

Some fundamental questions up to ML (CS-3035) mid-semester exam.
(but not limited to):

<https://www.javatpoint.com/clustering-in-machine-learning>

1. What is learning in computer science/machine learning?

'Learning' in machine learning refers to the determination of model parameters using the given training dataset.

2. Differentiate between weak-AI and Strong-AI.

	Strong artificial intelligence	Weak artificial intelligence
Definition	<ul style="list-style-type: none">the form of artificial intelligence, which has the same intellectual abilities as human, or even surpasses him in it	<ul style="list-style-type: none">Weak AI is generally developed or used for specific application domains.In a standard work on artificial intelligence, this is formulated as follows: "The assertion that machines could possibly act intelligently (called, weakness, act as if they are intelligent) is called the, weak AI 'hypothesis ..."
Capabilities and Domains	<ul style="list-style-type: none">Logical thinkingMaking decisions in case of uncertaintyTo planTo learnCommunication in natural languageUse all these abilities to achieve a common goal	<ul style="list-style-type: none">Expert systemsNavigation systemsVoice recognitionCharacter recognitionSuggestions for corrections in searches

3. Differentiate between Machine Learning and Deep Learning.

S.NO	Deep Learning	Machine Learning
1.	To be qualified for deep learning, there has to be at least three layers	Can be defined as a shallow neural network which consists one input and one output, with barely one hidden layer
2.	Requires large amount of unlabelled training data	Requires small amount of data
3.	Performs automatic feature extraction without the need for human intervention	Cannot perform automatic feature extraction, requires labelled parameters
4.	High-performance hardware is required	High-performance hardware is not required
5.	Can create new features	Needs accurately identified features by human intervention
6.	Offers end-to-end problem solution	Tasks are divided into small portions and then forms a combined effect
7.	Takes a lot of time to train	Takes less time to train

4. Explain different types of learning using suitable real-world examples.

Learning Problems

- 1. Supervised Learning

An example of supervised learning is text classification problems. In this set of problems, the goal is to predict the class label of a given piece of text. One particularly popular topic in text classification is to predict the sentiment of a piece of text, like a tweet or a product review.

- 2. Unsupervised Learning

Finding customer segments Clustering is an unsupervised technique where the goal is to find natural groups or clusters in a feature space and interpret the input data. There are many different clustering algorithms. One common approach is to divide the data points in a way that each data point falls into a group that is similar to other data points in the same group based on a predefined similarity or distance metric in the feature space. Clustering is commonly used for determining customer segments in marketing data. Being able to determine different segments of customers helps marketing teams approach these customer segments in unique ways. (Think of features like gender, location, age, education, income bracket, and so on.)

<https://www.springboard.com/blog/lp-machine-learning-unsupervised-learning-supervised-learning/>

- 3. Reinforcement Learning

The example of reinforcement learning is your cat is an agent that is exposed to the environment. The biggest characteristic of this method is that there is no supervisor, only a real number or reward signal.

✕	Supervised Learning	Unsupervised Learning	Reinforcement Learning >
Definition	The machine learns by using labelled data	The machine is trained on unlabelled data without any guidance	An agent interacts with its environment by producing actions & discovers errors or rewards
Type of problems	Regression & Classification	Association & Clustering	Reward based
Type of data	Labelled data	Unlabelled data	No pre-defined data
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output	Follow trial and error method
Popular algorithms	Linear regression, Logistic regression, Support Vector Machine, KNN, etc	K-means, C-means, etc	Q-Learning, SARSA, etc

5. Differentiate between supervised learning and unsupervised learning.

Unsupervised Learning	Supervised Learning
The machine is given huge sets of data that are not labelled as inputs to analyse.	The input is in the form of raw data that is labelled.
The machine needs to figure out the output on its own by identifying patterns in the raw data provided to it.	The machine is already fed with the required feature set to classify between inputs (hence the term 'supervised').
Divided into two types of problems – Association (where we want to find a set of rules that describe our data) and Clustering (where we want to find groups in our data).	Divided into two types of problems – Regression (outputs are real values) and Classification (outputs are categories).
K-means for clustering problems and Apriori algorithm for association rule learning problems.	Linear regression for regression problems, Random Forest for classification and regression problems, Support Vector Machines for classification problems

Three types of analysis

- Univariate analysis
 - the examination of the distribution of cases on only one variable at a time (e.g., college graduation)
- Bivariate analysis
 - the examination of two variables simultaneously (e.g., the relation between gender and college graduation)
- Multivariate analysis
 - the examination of more than two variables simultaneously (e.g., the relationship between gender, race, and college graduation)

Univariate Data

- Involving **a single variable**.
- Does **not deal** with causes or relationships.
- The major purpose of univariate analysis is **to describe**.
- Univariate data uses central tendency: mean, mode, median.
- Its use **dispersion** method like **range, variance, max, min, quartiles, standard deviation**.
- frequency distributions
- Its result show in **bar graph, histogram, pie chart, line graph, box-and-whisker plot**

Bivariate Data

- Involving **two variables**.
- **Deals** with causes or relationships.
- The major purpose of bivariate analysis is **to explain**.
- Bivariate data uses analysis of two variables simultaneously.
- Its use **Correlations**
 - comparisons, relationships, causes, explanations.
- Its result show in **tables where one variable is contingent** on the values of the other variable.

18. Discuss different types of distance metrics using suitable expressions.

There are 4 types of Distance Metrics in Machine Learning.

1. Euclidean Distance
2. Manhattan Distance
3. Minkowski Distance
4. Hamming Distance

Euclidean Distance -

Euclidean Distance represents the shortest distance between two points

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

Manhattan Distance -

Manhattan Distance is the sum of absolute differences between points across all the dimensions.

$$d = |p_1 - q_1| + |p_2 - q_2|$$

Minkowski Distance

Minkowski Distance is the generalized form of Euclidean and Manhattan Distance.

$$D = \left(\sum_{i=1}^n |p_i - q_i|^p \right)^{1/p}$$

Hamming Distance

Hamming Distance measures the similarity between two strings of the same length. The Hamming Distance between two strings of the same length is the number of positions at which the corresponding characters are different.

Hamming Distance

$$D_H = \sum_{i=1}^k |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

19. Explain the Minkowski Distance is a generalization of the Manhattan and Euclidean Distance metrics.

Minkowski Distance is the generalized form of Euclidean and Manhattan Distance.

Minkowski Distance calculates the distance between two points. It is a generalization of the Euclidean and Manhattan distance measures and adds a parameter, called the “order” or “ p ”, that allows different distance measures to be calculated.

The Minkowski distance measure is calculated as follows:

$$D = \left(\sum_{i=1}^n |p_i - q_i|^p \right)^{1/p}$$

Where “ p ” is the order parameter.

When p is set to 1, the calculation is the same as the Manhattan distance. When p is set to 2, it is the same as the Euclidean distance.

- $p=1$: Manhattan distance.
- $p=2$: Euclidean distance.

Intermediate values provide a controlled balance between the two measures.

It is common to use Minkowski distance when implementing a machine learning algorithm that uses distance measures as it gives control over the type of distance measure used for real-valued vectors via a hyperparameter “ p ” that can be tuned.

22. What is KNN classification? Why is it called lazy learning?

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.

K-NN is a lazy learner because it doesn't learn a discriminative function from the training data but memorizes the training dataset instead. There is no training time in K-NN. The prediction step in K-NN is expensive. Each time we want to make a prediction, K-NN is searching for the nearest neighbors in the entire training set. An eager learner has a model fitting or training step. A lazy learner does not have a training phase.

24. Explain the advantages and disadvantages of KNN algorithm.

Some Advantages of KNN

- Quick calculation time
- Simple algorithm – to interpret
- Versatile – useful for regression and classification
- High accuracy – you do not need to compare with better-supervised learning models
- No assumptions about data – no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.

Some Disadvantages of KNN

- Accuracy depends on the quality of the data
- With large data, the prediction stage might be slow
- Sensitive to the scale of the data and irrelevant features
- Require high memory – need to store all of the training data
- Given that it stores all of the training, it can be computationally expensive

25. How many types of clustering are available in ML? Explain each type with examples.

Types of Clustering Methods :

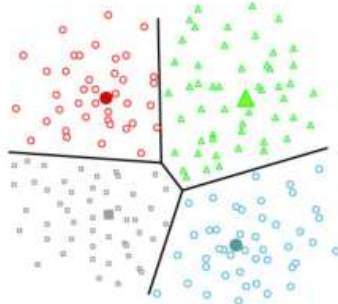
The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

1. **Partitioning Clustering**
2. **Density-Based Clustering**
3. **Distribution Model-Based Clustering**
4. **Hierarchical Clustering**
5. **Fuzzy Clustering**

Partitioning Clustering

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the K-Means Clustering algorithm.

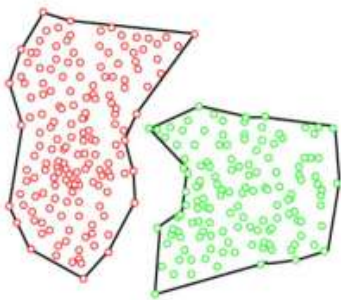
In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



Density-Based Clustering

The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.

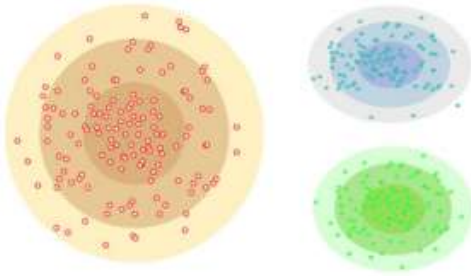
These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.



Distribution Model-Based Clustering

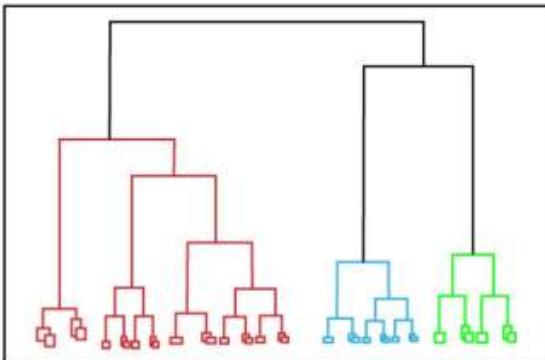
In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly Gaussian Distribution.

The example of this type is the Expectation-Maximization Clustering algorithm that uses Gaussian Mixture Models (GMM).



Hierarchical Clustering

Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm**.



Fuzzy Clustering

Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster. Each dataset has a set of membership coefficients, which depend on the degree of membership to be in a cluster. Fuzzy C-means algorithm is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm.

27. Explain the advantages and disadvantages of KMeans clustering.

K-Means Advantages :

- 1) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.
- 2) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

K-Means Disadvantages :

- 1) Difficult to predict K-Value.
- 2) With global cluster, it didn't work well.
- 3) Different initial partitions can result in different final clusters.
- 4) It does not work well with clusters (in the original data) of Different size and Different density

28. Differentiate between classification and regression.

Classification	Regression
<ul style="list-style-type: none">• Classification is the task of predicting a discrete class label• In a classification problem data is labelled into one of two or more classes• A classification problem with two classes is called binary, more than two classes is called a multi-class classification• Classifying an email as spam or non-spam is an example of a classification problem	<ul style="list-style-type: none">• Regression is the task of predicting a continuous quantity• A regression problem requires the prediction of a quantity• A regression problem with multiple input variables is called a multivariate regression problem• Predicting the price of a stock over a period of time is a regression problem

Classification vs Regression

- | | |
|---|---|
| <ul style="list-style-type: none">• Classification means to group the output into a class.• classification to predict the type of tumor i.e. harmful or not harmful using training data• if it is discrete/categorical variable, then it is classification problem | <ul style="list-style-type: none">• Regression means to predict the output value using training data.• regression to predict the house price from training data• if it is a real number/continuous, then it is regression problem. |
|---|---|

29. Why do we call the Linear Regression a linear model?

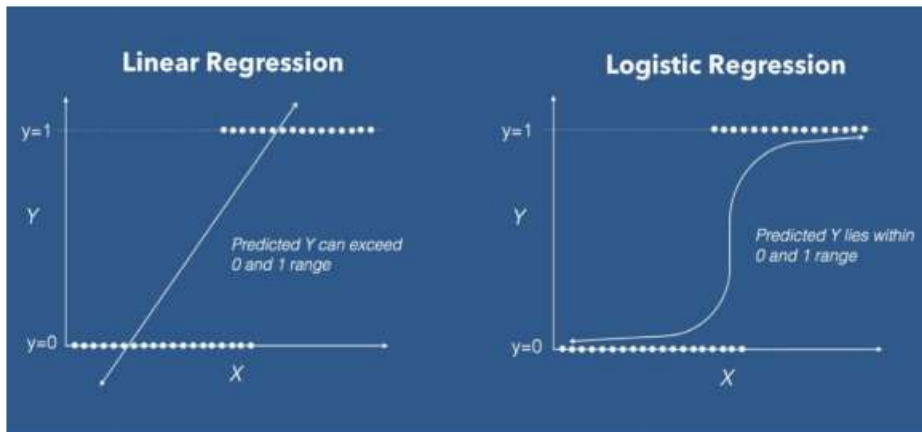
Linear regression is called linear because you model your output variable (lets call it $f(x)$) as a linear combination of inputs and weights (lets call them x and w respectively).

32. What is Polynomial Regression?

- Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + \dots + b_nx_1^n$$

- It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
- It is a linear model with some modification in order to increase the accuracy.
- The dataset used in Polynomial regression for training is of non-linear nature.
- It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.



■ Linear Regression	■ Logistic Regression
<ul style="list-style-type: none"> ■ Target is an interval variable. ■ Input variables have any measurement level. ■ Predicted values are the mean of the target variable at the given values of the input variables. 	<ul style="list-style-type: none"> ■ Target is a discrete (binary or ordinal) variable. ■ Input variables have any measurement level. ■ Predicted values are the probability of a particular level(s) of the target variable at the given values of the input variables.

34. Justify the name “Logistic Regression”.

Logistic Regression is one of the basic and popular algorithm to solve a classification problem. It is **named** as '**Logistic Regression**', because its underlying technique is quite the same as **Linear Regression**. The term “**Logistic**” is taken from the **Logit** function that is used in this method of classification.

35. Why do we use a logistic function in Logistic Regression?

36. Derive the log-loss cost function of Logistic Regression using step by step explanation.

37. What are the advantages and disadvantages/drawbacks of Linear Regression algorithm?

Advantages	Disadvantages
Linear Regression is simple to implement and easier to interpret the output coefficients.	On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity to compared to other algorithms.	Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.	But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

39. What is regularization?

Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

The commonly used regularisation techniques are :

1. L1 regularisation
2. L2 regularisation
3. Dropout regularisation

Regularization

- The loss function becomes
- The loss function of regularization doesn't depend on data.
- Common regularizations are

- L2 Regularization: $J_{reg}(\vec{w}) = \lambda \sum_{i=1}^N w_i^2$

- L1 Regularization: $J_{reg}(\vec{w}) = \lambda \sum_{i=1}^N |w_i|$

- Elastic-Net Regularization: $J_{reg}(\vec{w}) = \lambda \sum_{i=1}^N \left(\frac{\alpha}{2} w_i^2 + (1-\alpha)|w_i| \right)$



40. Explain different types of regularization in ML using appropriate Examples.

L1 Regularization or Lasso Regularization

L1 Regularization or Lasso Regularization adds a penalty to the error function. The penalty is the sum of the **absolute** values of weights.

$$\text{Min}(\sum_{i=1}^n (y_i - w_i x_i)^2 + p \sum_{i=1}^n |w_i|)$$

p is the tuning parameter which decides how much we want to penalize the model.

L2 Regularization or Ridge Regularization

L2 Regularization or Ridge Regularization also adds a penalty to the error function. But the penalty here is the sum of the **squared** values of weights.

$$\text{Min}(\sum_{i=1}^n (y_i - w_i x_i)^2 + p \sum_{i=1}^n (w_i)^2)$$

Similar to L1, in L2 also, p is the tuning parameter which decides how much we want to penalize the model.

49. Using K-mean Clustering Algorithm, Cluster the following data points: (5,7), (11,45), (10, 6), (18,29), (10,25), (4,3); where K=2 and Euclidean Distance.

	x	y
①	5	7
②	11	45
③	10	6
④	18	29
⑤	10	25
⑥	4	3

$K_1 \rightarrow \{ \{5, 7\} \}$
 $K_2 \rightarrow \{ \{11, 45\} \}$

for ③

$$r_1 = \sqrt{(18-5)^2 + (29-7)^2} = 5.07$$

$$r_2 = \sqrt{(10-11)^2 + (6-45)^2} = 39.01$$

New centroid for $K_1 = \left(\frac{5+10}{2}, \frac{7+6}{2} \right)$
 $= (7.5, 6.5)$

for ④

$$r_1 = \sqrt{(18-7.5)^2 + (29-6.5)^2} = 24.87$$

$$r_2 = \sqrt{(18-11)^2 + (29-45)^2} = 17.46$$

$K_1 \rightarrow \{ \{5, 7\}, \{10, 6\} \}$
 $K_2 \rightarrow \{ \{11, 45\}, \{18, 29\} \}$

New centroid for $K_2 = \left(\frac{11+18}{2}, \frac{45+29}{2} \right)$
 $= (14.5, 37)$

for ⑤

$$r_1 = \sqrt{(10-7.5)^2 + (25-6.5)^2} = 18.67$$

$$r_2 = \sqrt{(10-14.5)^2 + (25-37)^2} = 12.81$$

$K_1 \rightarrow \{ \{5, 7\}, \{10, 6\} \}$
 $K_2 \rightarrow \{ \{11, 45\}, \{18, 29\}, \{10, 25\} \}$

New centroid for $K_2 = \left(\frac{11+18+10}{3}, \frac{45+29+25}{3} \right)$
 $= (12.25, 31)$

for ⑥

$$r_1 = \sqrt{(4-7.5)^2 + (3-6.5)^2} = 4.95$$

$$r_2 = \sqrt{(4-12.25)^2 + (3-31)^2} = 25.85$$

$K_1 \rightarrow \{ \{5, 7\}, \{10, 6\}, \{4, 3\} \}$
 $K_2 \rightarrow \{ \{11, 45\}, \{18, 29\}, \{10, 25\} \}$

K_1 cluster has { ①, ③, ⑥ }
 K_2 cluster has { ②, ④, ⑤ }

50. Using KNN algorithm and the given data set, predict the label of the test data point (3,7), where K=3 and Euclidean distance.

X Y Label

7 7 1

7 4 1

3 4 2

1 4 2

	X	Y	Label
i)	7	7	1
ii)	7	4	1
iii)	3	4	2
iv)	1	4	2

$$D(x, i) = \sqrt{(3-7)^2 + (7-7)^2} = 4 \rightarrow N3 - 1$$

$$D(x, ii) = \sqrt{(3-7)^2 + (7-4)^2} = 5$$

$$D(x, iii) = \sqrt{(3-3)^2 + (7-4)^2} = 3 \rightarrow N1 - \text{TRUE } 2$$

$$D(x, iv) = \sqrt{(3-1)^2 + (7-4)^2} = 3.6 \rightarrow N2 - 2$$

Number of 2 > Number of 1

\therefore The label for (3,7) = 2