How would you define Machine Learning?

Machine Learning is the science (and art) of programming computers so they can learn from data.

Can you name four types of problems where it shines?

- Prediction
- Classification
- Generative Text
- Visualization of data

What is a labeled training set?

It's a dataset of your data points where you already know the result, ex. photos of cat and dog and you labeled each phot with the type of the animal.

What are the two most common supervised tasks?

Prediction and Classification

Can you name four common unsupervised tasks?

- Clustering
- Anomaly detection
- Visualization

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

reinforcement learning

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

Clustering

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

Supervised learning

What is an online learning system?

Online means that you keep teaching the AI on the fly for the new scenarios that will appear, unlike the batch where you teach the AI on a dataset and deploy.

What is out-of-core learning?

if your memory doesn't fit for full dataset because it's too large, the process of learning on chunks is called out-of-core learning.

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

k-Nearest Neighbors

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

Model parameters are the internal coefficients learned from the training data to make predictions, while hyperparameters are the external settings configured before training that govern the training process and model structure.

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

it's trying to generalize your data, the most comment is the loss function or the cost function, they make the prediction via mathematical equations.

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- Irrelevant Features
- Overfitting the Training Data
- Underfitting the Training Data

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

Overfitting

- dataset size
- complexity of the model
- cross validation

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

unseen data that we use to evaluate the performance.

What is the purpose of a validation set?

to tune the hyperparameters and to choose the best algorithm or model

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

it will over-fit the production data and will give wrong evaluation of the model.

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

Repeated cross-validation is a technique where we test our model multiple times with different random splits of the data to get a more reliable and accurate estimate of how well it will perform on new data, rather than relying on just one split.