Winter Session: January 21, 2025 Day 2

1. Explanation of Clustering

Clustering is a type of unsupervised machine learning where the goal is to group a set of datapoints into subsets; called cluster; such that datapoints within the same cluster are more similar to each other than to those in other clusters. Clustering is often used when there are no predefined labels or categories in the dataset; and the goal is to discover hidden patterns or structure in the data.

2. Representation of Clustering (Plotting)

"make_blobs" is a function that is able to generate synthetic data; and it returns two variables X and y. X contains the feature coordinates: a 2D array (matrix) that holds the coordinates of the generated points.

Y contains the Cluster labels. For example; if the function creates three clusters; y might look [0, 1, 2 ---]

3. Explanation of K-means algorithms

The steps that you need to follow to implement K-means algorithms are below:

Step 0: choose the initial center of the clusters randomly

Step 1: Assign each datapoint to a cluster based on which cluster center is closest

Step 2: Determine new cluster-centers by calculating the mean position of all the datapoints assigned to the cluster

Step 3: Compute the distance between the old and the new cluster centers

Repeat Steps 1-3 until the difference between the old and the new centroid is sufficiently small

4. Explanation of Classifier

A "Classifier" is a machine learning or algorithm used to categorize data into predefined classes or groups. It maps input data (features) to a specific category or labels based on patterns it has learned during training

5. Evaluation of model prediction

Evaluating a model's predictions involves assessing how well the model performs on a dataset and how accurately it predicts the desired outcomes.

Key Metrics for model evaluation are the follo

1. Accuracy

Definition: The percentage of correctly predicted instances out of the total instances

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

It works well when the dataset is balanced (equal distributions of classes)

2. Precision

The percentage of true positive predictions among all positive predictions.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

3. Recall (Sensitivity)

The percentage of true positive predictions among all positive instances

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

4 F1_Score

The harmonic mean of precision and recall ; balancing the two metrics:

$$\text{F1\_Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Confusion Matrix

A matrix summarizing the performance of a classification model.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual negative	False Positive (FP)	True Negative (TN)

The Confusion matrix helps visualize the distribution of predictions and identify errors.

6. ROC and AUC

ROC Curve: Plots the True Positive Rate against False Positive rate

AUC Curve: Measures the ability to distinguish between classes. Higher AUC indicates better performance

7. Cross- Validation:

1. Split the Data: divide your dataset into training; validation; and test sets

2. Train the model : Fit the model to the training data

3. Make predictions: predicted class labels as well as predicted probabilities

Remark: It is important to make that the method that we have just mention are one only of the evaluation goal. These evaluation helps us to provide a point estimate of model's performance

8. What are the other goals for the evaluations of model's performance ?

- To provide information about the variance around that point estimate
- To find the best parameters for a given model (hyperparameters)
- To determine which of several models (Logistic Regression ; Random forests ; Neural Network).

9. Why High Variance indicates overfitting or Sensitivity to training data ?

High variance in ML model typically means that the model performs well on the training data but struggles to generalize to new ; unseen data.

Remark : "Hyperparameters" are settings or configurations that are set before training the model and cannot be learned from data.
They control aspects of the model such as its complexity ;
how it learns or how predictions are made.

10. What is the Classification metrics ?

"Classification metrics" are used to evaluate the performance of a classification model by comparing its predictions against the actual labels. These metrics helps us understand how well the model is performing in terms of making correct predictions.

Type of tests that can be conducted :

- Accuracy
- Precision

- Recall
- F1 - score

11. What is the meaning of Regression metrics ?

Regression metrics are used to evaluate the performance of a regression model, which predicts a continuous numerical value rather than a categorical label. These metrics quantify how well the model's predictions match the actual target value.

The common regression metrics are:

1. Mean Absolute Error
2. Mean Square Error
3. Root Mean Square Error
4. R-squared (R^2)
5. Adjust R^2

Remark: In machine learning; we train a model on one dataset with the goal of making accurate predictions in another dataset. In other words; we aim to create a model that has good generalization. Our evaluation metrics should ideally give us a good sense of how well our model to perform on data it has never been seen before.

12. Is it good idea to train all your dataset?

By fitting on model on all our data and evaluating performance on the same dataset; we have created a model that captures our training data perfectly; but is unlikely to generalize to new data (i.e: we have overfit to the training data).

13. Is it good idea to split the data in Test / Train data?

yes; it is a good idea to split the data into test / train data. you train your model on one subset and evaluate the model on the other. The model's performance on the test set is referred to as the testing accuracy because we are evaluating the model on an independent dataset that was not used during model training

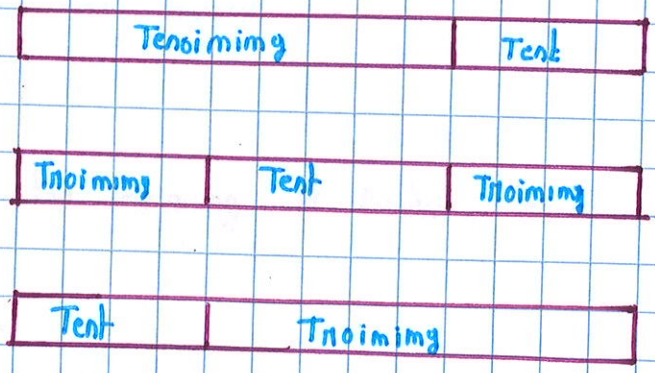
For this reason; testing accuracy is usually a better estimate of out of sample performance than training accuracy.

14. What is K-fold cross validation ?

Here are the steps for K-fold cross validation:

1. split the dataset into K equal partitions
2. use fold 1 as the testing set and the union of the other folds as the training set
3. Calculate testing accuracy
4. Repeat steps 2 and 3 K times ; using a different fold as the testing set each time
5. Using the average testing accuracy as the estimate of out of sample of accuracy

Remark: Each fold acts as validation only once ; and the model is retrained from scratch for each fold.



15. What in the Cross Validation for Parameter Tuning ?

"Cross Validation for parameter tuning" is a method used to systematically find the best hyperparameters for a machine learning model.

Poor hyperparameter can lead to underfitting or overfitting

Good hyperparameter improves the model's performance and generalizability

Selection of hyperparameter is done by trial and error. There is no method that you can use it to come up with the optimal value of hydroparameters.

16. What is the Cross Validation for Model Selection ?

"Cross Validation for model selection" is a method used to compare multiple machine learning models and identify the one that performs best on a given dataset.

17. What is the meaning of Nested - Cross validation in two loops ?

The main problem is that when you are tuning hyperparameters, you often evaluate the model using the same validation data you used

for tuning. This can lead to data leakage; overfitting to the validation data; and an overly optimistic estimate of the model's performance.

Nested Cross Validation is about two loops:

Outer loop: Tests the model's performance on unseen data

Inner loop: Tunes the hyperparameters for the model

18. What is logistic regression models?

Logistic Regression is a statistical method used in machine learning to model the probability of a binary outcome (eg success / failure; yes / No, 1/0). based on one or more independent variables (features).

Despite its name; logistic regression is used for classification problems, not regression.

Logistic regression starts by computing a linear combination of the input features.

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

where :

w : Weights

x : inputs

b : bias (intercept)

19. What is the meaning of model tuning ?

Model tuning refers to the process of optimizing a machine learning model's hyperparameters to achieve the best performance on a given dataset. Hyperparameters are parameters that are not learned during training but are set before the training process.

Remark: Model deployment is the process of taking machine learning model that has been trained ; tested ; and validated and making it available for use in a production environment where it can interact with real world data and provide predictions.