TOPIC .... DATE ....

Reet Chauhan
022200300003001
M.Sc. DFIS

AI ASSIGNMENT

Q1 Compare and contrast with example supervised and unsupervised learning algorithms.

Ans Supervised and unsupervised learning are two approaches in Machine Learning.

SUPERVISED LEARNING :-

Supervised learning involves training a model using labelled data, i.e., with the input features and target/label values. The goal is to map a function to predict or classify outcomes for new inputs.

Some of the supervised learning algorithms are :-

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. Support Vector Machine
5. K-Nearest Neighbour
6. Naive-Bayes Classifier

These are some of the Models that can perform prediction and/or classification tasks.

Example of Supervised learning :-

Let's predict house prices based on area. Steps to implement linear regression on such data are :-

1. collect the data :-

TOPIC .... DATE ....

The data will be in a format like,

Area (in sq. ft.)	Price (in Lac)
5000	5.5
800	9
1000	9.8
1300	12.5
...	...

② Preprocess the data:-
During preprocessing, check for null values and handle them if present. [Categorical Datas are to be encoded] Then split the dataset into training and testing set.

③ Train the Model:-
Feed the training set to the linear Regression model. This Model works on the function of a line, i.e.,
$y = mx + b$,
m being Slope; b is ~~coeff~~ Intercept, x is input feature and y being the output.
Training the model with help in determining slope and Intercept. Once the slope and intercept is determined, the prediction for the test set is made.

④ Evaluate the Model:-
Various metrics like mean square error, precision, R2 score etc. can evaluate how good the model is

⑤ Prediction for unknown (new) data:-
Once evaluated, the model is ready to be used for real-time new data to make prediction.

TOPIC .... DATE..............................

like the implemented linear regression, There are various other supervised learning Models used for various different problems.

UNSUPERVISED LEARNING :-

Unsupervised learning involves training a model using unlabelled data, i.e., only input feature are available. The goal is to map a function to extract meaningful information from the data by clustering, reducing dimensions and/or finding patterns/association.

Some of the unsupervised learning algorithms are :-

1. K-Mean clustering
2. Hierarchial clustering
3. DBSCAN
4. Principle component Analysis
5. t-SNE
6. Gaussian Mixture Model

These are some of the models that can perform clustering and/or association tasks.

Example of Unsupervised learning :-

Let's make clusters of customers on the bases of purchases. Steps to implement K-Mean clustering on such data.

1. Collect the data :-

The data will be in format like,

Transaction_ID	Purchase_Amt	Frequency	Product_Ctgry
1	5000	5	Electronics
2	2000	3	Grocery
3	10000	2	Electronics
4	4500	10	Clothing
...	...	...	...

TOPIC .... DATE ....

② Preprocess the Data :-
Like the supervised learning, the pre-processing involves handling nulls, encoding categorical data, but not ~~and~~ spliting data into test and train as no label available.

② Train the ~~data~~ Model :-
To train a K-Mean Clustering model, first a K is to be defined. K can be pre-defined or can be defined by methods like elbow method. Then the K-Mean Clustering model is implemented on the model.

④ Evaluate the Model :-
The evaluation metrics of unsupervised learning are different from supervised learning. Silhouette Score, Calinski-Harabasz index etc can be used to evaluate the model.

⑤ Deploy the Model :-
The Model with good evaluation can then be deployed to ~~predict~~ work in real-time.

Just like the implemented K-Means Clustering, various other unsupervised learning models are used to solve various tasks.

TOPIC .... DATE..............................

Q2 Discuss Various Metrics used for classification and Regression Model with their Mathematical Relation.

Ans There are various evaluation Metrics used to evaluate the performance of a supervised Learning Model. These Metrics are different for classification and Regression problems.

Some metrics for classification are as follow:-

① Accuracy:- It measures the overall correctness of the Model.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

TP :- True positive , TN :- True Negative
FP :- False Positive , FN :- False Negative

② Precision :- It measures the proportion of true positive out of all positive predictions.

$$\text{Precision} = TP/(TP+FP)$$

③ Recall :- Also called Sensitivity. It measures the proportion of true positive predictions out of all actual positive instances.

$$\text{Recall} = TP/(TP+FN)$$

④ Specificity :- It measures the proportion of true negative prediction out of all actual negative.

$$\text{Specificity} = TN/(TN+FP)$$

TOPIC .... DATE.........................

(5) **F1 Score** :- It is harmonic Mean of Precision and Recall, providing balanced measure of a model.

$$F1\ Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)}$$

(6) **Confusion Matrix** :- It is a tabular representation of the actual and predicted values.

		Predicted Positive	Predicted Negative
Actual	Positive	TP	FN
	Negative	FP	TN

These are few metrics to evaluate a classification model like logistic Regression, Naive-Bayes classifier etc.

Some metrics for regression are as follow:-

(1) **Mean Square Error** :- It calculates the average of the squared difference between the predicted and actual values.

$$MSE = \frac{1}{n} \sum (y_i - \bar{y_i})^2$$

where, n is no. of data points, y_i is the actual value and $\bar{y_i}$ is the predicted value.

(2) **Mean Absolute Error** :- It calculates the avg. of the absolute difference between the predicted and actual values.

TOPIC .... DATE..............................

$$MAE = \frac{1}{n} \sum |y_i - \bar{y}_i|$$

(3) Root Mean Square Error :- It is the square root of the Mean Square Error.

$$RMSE = \sqrt{MSE}$$

(4) R2 Score :- Also called R squared. It measures the proportion of the variance in the dependent variable from the independent variable.

$$R2 = 1 - (SSR/SST)$$

where SSR is sum of squared residues and SST is sum of squares (total).

(5) Mean Square Logarithmic Error :- It calculates the avg. of the logarithmic difference between the actual values and the predicted values.

$$MSLE = \frac{1}{n} \sum \left(\log(1+y_i) - \log(1+\bar{y}_i)\right)^2$$

These are few metrics to evaluate a prediction model like linear regression.

TOPIC .... DATE....................

Q3 Explain in detail various types of image segmentation and object detection.

Ans IMAGE SEGMENTATION :-

The task of dividing an image into multiple segments based on certain characters or features.
There are mainly 3 type of Image segmentation:-

(1) Semantic Segmentation:-
Semantic Segmentation involves assigning a semantic label to each pixels in an image, grouping similar pixels together based on their visual properties.
In Semantic Segmentation, the output is a pixel-level map where each pixel is assigned a label including the class it belongs to.
Example:-
An image with Multiple trees and vehicles will provide mask that categorize all trees into one class of tree and all vehicles into one vehicle class. irrespective of what vehicle it is, like car, cycle, bus etc.

(2) Instance Segmentation:-
Instance Segmentation not only label each pixel but also distuinguish between individual instances of object.
In Instance Segmentation, the output is a pixel-level map where each pixel is assigned a ~~label~~

TOPIC .... DATE............

a label indicating the object category it belongs to and a unique idenifier for the instance it corresponds to.

Example :-
An image with multiple people will locate Individual object within the crowd, identifying no. of instances in the picture, but cannot predict the region or object for each instance.

(3) Panoptic Segmentation :-
Panoptic Segmentation is a combination of both semantic and instance segmentation. In Panoptic Segmentation, the output is a pixel-level map where each pixel is assigned a semantic label and an instance ID (or "ignore" label).

Example :-
An image with people and vehicles will provide mask that categorize each individual person and each vehicle with separate instances IDs.

There are several different segmentation techniques :-

1. Threshold-based segmentation
2. Edge-based image segmentation
3. Region-based image segmentation
4. Clustering-based image segmentation
5. Artificial Neural Network-based image segmentation.

TOPIC .... DATE..............................

OBJECT DETECTION :-

The task involves identifing and locating objects within digital image and video frames.
The techniques to implement object Detection are as follows:-

① R-CNN :-

Region-based Convolutional Neural Network.
It is one of the pioneering Object detection method. It segments an input image into regions extract feature using CNN and classify using SVM.

② SPP-net :-

Spatial Pyramid Pooling Networks.
It improves upon R-CNN by introducing SPP. This allows the network to accept images of variable sizes and produce fixed length feature vectors.

③ Fast R-CNN :-

It builds upon R-CNN and address its limitations by sharing convolutional features across region proposal. It performs feature extraction for the entire image and then use RoI to align feature with the region proposals.

④ Faster R-CNN :-

It further improves the speed and accuracy by introducing the Region Proposal Network. This end-to-end network eliminates the need for external Region proposal algorithms.

TOPIC .... DATE..................................

⑤ YOLO :-

You only look Once Detector.
It is a one-stage object detection algorithm that divides an input image into grid and predicts bounding boxes and class probability for each grid cell.

⑥ SSD :-

Single Shot Multi-Box Detector.
It is another One-stage object detection algorithm that predict object bounding boxes and class probabilities at multiple scales and feature maps. It achieves a good balance between accuracy and speed.

Q4. Explain with suitable example, how face detection work in real time applications.

Ans Face/Facial Detection is an Artificial Intelligence based computer technology used to find human faces in digital images and videos.
~~It uses ML and ANN technology to do so.~~
Face detection application uses a combination of AI algorithms, Machine learning, Statistical Analysis and Image processing to find human faces within image and distinguish them ~~for~~ from non-face objects and other human body parts.

TOPIC .... DATE..................................

To achieve real-time face detection:-

① Image Acquisition:-

Acquiringthe digital media that is to be ~~or~~ analyzed.

② Preprocessing:-

Resizing, reducing noise, adjusting contrast etc. is done to enhance the quality of image for better results.

③ Feature Extraction:-

Extracting distinctive features like edges, lines, textures etc ~~of~~ that are characteristics of human face. To do so, various techniques such as Haar Cascades, CNN etc can be used.

④ Classification:-

The extracted features are then classified whether that region of the image contain a face or not. Classifier is trained on large set of data of labelled image. SVMs, Neural Networks etc are employed for classification

⑤ Post-Processing:-

Post processing is done to improve the accuracy of the model. Size filtering, confidence thresholding, Bounding box adjustment etc. are done to eliminate false positive results.

TOPIC .... DATE..............................

⑥ Output :-
The output is typically a set of bounding boxes that indicate position and size of face detected.

⑦ Caching and Optimization :-
To avoid redundant computation, face detection systems may cache intermediate results. The optimization reduce computation time.

Despite of the algorithms and techniques used by real-time face detection systems, the basic flow of proceeding remains the same as the one mentioned above.

Hardware acceleration, Multi-Scale processing, ROI filtering, parallel processing etc are some additional techniques ~~[illegible]~~ employed by various face detection systems to enhance their performance in real-time implementation.

The results of face detection are further used for tasks including face recognition, emotion detection etc.

TOPIC .... DATE..................

Q5 Discuss with example text vectorization and various methods of text matching.

Ans TEXT VECTORIZATION:-

It is a process of converting text data into numerical vectors that ML Algorithm can understand and process.
Some popular Text Vectorization methods are:-

① Binary Term frequency:-
The presence or absence of a term in a document is represented by 0 or 1. This technique ignores the frequency of terms.
For Example, The following two documents:
DOC 1: "The cat is black"
DOC 2: "The dog is brown"
Binary Term ~~ted~~ frequency vectorization of the Documents is:
DOC1 :[1, 1, 1, 1, 0, 0]
DOC2 :[1, 0, 1, 0, 1, 1]
Here, the vector elements represent present (1) or absence (0) of the terms "The", "Cat", "is", "black", "dog" and "brown" respectively

② Bag of word (BoW) Term frequency:-
The presence and frequency of terms in a document is considered in BOW term frequency.
For Example, for the above 2 documents:

TOPIC .... DATE..............................

BoW vectorization is:
Doc1: [1, 1, 1, 1, 0, 0]
Doc2: [1, 0, 1, 0, 1, 1]
Here, the vector elements represent the frequency of each terms as "The", "cat", "is", "black", "dog" and "brown" respectively.

③ (L1) Normalized Term Frequency :-
Scales the term frequency in a document resulting in a vector with values ranging between 0 and 1.
For example, ~~(L1)~~ for the above 2 document, (L1) Normalized Term frequency vectorization is:
Doc1: [0.25, 0.25, 0.25, 0.25, 0, 0]
Doc2: [0.25, 0, 0.25, 0, 0.25, 0.25]
Here, the vector elements represent the frequencies of each term.

④ (L2) Normalized TF-IDF :-
Term Frequency - Inverse Document Frequency combines the Term Frequency with the inverse document Frequency, resulting in a vector with values ranging between 0 and 1.
For example, for the above 2 documents, (L2) Normalized TF-IDF vectorization is:
Doc1: [0.267, 0.267, 0.267, 0.267, 0, 0]
Doc2: [0.267, 0, 0.267, 0, 0.267, 0.267]
Here, Vector elements represent frequencies.

TOPIC .... DATE..............................

(5) Word2Vec:-

It is an advanced technique that represents words in a vector space based on their semantic meaning. Word2Vec represents words as dense, low-dimensional vectors, capturing their relationships and similarities
For Example,
It can learn that "King" is to "queen" as "man" is to "woman".

TEXT MATCHING:-

It is the task of determining similarity or relatedness between two pieces of text.
It involves comparing textual data and identifying patterns, relations etc. to determine if two texts are similar.
Methods of Text Matching are:-

(1) Exact Matching:-
comparing exact sequence of characters or words in two texts.
Example, Searching for a word like "Hello" in collection of articles.

(2) Vector-Space Models:-
Text Vectorization using TF-IDF, Word2Vec, ~~etc~~ BoW etc.
Example; vectorization of "Hello world" and "Hello everyone" is [1,1,0] and [1,0,1] respectively.

Title ______ Date ______
Page ______

(3) Semantic Similarity :-
Aims to capture the underlying meaning of text and measure similarity.
Example, In sentences,
"A car crashed into a tree" and
"An automobil hit a tree".
A Semantic Model can understand that "car" and "automobile" have similar meaning and "hit" and "crashed" indicate similar event.

(4) Phonetic Matching :-
Find Similarities based on the pronunciation rather than spelling or meaning.
Example, Phonetic Matching model will match "Smith" with "Smyth" and "Smithe" as they have similar pronunciations.

(5) Cosine Similarity :-
The text is vectorized and the cosine of the angle between them is calculated to determine the similarity.
Example, in the sentences vectorized earlier using BOW as [1,1,1,1,0,0] and [1,0,1,0,1,1] can be computed using the formula
$\cos\theta = (A \cdot B) \div (\|B\| \times \|A\|)$
Here,
$A \cdot B = 1\times1 + 1\times0 + 1\times1 + 1\times0 + 0\times1 + 0\times1 = 2$
$\|A\| = \sqrt{1^2+1^2+1^2+1^2+0^2+0^2} = 2$
$\|B\| = \sqrt{1^2+0^2+1^2+0^2+1^2+1^2} = 2$
$\cos\theta = 2 \div (2\times2) = 1 \div 2 = 0.5$

CLASSTIME

Title ____ Date ____
Page ____

Q6 Explain how classification Method work for text data by citing relevant example.

Ans Sample Dataset be "Review.csv"
"I love the Movie.", "Positive"
"The film wasn't good.", "Negative"
"The direction was amazing.", "Positive"
"The movie was engaging.", "Positive"
"The dialogues were cliche.", "Negative"

For the Sample Data, the text classification will be as follow:-

① Gather the data:-
First Step is to collect the data, i.e., load the "Review.csv" file.

② Preprocess the data:-
The Preprocessing will include:-
→ converting all to lower case.
→ Removing special characters.
→ Handling contractions like "wasn't" to "was not."
→ Removing stop words like "is", "was", "the"
And then splitting the dataset into test and train sets.

③ Feature Extraction:-
First we extract feature and target, label the target ("positive" as 1 and "negative" as 0) and vectorize the text using BOW. To do so, we:

CLASSTIME

Title
Date
Page

→ Create a vocabulary by tokenizing the text into individual words.
→ Assign index to each word in vocabulary.
→ Then vectorize each element

(4) Train the Model:-
Implement a classification model like Naive-Bayes algorithm for text classification.
→ Calculate prior probability of each sentiment class (0 or 1) based on train set.
→ Calculate likelihood for each word in vocabulary given each sentiment class.
→ Combine prior probability and likelihood to estimate posterior probability
→ Train the model by fitting calculated probabilities.

formula:- $P(C|X) = P(C) \times P(X|C)$

where $P(C|X)$ → Posterior Probability
$P(C)$ → Prior Probability
$P(X|C)$ → likelihood.

(5) Evaluate the Model:-
Predict the sentiment label for the test set using the trained Model and evaluate the Model using metrics like Accuracy, precision, recall, F1 score etc to check model performance

(6) Prediction:-
The model can now be used to predicted label for new, unseen data (movie reviews)'

CLASSTIME

Title ______ Date ______
Page ______

These steps can be followed for any text classification Dataset.
This example covered Naive-Bayes Algorithm for the classification but there are various other models that can be used such as logistic Regression, SVM, or deep learning neural Networks like RNN.

Q7 Discuss Over-fitting and under-fitting with example.

Ans OVERFITTING :-

Overfitting is a situation that occurs when a model tries to cover all the data points which may seem to make good predictions in the start but then starts caching noise and inaccurate values.
An Overfitting model has low bias and high variance.

Reasons of Overfitting :-
1. Giving ~~larger~~ smaller train set and giving more training to the model.
2. Defining a complex Model Architecture.
3. Lack of Pre-processing the noisy data.
4. Lack of Model Evaluation.
5. Lack of Regularization.

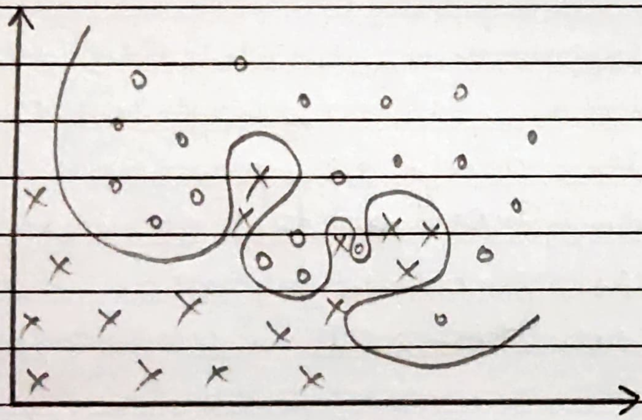
CLASSTIME

Title Date
Page

Techniques to Reduce Overfitting :-

1. Provide a larger, relevant train set.
2. Simplify the model architecture by reducing layers, nodes or parameters.
3. Remove outliers and smooth the noisy data points by data cleaning and pre-processing.
4. Monitor the model's performance.
5. Apply Regularization methods like L1 or L2 regularization.

Example of Overfitting :-



The curve in the graph represent an overfitted model which seems to appear as "too good to be true" and gives a complex decision boundry.

UNDERFITTING :-

Underfitting is a situation that occurs when a model is not able to capture the underlying trend of the data, i.e., it perform well on train set but not the test set.
An underfitting model has low variance and high bias.

CLASSTIME

Title ______ Date ______ Page ______

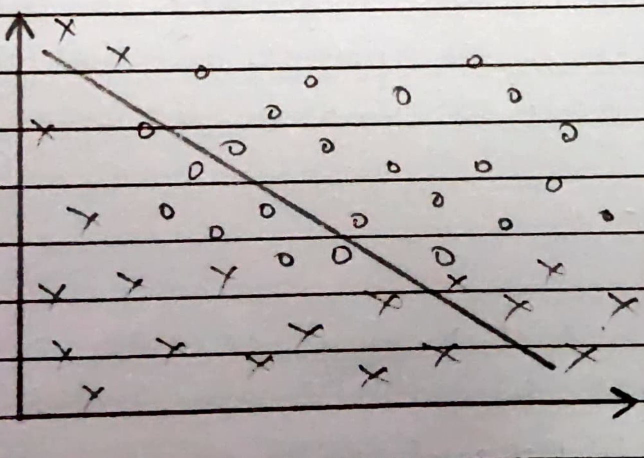
Reasons of Underfitting :-

1. Low model complexity.
2. Model trained for insufficient number of iteration/time.
3. Data Scaling and Normalization of Sensitive data.
4. Data Imbalance, i.e., one class has more samples than the other class.
5. Model Regularization Oversight

Techniques to prevent Underfitting:-

1. Add a few more layer or nodes to a simple model
2. Increase the number of epochs to allow the model to learn the patterns and adjust.
3. Apply appropriate scaling technique.
4. Oversampling the minority class or undersampling the majority class to bring balance.
5. Regularization can also help in avoiding underfitting

Example of underfitting:-



The graph is too simple to handle the data and hence the decisionboundry lack accuracy.

CLASSTIME

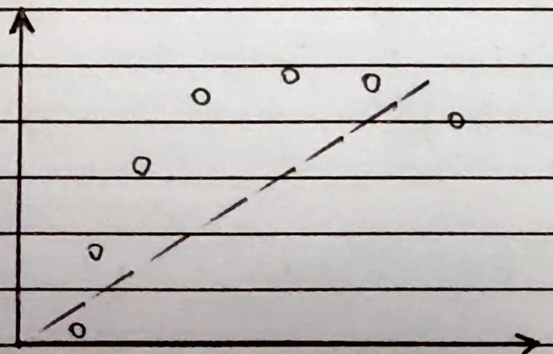
Title______________________ Date______
Page______

Q7 Explain in detail bias-variance trade-off with suitable example.

Ans Bias and Variance are the prediction errors and the balance between the bias and the variance error is known as bias-variance trade-off.

BIAS :-

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. The high difference between the predicted values made by the model and actual value/expected value signifies high bias error and thus implies that less attention was payed while training and oversimplifies the model. High bias gives an underfitting model. Example,



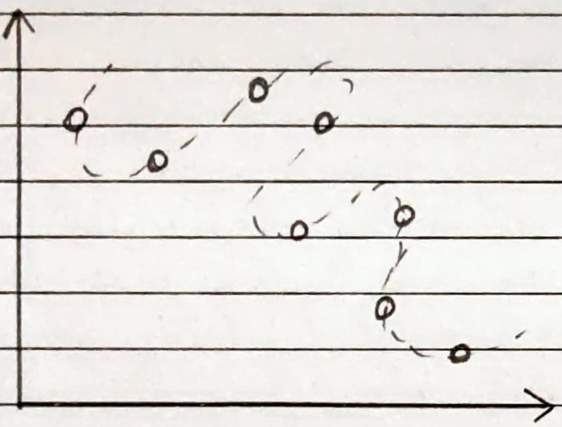
The graph shows that the model lacks accuracy and hence predict values with high error.

CLASSTIME

Title
Date
Page

VARIANCE :-

Variance is the variability of model prediction for a given data point or value which tells us spread of our data. When a model varies too much, ~~a~~ it signifies ~~to~~ that the model has high variance that further implies that data isn't trained enough and the model is too complex. High variance gives an overfitting model. Example,



The graph shows that the model isn't provided enough data to train and hence is giving a way complex model with lot of variance.

COMBINATION OF BIAS-VARIANCE :-

1. Low Bias, Low Variance :- Shows an ideal Machine learning model.

2. Low Bias, High Variance :- predictions are inconsistent and accuracy is average. Leads to Overfitting.

CLASSTIME

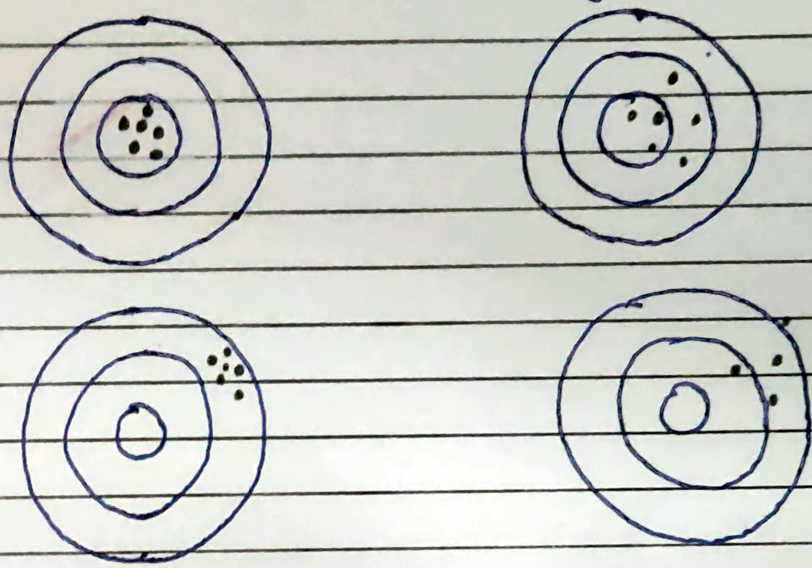
③ **High bias, Low Variance:-** predictions are consistent, and accuracy is low. Leads to underfitting.

④ **High bias, High Variance:-** predictions are inconsistent, and accuracy is low.

Low Variance | ~~Low~~ High Variance

Low Bias

High Bias



BIAS-VARIANCE TRADE-OFF:-

The Balance between Bias and Variance is bias-variance trade-off. This helps in selecting values of Regularization constant. It helps in ~~be~~ building a model with good balance by minimizing total error.

$$TE = (Bias)^2 + Variance + Irreducible\ Error$$