Task 11

Exercises

1. How would you define Machine Learning?

Answer: Machine Learning is a field of Artificial Intelligence that uses statistical techniques to give computer systems the ability to learn from data and improve from experience without being explicitly programmed.

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

Answer: The four types of problems are:

- Classification
- Regression
- Clustering
- Anomaly Detection

3. What is a labeled training set?

Answer: A labeled training set is a dataset used to train machine learning models, consisting of input-output pairs where each input is associated with known output label.

4. What are the two most common supervised tasks?

Answer: The two most common supervised tasks are:

- Classification
- Regression

5. Can you name four common unsupervised tasks?

Answer: The four common unsupervised tasks are:

- Clustering
- Dimensionality reduction
- Anomaly detection
- Association rule learning

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

Answer: To allow a robot to walk in various unknown terrains, we would use a Reinforcement Learning algorithm. RL algorithms are particularly well-suited for this type of task because they involve learning how to make sequences of decisions by interacting with an environment to achieve a goal.

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

Answer: To segment our customers into multiple groups, we would use a Clustering algorithm. Clustering is an unsupervised learning technique that identifies patterns or groupings within data based on similarities. Some common clustering algorithms include:

- K-means Clustering: Divides the dataset into K clusters based on the nearest mean.
- Hierarchical Clustering: Builds a hierarchy of clusters either in a bottom-up agglomerative or top-down divisive approach.

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

Answer: Spam detection is typically framed as a supervised learning problem.

9. What is an online learning system?

Answer: An online learning system is a model that learns incrementally by processing one instance at a time, which is useful for situations where data arrives in a sequential manner.

10. What is out-of-core learning?

Answer: Out-of-core learning is a method used to train models on datasets that are too large to fit into memory, by using data streaming techniques to process the data in small chunks.

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

Answer: A type of learning algorithm that relies on a similarity measure to make predictions is an Instance-based learning algorithm. One of the most common instance-based learning algorithms is the k-Nearest Neighbors (k-NN) algorithm.

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

Answer:

Model Parameter:

- These are the internal variables of the model that are learned from the training data.
- Weights in a linear regression model, coefficients in a logistic regression model, and split points in a decision tree.
- They are automatically adjusted by the learning algorithm during the training process to minimize the cost function and improve the model's performance.

Hyperparameter:

- These are external configurations set before the training process begins and are used to control the learning process.
- Learning rate in gradient descent, the number of neighbors (k) in k-NN, the number of hidden layers in a neural network, and the maximum depth of a decision tree.
- They are not learned from the data but are typically set manually by the practitioner. Hyperparameters are often tuned through processes like grid search, random search, or cross-validation to find the best combination for the model's performance.

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 search for the best model parameters that minimize a cost function. The most common strategy is optimization (e.g., gradient descent). They make predictions by applying the model to new data using the learned parameters

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

Answer: The four main challenges are:

- Insufficient quantity of training data
- Poor quality of data
- Non-representative training data
- Overfitting and underfitting

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: The model is overfitting, Possible solutions:

- Simplifying the model (reducing its complexity)
- Using regularization techniques
- Gathering more training data

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

Answer: A test set is a separate dataset used to evaluate the performance of a trained model to ensure it generalizes well to new, unseen data.

17. What is the purpose of a validation set?

Answer: A validation set is used during model training to tune hyperparameters and make decisions about model selection to avoid overfitting.

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

Answer: Tuning hyperparameters using the test set can lead to overfitting on the test data, resulting in an overly optimistic estimate of the model's performance.

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

Answer: Repeated cross-validation involves running multiple rounds of cross-validation and averaging the results. It is preferred over using a single validation set because it provides a more reliable estimate of model performance by reducing variance and minimizing the risk of overfitting to a particular validation split.