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

* Machine learning is a method of data analysis that automates analytical model building.
* It is a file that has been trained to recognize certain types of patterns.

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

1. spam detection in email
2. fraudulent credit card transactions
3. automatically driving vehicles.
4. cancer diagnosis

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

It used to train the algorithm, and then you use the trained model on the test set to predict the response variable values that are already known

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

1. Regression
2. classification

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

1. Clustering
2. dimensionality reduction
3. visualization

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

Reinforced Learning

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

K-Means clustering

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

A spam detection is a classic supervised learning problem-Because at the end of the day my goal is to know if a user is either a spammer or not - a standard binary classification task.

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

An online learning system learns from new data on-the-fly. As a result, the system is trained incrementally either by using one example at a time or using a mini-batch approach. This keeps each learning step cheap and memory efficient.

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

Out-of-core learning is used when a dataset is too large to fit into a computer's memory. The algorithm loads part of the data, runs a training step, then repeats the process until it has run on all the data.

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

instance-based algorithm

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

Basically, parameters are the ones that the “model” uses to make predictions etc. For example, the weight coefficients in a linear regression model. Hyperparameters are the ones that help with the learning process. For example, number of clusters in K-Means, shrinkage factor in Ridge Regression

**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?**

The goal for a model-based algorithm is to be able to generalize to new examples. To do this, model based algorithms search for optimal values for the model's parameters, often called theta. This searching, or "learning", is what machine learning is all about. Model-based system learn by minimizing a cost function that measures how bad the system is at making predicitons on new data, plus a penalty for model complexity if the model is regularized. To make a prediction, a new instance's features are fed into a hypothesis function which uses the minimized theta found by repeatedly running the cost function.

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

1. lacking in data
2. overfitting the data
3. underfitting the data
4. nonrepresentative data

**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** poor on the **training set and** equally poor according to the cross validation scores, it is underfitting

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

The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

*A set of examples used only to assess the performance of a fully-specified classifier.*

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

 The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

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

* The split between train / dev / test should **always be the same** across experiments
  + otherwise, different models are not evaluated in the same conditions
  + we should have a **reproducible script** to create the train / dev / test split
* we need to test if the **dev** and **test** sets should come from the same distribution

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

You'll be much more likely to overfit.

You need to have at least a vague idea of how your model will do on previously unseen instances. The training set itself cannot do that, unless you split it into training and validation set. I recommend you use at least 10-fold crossvalidation.