

too slow

stuck in solution

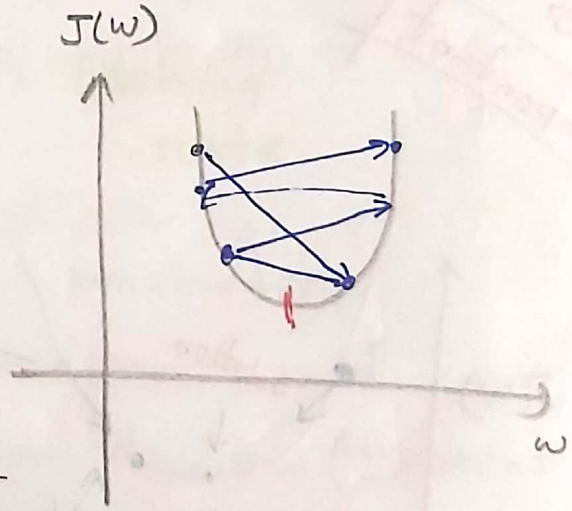
/ long time to reach



if α is too large

Gradient descent may overshoot and never reach the minimum

→ fail to converge, diverge



CSE 16 Fin

Regression	classification
① output variable must be continous nature on real value	② output variable must be descrete value
② we try to find best fit line which can predict the output more accurately	② we try to find decision boundary which can divide dataset into different classes
③ weather prediction, House price pediction	③ Identification of spam, speech Recognition, cancel Identification
④ type :- linear nonlinear	④ type :- Binary Multiclass

CS CamScanner

what are causes of overfitting

(b) CSE

overfitting

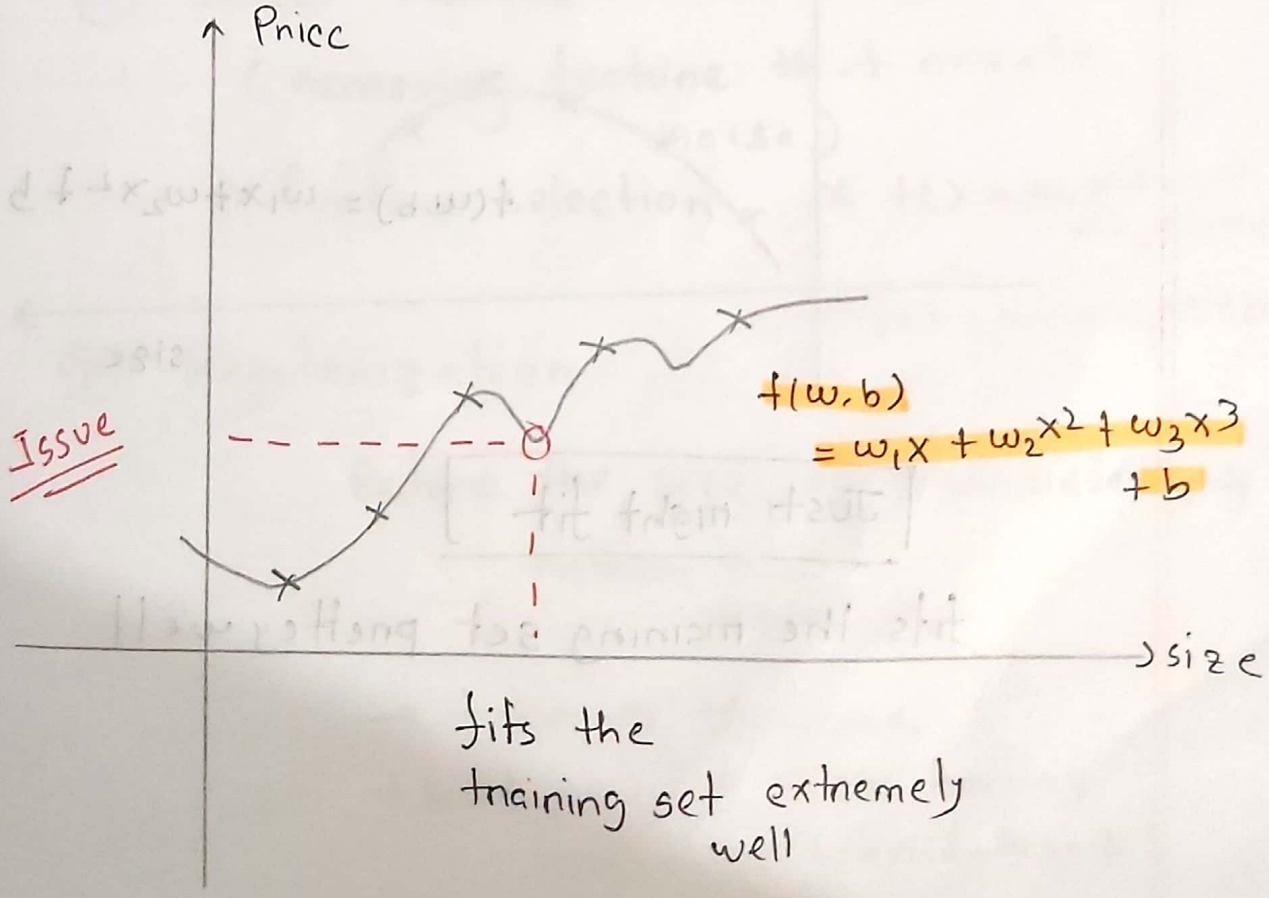
→ when model tries to cover all the data points / more than required data in dataset (unnecessary - irrelevant data)

→ caching noises and inaccurate values

→ Low bias (very complex, random fluctuations)

→ High variance (change prediction significantly)

→ model has too many parameters / feature



fits the training set extremely well

111 How do filters extract features in CNN?

checking a if a digit has a loopy filter?
cincle

-1	1	1	1	-1
-1	1	-1	1	-1
-1	1	1	1	-1
-1	-1	-1	1	-1
-1	-1	-1	1	-1
-1	-1	1	-1	-1
-1	1	-1	-1	-1

*

3x3 filter

1	1	1
1	-1	1
1	1	1

filter for loopy cincle

that means it has cincle

-0.11	1	-.11
-.55	.11	-.33
-.33	-.33	-.33
-.22	-.11	-.222
-.33	-.33	-33

$(-1\times1)+(1\times1)+(1\times1)$
$+(-1\times1)+(1\times-1)+(-1\times1)$
$+(-1\times1)(1\times1)+(1\times1) = -1/9$
$= .11$

feature map

what is Pooling ?

netenes to neducing the spatial dimession (width×height) of Input which helps control the number of panameten and computation in the netwonk while pnesenving Impontant featunc.

1. Reduce dimmesion
2. Reduce computation
3. Reduce Ovenfitting (less panametens)
4. Model is tolenant towands vaniation, distontion

CS CamScanner

Q5 How does pooling simplify them in CNN?

Max pooling

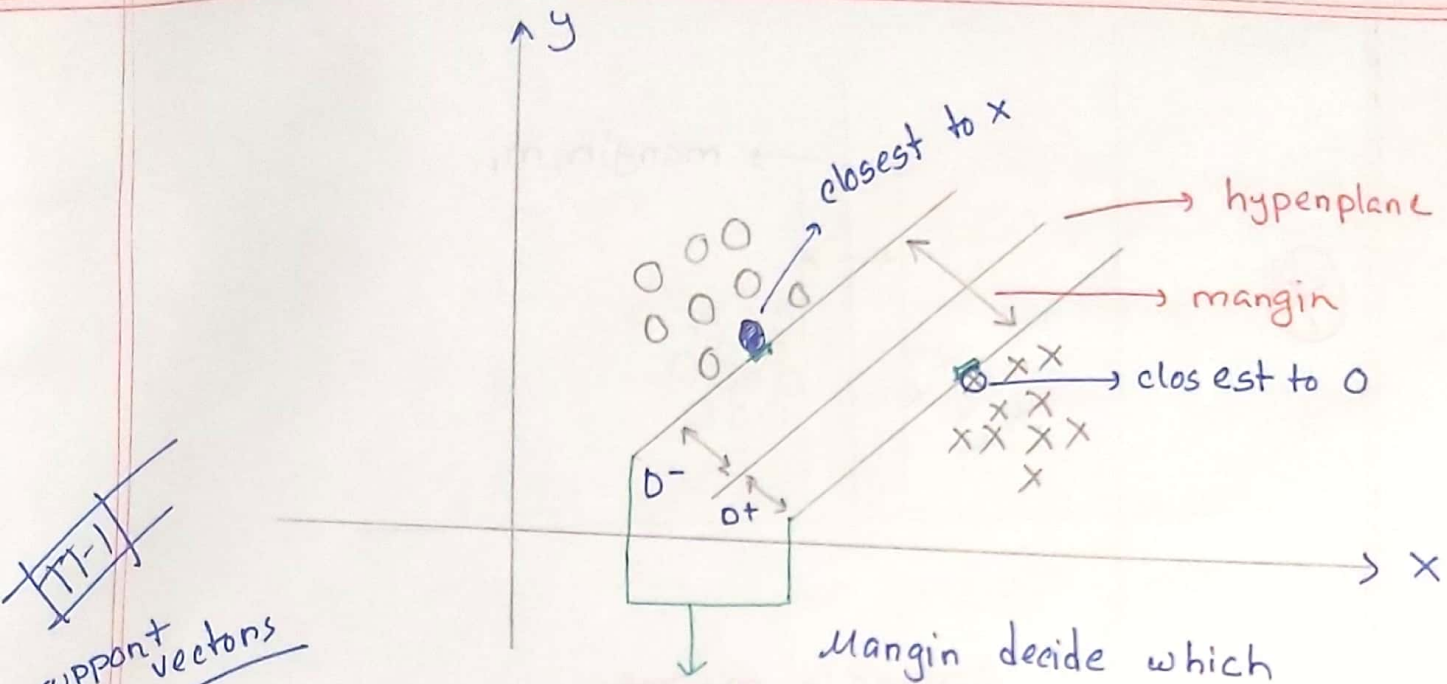
5	1	3	4
8	2	9	2
1	3	0	1
2	2	2	0

→

8	9
3	2

2x2 filter

Q. what are support Vectors?

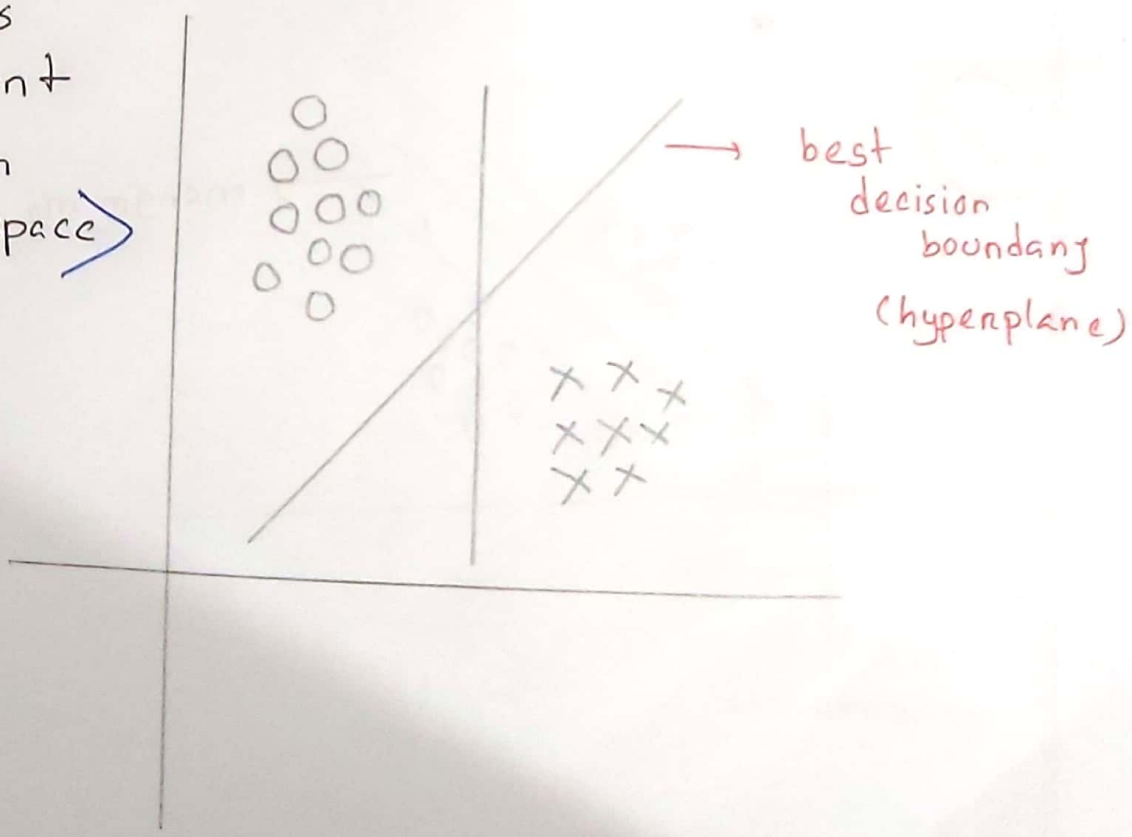


support vectors

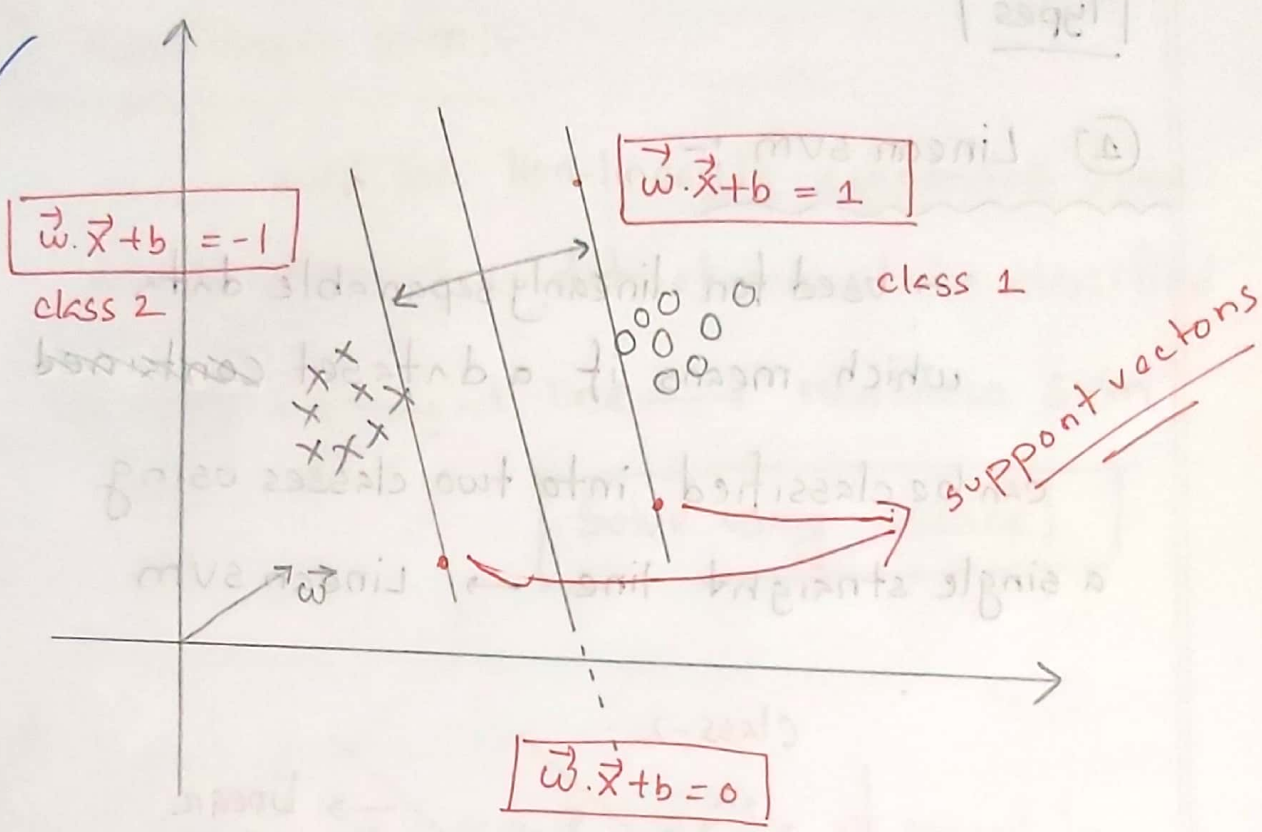
Mangin decide which hypERplane to choose?

Support vectors

SV are the data points that lie closest to the decision boundary (hyperplane) that seperates different classes in feature space



Q. How do support vector help to find the optimal mangin of model?



$(w \cdot x + b) \geq 1$; $f_{w,b} = 1$ $\forall x$ of class 1

$(w \cdot x + b) \leq -1$: $\forall x$ of class 2

$\therefore$ if multiple hyperplanx $= \frac{2}{\|\vec{w}\|}$ → hypenplan that has maximum value of this will choosen

(*) what is clustering? Pros and cons of Kmeans clustering?

Ans :- Clustering is a unsupervised machine learning technique used for grouping a set of objects or data into subset/cluster based on their similarities.

K means clustering is one of the most popolar and widely used clustering Algorithm where it partitions a dataset into K clusters with each cluster represented by its centroid.

The algorithm works iteratively by assigning datapoints to nearst centroid. and update centroids based on the mean of points assigned.

CS CamScanner

pros of k means

1. Simplicity
2. Fast convengence
3. scalibility
4. Interpretibility
5. Versatility

cons of kmeans

1. sensitivity to intial cluster
2. Fixed Number of cluster
3. struggle to identify Non gobular clusters
4. Assumption of circular / sperical cluster
   (always গোলাকার thinking)
5. Outliner affect centenoid position

CamScanner

	Height	weight
A	185	72
B	170	56
C	1568	60
D	179	68
E	182	72
F	188	77
G	180	71
H	180	70
I	183	84
J	180	88
K	180	67
L	177	76
M		

∴ Divide the data into 2 cluster

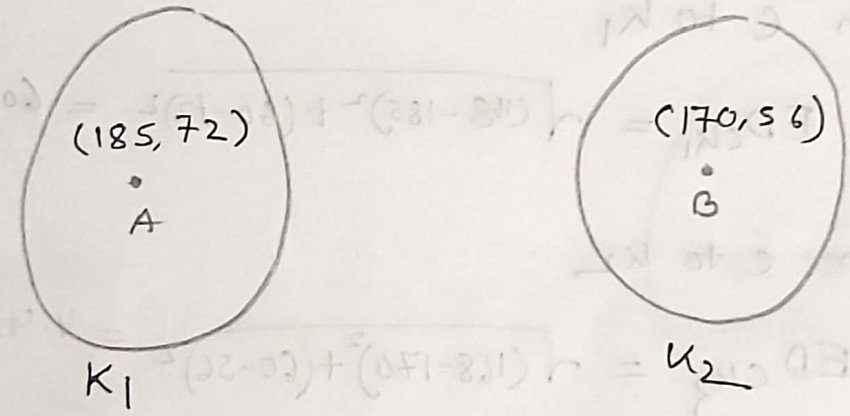
step-1 :- Given $k=2$

There will be 2 cluster in dataset

Assuming

A as a centroid of cluster 1 ($K1$)

B as a centroid of cluster 2 ($K2$)



Now based on this map other values to cluster.

Step-2 :- Take Variable C (next)

we need to which cluster for it.

For that Find Equclidean distance of C from each cluster

$\therefore$ Eq distance $= \sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$

$\therefore$ From C to K_1

$$ED_{CK_1} = \sqrt{(168-185)^2 + (60-72)^2} = 20.80$$

$\therefore$ From C to K_2

$$ED_{CK_2} = \sqrt{(168-170)^2 + (60-56)^2} = 4.48$$

$\therefore$ ED_{CK_2} is smaller. C belongs to K_2 clustering.

$K_2 = \{B, C\}$

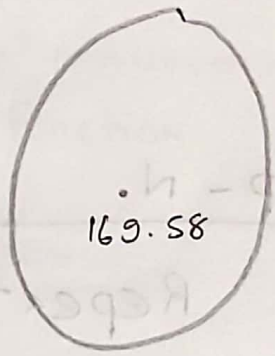
Step-3 :- Find the new centroid of K_2

$$K_2 = \left\{ \frac{x_1 + x_2 + \dots x_n}{n} , \frac{y_1 + y_2 + \dots y_n}{n} \right\}$$

$$= \frac{170 + 185}{2} , \frac{60 + 56}{2}$$

$$= 169, 58$$

$K_1 = \{A\}$



updated centroid

$K_2 = \{B, C\}$

step-4 NOW DO for D Variable

ED

$$K_1D = \sqrt{(178-185)^2+(68-72)^2} = 6.32$$

ED

$$K_2D = \sqrt{(178-169)^2+(68-58)^2} = 14.14$$

$\therefore$ D belongs to cluster K_1

$\therefore$ $K_1 = \{A, D\}$

step - N

Repeat (3.4) for all

$\boxed{\text{Final ans}}$

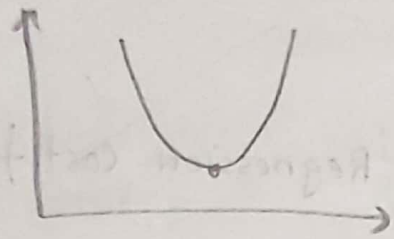
$K_1 = \{A, D, E, F, G, H, I, K, L, M\}$

$K_2 = \{B, C\}$

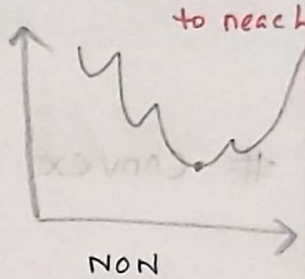
TT-1

(TT) (A) Does gradient descent require a convex cost function to coverage?

Yes



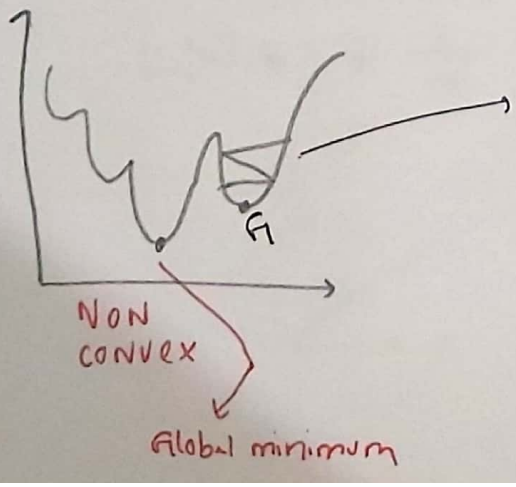
CONVEX



NON CONVEX

why convexity is important gradient descent

1. help to reach global minimum of cost function
2. shape of cost function
3. Learning rate or step size of gradient descent



Here gradient descent stuck at local minimum G and failed to reach global minimum due to NONCONVEX

NON convex Logistic Regnesseion cost function

$$J(\vec{w},b) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(f_{w,b}(x^{(i)}) - (y^{(i)}) \right)^2$$

Convex Logistic Regnession Cost function

$$J(\vec{w},b) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{L\left(f_{w,b}(\vec{x}^{(i)}), y^{(i)} \right)}_{\text{Loss function}}$$

CSE

$\boxed{\circledast \text{ what is loss function?}}$

loss function is a measurement of finding how well a given machine learning model fits into specific data set

Q-1

can we use mean squared Error for calculating gradient discent of Logisitic Regression to converge to global optima?

mean squared error (MSE) $= \frac{1}{2}(y-\hat{y})^2$

↓

if we create cost of function of Logistic Regression using MSE

$$J(\vec{w},b) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(f_{w,b}(x^{(i)}) - y^{(i)}\right)^2$$

which is a nonconvex

To converge to global optima we need a convex cost function

$$J(\vec{w},b) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{L\left(f_{w,b}(\vec{x}^{(i)}) - y^{(i)}\right)}_{\text{Loss function}}$$

$$= \begin{cases} -\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) \\ -\log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right) \end{cases}$$

TH

"Improper use Regularization leads to significant model failures"

① Over - Regularize

→ imposes heavy penalty on model's parameters

→ becomes too simple / Grid

→ results underfitting

→ overly generalized

→ perform poorly

② Under Regularize

→ overfitting

→ includes noise + fluctuations

→ perform overly well

→ fail to generalize unseen, new data

TH How does logistic regression differ from linear regression in term of nature of dependent variable and type of problem solves.

16 Final

Definition 16

Linear Regression	Logistic Regression
dependent variable is continous and Real value	dependent variable is binary or categorical
Represent a quantity that varies along a continous range	represents the probability of a particular category/class
predicting house, temperature, sales	predicting whether an email is spam or not
is suitable for problem where the goal is to predict a continous outcome	is suitable for problem where the goal is to classify or categorize (binary classification
$f(\vec{w}, b) = \vec{w} \cdot \vec{x} + b$	$f(\vec{w}, b) = \frac{1}{1+e^{-(\vec{w} \cdot \vec{x} + b)}}$

write definition using this

* Evalute the effectiveness of CNNs in image classifycation task?

1. Hienarchical feature Learning
   (filtens for detecting feature)

2. Parameter sharing
   (reduce total no parameter)

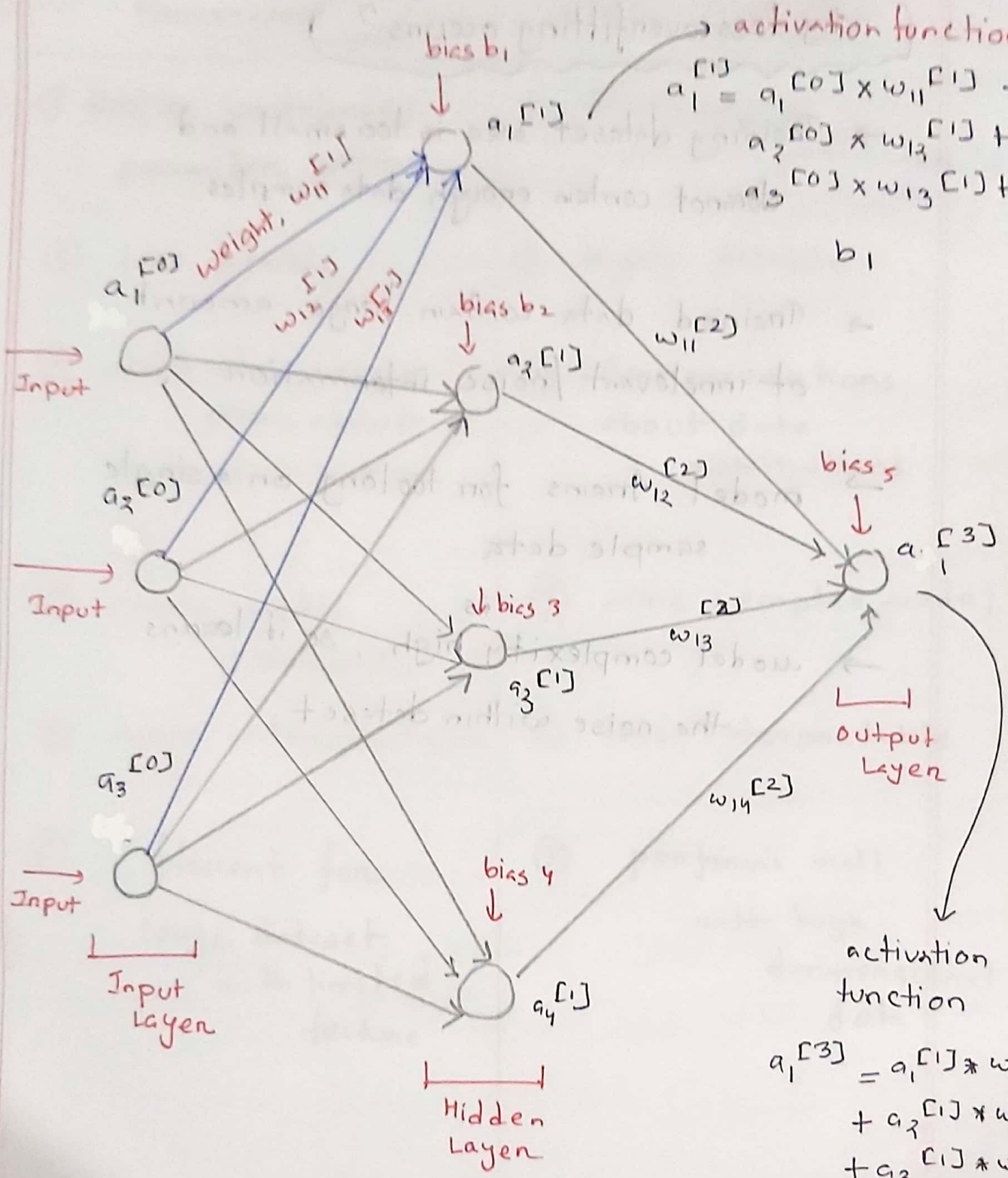
4. pooling Layers
   → downside dimmesion
   → reduce computation load

3. mapping using Activation function

* problem solves:- Image recognition,
  spech recognition,
  object dection
  natural languag processing

mathematical breakdown 17-1

forward propagation process → neural Network



why does overfitting occurs?

→ Training dataset size is too small and doesnot contain enough data samples

→ Trained data contain large amount of irrelevant / noise information

→ model trains for too long on a single sample data.

→ model complexity is high, so it learns the noise within dataset

CamScanner

Parametric model	non-parametric model
① Require predefinned parameter setting	① donot rely on specific parameter
② Less flexible	② Highly Flexible
③ make strong assumptions about data distribution	③ Few assumptions about data distributions
④ simple model structure	④ more complex model
⑤ more interpretable	⑤ less interpretable
⑥ Efficient for large dataset with limited texture	⑦ performs well with high dimensional data

CamScanner

Truth Label

Predicted labels	1	0
1	TP True positive	FP False positive
0	FN False Negative	True Negative

— o —

Predicted

Actual	Negative	positive
Negative	True Negative TN	False positive FP
positive	False Negative FN	True positive TP

Accuracy

Accuracy is metric that measure how often a machine learning model connectly predict outcome by diving the number of connect predictions by the number of total prediction

$$\text{Accuracy} = \frac{\text{connectly predicted Label}}{\text{Total Label}}$$

$$= \frac{TP + TN}{TP + TN + FP + FN}$$

where to use

→ inital model assesment

→ Balanced datasheets / datasets

→ when misclassification costs are equal

Precision

Precision is a metric that measures how often machine learning model correctly predicts the positive class

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

Recall is a metric that measures how often a machine learning model correctly identifies positive instance (true positive) from all the actual positive samples in dataset.

CamScanner

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score

F1 is harmonic mean of precision and Recall.

$$F_1\text{-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

CamScanner