CM405PC INTRODUCTION TO MACHINE LEARNING
II YEAR B.TECH CSM II-SEM

UNIT1 QUESTION BANK

Q.NO.	UNIT-1	CO	BT	MARKS
1	Define machine learning. Briefly explain the importance and Evolution of machine learning.	1	2	5
2	Explain the need for machine learning	1	2	5
3	List out the applications of machine learning in detail.	1	3	5
4	What are the different types of Learning methods/models in ML?	1	4	5
5	What are the main key differences between supervised and unsupervised machine learning?	1	2	5
6	Differentiate between online and batch learning.	1	2	5
7	Explain the main Challenges of Machine Learning.	1	2	5

1	Define machine learning. Briefly explain the importance and Evolution of machine learning.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. As it is evident from the name, it gives the computer that makes it more similar to humans: ***The ability to learn***.

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning

Machine learning is an important component of the growing field of data science. Through the

use of statistical methods, algorithms are trained to make classifications or predictions.

The term Machine Learning was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence, and stated that "it gives computers the ability to learn without being explicitly programmed".

And in 1997, Tom Mitchell gave a "well-posed" mathematical and relational definition that "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

for example, traffic patterns at a busy intersection (task T), you can run it through a machine learning algorithm with data about past traffic patterns (experience E) and, if it has successfully "learned", it will then do better at predicting future traffic patterns (performance measure P).

Example: playing checkers.

E =the experience of playing many games of checkers

T = the task of playing checkers.

P = the probability that the program will win the next game

The highly complex nature of many real-world problems, though, often means that inventing specialized algorithms that will solve them perfectly every time is impractical, if not impossible. Examples of machine learning problems include, "Is this cancer?", "Which of these people are good friends with each other?", "Will this person like this movie?" such problems are excellent targets for Machine Learning, and in fact, machine learning has been applied to such problems with great success.

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.

Machine learning is much similar to data mining as it also deals with the huge amount of the data

2	Explain the need for machine learning

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement

directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

3	List out the applications of machine learning in detail.

1. **Image Recognition:**

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's **face detection** and **recognition algorithm**.

It is based on the Facebook project named "**Deep Face**," which is responsible for face recognition and person identification in the picture.

2. **Speech Recognition**

While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. **Traffic prediction:**

If we want to visit a new place, we take help of Google Maps, which shows us the correct path

with the shortest route and predicts the traffic conditions.

4. **Product recommendations:**

Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. **Self-driving cars:**

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. **Email Spam and Malware Filtering:**

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning.

7. **Virtual Personal Assistant:**

We have various virtual personal assistants such as **Google assistant**, **Alexa**, **Cortana**, **Siri**. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

8. **Online Fraud Detection:**

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as **fake accounts**, **fake ids**, and **steal money** in the middle of a transaction. So to detect this, **Feed Forward Neural network** helps us by checking whether it is a genuine transaction or a fraud transaction.

4	What are the different types of Learning methods/models in ML?

At a broad level, machine learning can be classified into four types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Semi supervised learning**
4. **Reinforcement learning**

1) Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

Some of the supervised learning applications are:

- Sentiment analysis (Twitter, Facebook, Netflix, YouTube, etc)
- Natural Language Processing
- Image classification

- Predictive analysis
- Pattern recognition
- Spam detection
- Speech/Sequence processin

Both classification and regression problems may have one and more input variables and input variables may be any data type, such as numerical or categorical.

An example of a classification problem would be the MNIST handwritten digits dataset where the inputs are images of handwritten digits (pixel data) and the output is a class label for what digit the image represents (numbers 0 to 9).

An example of a regression problem would be the Boston house pricesdataset where the inputs are variables that describe a neighborhood and the output is a house price in dollars.

2) Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

- **Clustering**
- **Association**

An example of a clustering algorithm is k-Means where *k* refers to the number of clusters to discover in the data.

Unsupervised Learning applications are:

1. Similarity detection
2. Automatic labeling
3. Object segmentation (such as Person, Animal, Films)

The goal in such unsupervised learning problems may be to discover groups of similar examples

within the data, where it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization.

3) Semi supervised Learning

Semi-Supervised Learning

Semi-supervised learning is supervised learning where the training data contains very few labeled examples and a large number of unlabeled examples.

The goal of a semi-supervised learning model is to make effective use of all of the available data, not just the labeled data like in supervised learning.

In semi-supervised learning we are given a few labeled examples and must make what we can of a large collection of unlabeled examples. Even the labels themselves may not be the oracular truths that we hope for

in many practical applications labeled data is very scarce but unlabeled data is plentiful. "Semisupervised" learning attempts to improve the accuracy of supervised learning by exploiting information in unlabeled data. This sounds like magic, but it can work!

4) Reinforcement Learning

Reinforcement learning is learning what to do — how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them.

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

An example of a reinforcement problem is playing a game where the agent has the goal of getting a high score and can make moves in the game and received feedback in terms of

punishments or rewards.

In many complex domains, reinforcement learning is the only feasible way to train a program to perform at high levels. For example, in game playing, it is very hard for a human to provide accurate and consistent evaluations of large numbers of positions, which would be needed to train an evaluation function directly from examples. Instead, the program can be told when it has won or lost, and it can use this information to learn an evaluation function that gives reasonably accurate estimates of the probability of winning from any given position.

5	What are the main key differences between supervised and unsupervised machine learning?

	Supervised Learning	Unsupervised Learning
Data Set	An example data set is given to the algorithm. This contains data that is already divided into specific categories/clusters (labeled data).	The algorithm is given data that does not have a previous classification (unlabeled data).
Use Cases	Well suited for making **classifications**, i.e. dividing the data according to the given clusters. Also, well suited for **regression analysis**, for example, to make predictions about the frequency of questions asked.	Well suited for **clustering**. In this way, patterns in data can be identified. Also, suitable for **association** analysis, in order to recognize which relationships exist among the data.

6	Differentiate between online and batch learning.

Batch learning represents the training of machine learning models in a batch manner. In other

words, batch learning represents the training of the models at regular intervals such as weekly, bi-weekly, monthly, quarterly, etc. The data gets accumulated over a period of time. The models then get trained with the accumulated data from time to time at periodic intervals. Batch learning is also called **offline learning**. The models trained using batch or offline learning are moved into production only at regular intervals based on the performance of models trained with new data.

Building offline models or models trained in a batch manner requires training the models with the entire training data set. Improving the model performance would require re-training all over again with the entire training data set. These models are static in nature which means that once they get trained, their performance will not improve until a new model gets re-trained. Offline models or models trained using batch learning are deployed in the production environment by replacing the old model with the newly trained model.

There can be various reasons why we can choose to adopt batch learning for training the models. Some of these reasons are the following:

- The business requirements do not require frequent learning of models.
- The data distribution is not expected to change frequently. Therefore, batch learning is suitable.
- The software systems (big data) required for batch learning is not available due to various reasons including the cost. The fact that the model is trained with a lot of accumulated data takes a lot of time and resources (CPU, memory space, disk space, disk I/O, network I/O, etc.).
- The expertise required for creating the system for incremental learning is not available.

If the models trained using batch learning needs to learn about new data, the models need to be retrained using the new data set and replaced appropriately with the model already in production based on different criteria such as model performance. The whole process of batch learning can be automated as well. The disadvantage of batch learning is it takes a lot of time and resources to re-training the model.

The criteria based on which the machine learning models can be decided to train in a batch manner depends on the model performance. Red-amber-green statuses can be used to determine the health of the model based on the prediction accuracy or error rates. Accordingly, the models can be chosen to be retrained or otherwise. The following stakeholders can be involved in reviewing the model performance and leveraging batch learning:

- Business/product owners
- Product managers
- Data scientists
- ML engineers

In online learning, the training happens in an incremental manner by continuously feeding data as it arrives or in a small group. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.

Online learning is great for machine learning systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously. **It is also a good option if you have limited computing resources:** once an online learning system has learned about new data instances, it does not need them anymore, so you can discard them (unless you want to be able to roll back to a previous state and "replay" the data) or move the data to another form of storage (warm or cold storage) if you are using the data lake. This can save a huge amount of space and cost.

Online learning algorithms can also be used to train systems on huge datasets that cannot fit in one machine's main memory (this is also called ***out-of-core learning***). The algorithm loads part of the data runs a training step on that data and repeats the process until it has run on all of the data.

One of the key aspects of online learning is the **learning rate.** The rate at which you want your machine learning to adapt to new data set is called the learning rate. A system with a high learning rate will tend to forget the learning quickly. A system with a low learning rate will be more like batch learning.

One of the big disadvantages of an online learning system is that if it is fed with bad data, the system will have bad performance and the user will see the impact instantly. Thus, it is very important to come up with appropriate data governance strategy to ensure that the data fed is of high quality. In addition, it is very important to monitor the performance of the machine learning system in a very close manner.

Data governance needs to be put in place across different levels such as the following when choosing to go with online learning:

- Feature extraction

- Predictions

The following are some of the challenges for adopting an online learning method:

- Data governance
- Model governance includes appropriate algorithm and model selection on-the-fly

Online models require only a single deployment in the production setting and they evolve over a period of time. The disadvantage that the online models have is that they don't have the entire dataset available for the training. The models are trained in an incremental manner based on the assumptions made using the available data and the assumptions at times can be sub-optimal.

7	Explain the main Challenges of Machine Learning.

There are a lot of challenges in machine learning

1. Poor Quality of Data

Data plays a significant role in the machine learning process. One of the significant issues that machine learning professionals face is the absence of good quality data. Unclean and noisy data can make the whole process extremely exhausting. We don't want our algorithm to make inaccurate or faulty predictions. Hence the quality of data is essential to enhance the output. Therefore, we need to ensure that the process of data preprocessing which includes removing outliers, filtering missing values, and removing unwanted features, is done with the utmost level of perfection.

2. Underfitting of Training Data

This process occurs when data is unable to establish an accurate relationship between input and output variables. It simply means trying to fit in undersized jeans. It signifies the data is too simple to establish a precise relationship. To overcome this issue:

- *Maximize the training time*
- *Enhance the complexity of the model*
- *Add more features to the data*
- *Reduce regular parameters*
- *Increasing the training time of model*

3. Overfitting of Training Data

Overfitting refers to a machine learning model trained with a massive amount of data that

negatively affect its performance. It is like trying to fit in Oversized jeans. Unfortunately, this is one of the significant issues faced by machine learning professionals. This means that the algorithm is trained with noisy and biased data, which will affect its overall performance. Let's understand this with the help of an example. Let's consider a model trained to differentiate between a cat, a rabbit, a dog, and a tiger. The training data contains 1000 cats, 1000 dogs, 1000 tigers, and 4000 Rabbits. Then there is a considerable probability that it will identify the cat as a rabbit. In this example, we had a vast amount of data, but it was biased; hence the prediction was negatively affected.

We can tackle this issue by:

- *Analyzing the data with the utmost level of perfection*
- *Use data augmentation technique*
- *Remove outliers in the training set*
- *Select a model with lesser* features

4. Machine Learning is a Complex Process

The machine learning industry is young and is continuously changing. Rapid hit and trial experiments are being carried on. The process is transforming, and hence there are high chances of error which makes the learning complex. It includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, and a lot more. Hence it is a really complicated process which is another big challenge for Machine learning professionals.

5. Lack of Training Data

The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. Less amount training data will produce inaccurate or too biased predictions. Let us understand this with the help of an example. Consider a machine learning algorithm similar to training a child. One day you decided to explain to a child how to distinguish between an apple and a watermelon. You will take an apple and a watermelon and show him the difference between both based on their color, shape, and taste. In this way, soon, he will attain perfection in differentiating between the two. But on the other hand, a machine-learning algorithm needs a lot of data to distinguish. For complex problems, it may even require millions of data to be trained. Therefore we need to ensure that Machine learning algorithms are trained with sufficient amounts of data.

6. Slow Implementation

This is one of the common issues faced by machine learning professionals. The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results. Further, it requires constant monitoring and maintenance to deliver the best output.

7. Imperfections in the Algorithm When Data Grows

The best model of the present may become inaccurate in the coming Future and require further rearrangement. So you need regular monitoring and maintenance to keep the algorithm working. This is one of the most exhausting issues faced by machine learning professionals.