1. How would you define Machine Learning?

Answer: Machine learning is the field of study which gives computer the capability to learn from experience without explicitly programmed. The most common example is spam filtering where the system learns from training dataset and classifies the spam and non spam emails using some performance measure namely accuracy.

2. Can you name four types of problems where it shines?

Answer:

1. In problems where there are too many hard and fast rules to program a system whether it is spam or not. Instead of explicitly programming the system we can use machine learning to achieve a same or better level of success in classifying the problem.
2. For problems where it is impossible for the traditional programming to solve a complex problem such as speech recognition.
3. To find out underlying trends inside huge amount of data. This can be used to find out insights regarding business problems and take decision accordingly.
4. To detect outliers in the data or forecast future trends by analyzing historic data of huge volume.

3. What is a labeled training set?

Answer: Labeled training data can be defined as the output variable or values for certain input features. For example, there can be n number of features that can be used to find out whether am email is spam or not. In this case, there will be an attribute which would tell us whether the email is actually spam or not (1 or 0). This labeled data would be used to train the system in predicting future emails. This is part of supervised learning problem.

4. What are the two most common supervised tasks?

Answer:

1. Classification: The training data fed to the algorithm includes the desired output in the form of labeled data which consists of discrete values like 1 or 0. The system output can either be a binary class (1 or 0) or multi class classification problem (A, B, C, or D).
2. Regression: The training set consists of labeled data to train the algorithm however the system output is a continuous value like age of a person, salary of a person or price of second hand cars.

5. Can you name four common unsupervised tasks?

Answer:

1. Clustering to detect customers who belong to a similar group.
2. Visualization techniques where we feed a lot of unlabeled data and receive an output in a 2D or 3D plot.
3. Dimensionality reduction, is where we simplify the data without losing too much information.
4. Anomaly detection technique is used to remove the outliers from the dataset before feeding it to the other system.

6. What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?

Answer: Reinforcement learning is used to train the robots where we either penalize the system whenever it makes mistakes or give a reward when it takes a good decision. The idea is to avoid making mistakes by penalizing them.

7. What type of algorithm would you use to segment your customers into multiple groups?

Answer: We’ll use an unsupervised learning algorithm name K-Means, DBScan or Hierarchical clustering which divides the data points in a number of pre-defined clusters based on the trend or similarity of their transactions.

8. Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?

Answer: Spam detection can be framed as a supervised learning problem, since we would have a set of labeled training data telling us which historic emails were marked as spam or non-spam. There can be many factors which would help us mark a spam email like, use of certain keywords like offers, discount, 4you, third party unsecure URL, etc. along with the dependent output label with value like 1 or 0. The algorithm will find pattern amongst the input features and predict whether the new email is spam or ham.

9. What is an online learning system?

Answer: Instead of training the production system again and again offline using huge amount of new training data. We can train the system incrementally in small batches. Each learning step is fast & cheap and is used for system which receives continuous flow of data like stock prices where we need to adapt to changes rapidly.

10. What is out-of-core learning?

Answer: It is a method of training online learning algorithms using huge amount of training data that cannot fit into machine’s main memory.

11. What type of learning algorithm relies on a similarity measure to make predictions?

Answer: Instance based learning relies on measure of similarity to make predictions. In this case, the system learns the training examples by heart and makes predictions based on similarity it finds in the training set. This is a naïve way of predicting the output as it is unable to generalize well on future data if such similar training instance is not present in the training dataset.

12. What is the difference between a model parameter and a learning algorithm’s hyperparameter?

Answer: Model parameter is a variable of the selected model which can be estimated by fitting given data to the model. For example, while plotting a regression line, we intend to map a function between the input variables x and output variable y in the form of a straight line. In this equation y = mx + c, m and c is the slope and intercept and are estimated by fitting a straight line to the data by minimizing the RMSE. These parameters are called model parameters.

Model hyperparameters are set before the model starts training. Like in random forest algorithm, we mentioned the depth of trees, minimum sample size, leaf size, etc. These are parameters are called model hyper parameters.

13. What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?

Answer: Model based learning algorithms, first fits the model using training examples and then use the model to make predictions. They search for model parameters which are m and c in case of binary linear regression problem.

First initialize the values of m and c, and to know if a model is performing well we define a cost function. The cost function helps us understand how well our model is performing and gradient descent algorithm helps us update the model parameters in order to minimize the residual values.

14. Can you name four of the main challenges in Machine Learning?

Answer:

1. Insufficient quantity of training data.
2. Non presentative training data.
3. Poor quality data.
4. Irrelevant features.

15. If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?

Answer: We can say our model is overfitting when it performs well on training data and unable to generalize well on the unseen data.

1. Simplify the model by selecting lesser more relevant features.
2. Gather more training data to train the model.
3. Reduce the noise and remove outliers from the training data.

16. What is a test set and why would you want to use it?

Answer: In order to test how well our model is able to generalize on unseen data. We take out a subset of data from our training examples probably a ratio of 30% and test the predictions of our model using this test data.

17. What is the purpose of a validation set?

Answer: We do not want our model to adapt to the training data when we are trying various models to check, which gives us the best predictions. We create validation set from our training examples and test the model’s predictions using the same. We can use k fold cross validation for this purpose and check how well our model is performing. We can further tweak the hyperparameters to improve its accuracies and at last confirm our predictions using test data set.

18. What can go wrong if you tune hyperparameters using the test set?

Answer: During hyperparameter tuning, we train multiple different models using 100 different values of these hyperparameters to find the model with the lowest generalization error. In this process our model adapts to the hyperparameters and training instances due to which it performs well on the training data. However, it is unable to generalize well on test data leading to overfitting.

To avoid this, we hold out a validation dataset from the training examples to evaluate several models and hyperparameters during training.

19. What is repeated cross-validation and why would you prefer it to using a single validation set?

Answer: If we use single validation set, if the validations set is too small, this would result in imprecise model evaluation and if it is too large, then there wouldn’t be much training examples to train the model on.

One way to solve this problem is by cross validation, where we have multiple small validation sets. Each model is evaluated using this small validation set and rest of the instances are used for training purpose. By averaging out the evaluation of each models, we get a much better measure of their performances.