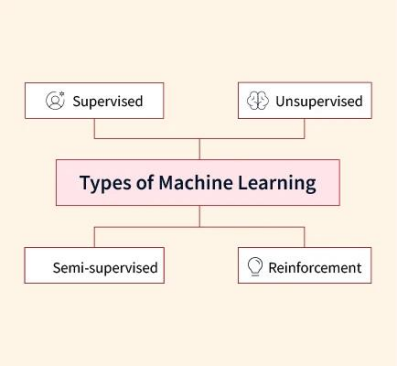
UNIT 2
Short Answer Questions

1. What is machine learning? Write its types.

A. Machine learning (ML) is a subset of artificial intelligence (AI) that allows computers to learn and improve from data without being explicitly programmed. ML uses algorithms to analyze data, identify patterns, and make predictions. The more data a machine learning model is exposed to, the better it performs.



2. Write down any five applications of supervised machine learning.

A. Applications of supervised machine learning are:

- Face recognition
- Fraud detection
- Visual recognition
- Risk assessment
- Spam detection
- Recommendation systems

3. Write down any five applications of unsupervised machine learning.

A. Applications of unsupervised learning are:

- Market basket analysis
- Semantic clustering
- Delivery store optimization
- Identify accident prone areas
- Text recognition
- Object recognition

4. How do you say a model is good or bad? Write its characteristics.

A. A good model should have strong predictive power and generalize well to unseen data.

Good Model Characteristics:

- Reliable
- Accurate
- Predictive
- Adaptable

5. What is accuracy? How to calculate accuracy?

Accuracy in machine learning is a metric that measures how well a model predicts outcomes. It's calculated by dividing the number of correct predictions by the total number of predictions. Accuracy can be calculated for classification models.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

We calculate accuracy from confusion matrix as

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

6. What is MSE? How to calculate MSE?

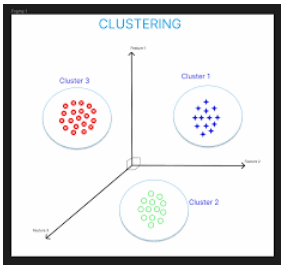
A. MSE means Mean Square Error. MSE measures the average squared difference between predicted and actual values. We calculate MSE for regression models.

The formula for calculating MSE is

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

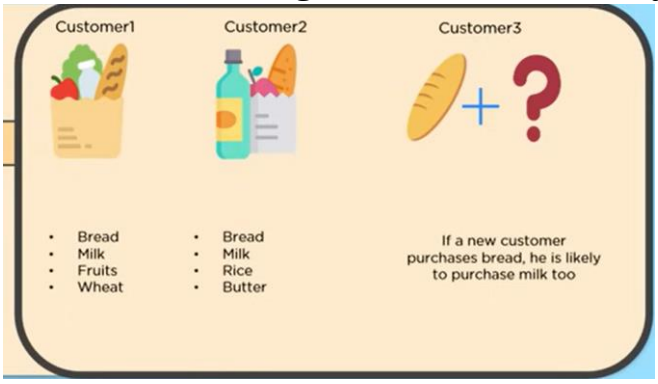
7. What is clustering?

A. Clustering in machine learning is a process that groups similar data points into clusters. It's used to organize data so that similar items are in the same group, and dissimilar items are in different groups.



8. What is association?

A. Association in machine learning is used in discovering interesting relations between variables in large databases i.e identifying patterns.



9. What are the problems faced when handling large data?

A. The problems faced when handling large data are:

- **Memory Limitations:** When data exceeds RAM, the OS swaps memory to disk, slowing performance; most algorithms aren't designed for large data, leading to out-of-memory errors.
- **Time Constraints:** Some algorithms run indefinitely or take unreasonably long to process even small data sets.
- **Bottlenecks:** Certain components of the computer may be underutilized due to deadlock like situation
- **Storage Speed Issues:** HDDs are slow, causing delays in data processing; SSDs help but are costly.

10. What are the different packages used for working with data in memory?

A. The different packages/tools used for working with large data is as follows:

- **Cython**- converts python code into C-like code, which runs faster.
- **Numexpr** – uses just in time compiler to run numerical expressions(NUMPY) efficiently.
- **Numba** – gives high performance on numerical and scientific computations.
- **Bcolz** – stores large numerical data in compressed format to save memory.
- **Blaze** – provides a consistent way to access different databases and data sources. Concerts python commands into sql queries or database friendly operations.
- **Theano**- compiles complex mathematical functions easily.
- **Dask** – enables parallel and distributed computing for large datasets by dividing data into chunks.