1. 머신러닝을 어떻게 정의할 수 있나요?

-> 머신러닝은 데이터로부터 학습할 수 있는 시스템을 만드는 것. 학습이란 어떤 작업에서 주어진 성능 지표가 더 나아지는 것을 의미함.

2. 머신러닝이 도움을 줄 수 있는 문제 유형 네 가지를 말해보세요.

-> 사람에게 통찰을 제공해야 하는 경우, 복잡한 문제, 수작업의 규칙 리스트를 대체하는 경우, 변화하는 환경에 적응하는 시스템을 만드는 경우

3. 레이블된 훈련 세트란 무엇인가요?

-> 각 샘플에 대해 원하는 정답을 담고 있는 훈련 세트

4. 가장 널리 사용되는 지도 학습 작업 두 가지는 무엇인가요?

-> 분류와 회귀

5. 보편적인 비지도 학습 네 가지는 무엇인가요?

-> 군집화, 축소, 연관 규칙, 시각화

6. 사전 정보가 없는 여러 지형에서 로봇을 걸어가게 하려면 어떤 종류의 머신러닝 알고리즘을 사용할 수 있나요?

-> 강화학습

7. 고객을 여러 그룹으로 분할하려면 어떤 알고리즘을 사용해야 하나요?

-> 비지도 학습의 군집 알고리즘. 어떤 그룹이 있어야 할지 안다면 지도 학습의 분류 알고리즘

8. 스팸 감지의 문제는 지도 학습과 비지도 학습 중 어떤 문제로 볼 수 있나요?

-> 지도 학습

10. 외부 메모리 학습이 무엇인가요?

-> 점진적 학습 시스템. 변화하는 데이터에 빠르게 적응하며 많은 양의 데이터를 훈련시킬 수 있음.

11. 예측을 하기 위해 유사도 측정에 의존하는 학습 알고리즘은 무엇인가요?

-> 인스턴스 기반 학습 시스템

12. 모델 파라미터와 학습 알고리즘의 하이퍼파라미터 사이에는 어떤 차이가 있나요?

-> 모델은 하나 이상의 파라미터를 사용해 새로운 샘플이 주어지면 무엇을 예측할지 결정함. 하이퍼파라미터는 모델이 아니라 알고리즘 자체의 파라미터임. 적용한 규제 정도를 예로 들 수 있음.

13. 모델 기반 알고리즘이 찾는 것은 무엇인가요? 성공을 위해 이 알고리즘이 사용하는 가장 일반적인 전략은 무엇인가요? 예측은 어떻게 만드나요?

-> 모델 파라미터의 최적값을 찾음. 예측을 만들려면 모델의 예측함수에 새로운 샘플의 특성을 주입함.

14. 머신러닝의 주요 도전 과제는 무엇인가요?

-> 부족한 데이터, 낮은 데이터 품질, 대표성 없는 데이터, 무의미한 특성, 언더피팅, 오버피팅 등이 있음.

15. 모델이 훈련 데이터에서의 성능은 좋지만 새로운 샘플에서의 일반화 성능이 나쁘다면 어떤 문제가 있는 건가요? 가능한 해결책 세 가지는 무엇인가요?

-> 과적합. 더 많은 데이터 수집, 모델의 단순화, 훈련 데이터의 잡음 감소의 방법이 있음.

16. 테스트 세트를 왜 사용해야 하나요?

-> 새로운 샘플에 대해 만들 일반화 오차를 추정하기 위해서 사용함.

17. 검증 세트의 목적은 무엇인가요?

-> 모델 비교에 사용됨. 가장 좋은 모델을 고르고 하이퍼파라미터를 튜닝함.

18. 훈련-개발 세트가 무엇인가요? 언제 필요하고 어떻게 사용해야 하나요?

-> 검증, 테스트 세트에 사용되는 데이터와 훈련 세트 사이에 데이터 불일치 위험이 있을 때 사용함. 훈련 세트 일부에서 모델을 훈련하고 훈련-개발 세트와 검증 세트에서 평가함. 훈련 세트에서 잘 동작하지만 훈련-개발 세트에서 성능이 나쁘다면 오버피팅의 가능성이 있음. 개발 세트에서 성능이 나쁘다면 훈련 데이터와 검증+테스트 데이터에 불일치가 있을 가능성이 높음. 검증+테스트 데이터에 더 가깝게 되도록 훈련 데이터를 개선해야 함.

19. 테스트 세트를 사용해 하이퍼파라미터를 튜닝하면 어떤 문제가 생기나요?

-> 오버피팅의 위험이 있고 일반화 오차를 낙관적으로 측정하게 됨.