**t1. What does one mean by the term "machine learning"?**

Machine Learning(ML) is a branch of artificial intelligence. It involves automation of various tasks by creating suitable models.

It is the art of programming that gives computers the ability to learn, without being explicitly programmed.

The ML models are created using training data and they evolve as the model learns the various parameters. ML is about making machines get better at some task by learning from data, instead of having to explicitly code rules.

**2. Can you think of 4 distinct types of issues where it shines?**

Some common applications where ML really shines are, spam filters used in emails, search engine optimizations, the algorithms used by social networking sites for showing advertisements that are user centric, algorithms used by streaming services like Netflix, YouTube for generating suggestions for users, finding fraudulent credit card transactions, speech recognition, etc.

These areas can be generalized as follows:

1. Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform bet‐ ter.
2. Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
3. Fluctuating environments: a Machine Learning system can adapt to new data.
4. Getting insights about complex problems and large amounts of data.

**3. What is a labeled training set, and how does it work?**

The examples that the system uses to learn are called the training set. In supervised learning, the training data you feed to the algorithm includes the desired solutions, also called as labels. Such training data is called labelled training set.

**4. What are the two most important tasks that are supervised?**

Most important tasks that are supervised are:

1. **Classification**: Example, spam filter that classifies a given email as spam or not spam.
2. **Regression**: Example, predicting price of a car, given set of features such as mileage, brand, age, etc.

**5. Can you think of four examples of unsupervised tasks?**

1. Clustering a website users into groups
2. Image detection, facial recognition
3. Find correlation between features so that they can combined to make simple data without losing much information. (dimensionality reduction)
4. Finding unusual credit card transactions to prevent fraud.

**6. State the machine learning model that would be best to make a robot walk through various unfamiliar terrains?**

Reinforcement learning, which is based on positive and negative rewards based system for every action performed can be used for making a robot that can walk through various unfamiliar terrains.

**7. Which algorithm will you use to divide your customers into different groups?**

If the training data doesn’t have any labels, clustering algorithm can be used for dividing customers into groups.

If groups are pre-defined, classification algorithms can be used.

**8. Will you consider the problem of spam detection to be a supervised or unsupervised learning problem?**

Traditionally, spam filtering is a classification problem, which labels the email as spam or not spam. The model learns this through labelled training data.

However, more recent researches also check for the possibility of applying unsupervised techniques as well to achieve better learning capability.

**9. What is the concept of an online learning system?**

In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches.

**10. What is out-of-core learning, and how does it differ from core learning?**

Out-of-core learning is used to train systems on huge datasets that cannot fit in one machine’s main memory. It uses online learning strategy where 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.

**11. What kind of learning algorithm makes predictions using a similarity measure?**

Instance based learning algorithm.

The system learns the examples by heart, then generalizes to new cases by comparing them to the learned examples (or a subset of them), using a similarity measure.

**12.What's the difference between a model parameter and a hyperparameter in a learning algorithm?**

A hyperparameter is a parameter of a learning algorithm (not of the model). As such, it is not affected by the learning algorithm itself; it must be set prior to training and remains constant during training.

Model parameter, on the other hand is a parameter of the model. There can be more than one model parameter and its value can change during the learning process.

**13.What are the criteria that model-based learning algorithms look for? What is the most popular method they use to achieve success? What method do they use to make predictions?**

Model-based learning algorithms look for performance measures to determine the quality of the parameters used in model creation.

The most popular methods used are:

1. Utility function or fitness function: - measures how good your model is
2. Cost function: - measures how bad the model is.

Typically, people use a cost function that determines the distance between predicted value and actual value in the training set. The goal is to learn the optimal model parameters that minimizes this distance.

Once the model is trained and the optimal parameter values are learned, we use these parameter values and the model to make predictions for new instances of data that are not included in the training set.

**14. Can you name four of the most important Machine Learning challenges?**

1. Insufficient Quantity of Training Data – Most ML models require a lot of data to work properly.
2. Nonrepresentative Training Data – the training data should be representative of the new cases you may want to generalize to.
3. Poor-Quality Data - if your training data is full of errors, outliers, and noise, it will make it harder for the system to detect the underlying patterns.
4. Irrelevant Features - the training data contains enough relevant features and not too many irrelevant ones.

**15. What happens if the model performs well on the training data but fails to generalize the results to new situations? Can you think of three different options?**

If the model performs well on the training data but fails to generalize the results to new situations, it is called overfitting. Overfitting happens when the model is too complex relative to the amount and noisiness of the training data. The possible solutions are:

* simplify the model by selecting one with fewer parameters
* gather more training data
* reduce the noise in the training data

**16. What exactly is a test set, and why would you need one?**

To evaluate the performance of a model, we train the model using a training set and then test the model using test set. To do this, we split the data into train and test set before we train our model. In this way, we can determine how well the model performs when it encounters data that it has never seen.

If we use same data used in training for testing the model, we cannot accurately measure the performance of model when it encounters previously unknown data and it may lead to overfitting of the model with training data.

**17. What is a validation set's purpose?**

Validation set is a set created by holding out part of training set. It is used for evaluating several candidate models and select the best one.

It’s a good idea to have separate data for tuning the hyperparameters of a model that is not the test set so that we obtain a model that generalizes in the best possible way.

This additional data split is called the ***validation*** set.

**18. What precisely is the train-dev kit, when will you need it, how do you put it to use?**

The validation and test data must be representative of the data that will be used in production. If we use web downloaded data for training, there is a possibility that the model we develop may overfit the training set, and the performance of your model on the validation set will be disappointing or it could be due to other reasons.

To find the actual cause of the issue, hold out part of the training pictures (from the web) in yet another set called the train-dev set. After the model is trained (on the training set, not on the train-dev set), you can evaluate it on the train-dev set: if it performs well, then the model is not overfitting the training set, so if performs poorly on the validation set, the problem must come from the data mismatch.

**19.What could go wrong if you use the test set to tune hyperparameters?**

If generalization error is measured multiple times on the test test and hyperparameters are tuned accordingly, that model is adapted to best fit for that data only. It is unlikely to perform as well on new data.

**References**:

Aurélien, G. (2017). Hands-on machine learning with scikit-learn & tensorflow. *Geron Aurelien*.