1주차 과제

노유민

1. 머신러닝을 어떻게 정의할 수 있나요?
데이터 사이의 관계를 나타내는 함수와 그 필요한 계수를 찾아내고 필요에 따라 새로운 데이터에 관한 예측을 하는 것. 아서 새뮤얼에 따르면 명시적인 프로그밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야, 톰 미첼에 따르면 작업 T에 대해 경험 E로 인해 성능 P가 향상되는 것을 의미한다.

2. 머신러닝이 도움을 줄 수 있는 문제 유형 네 가지를 말해보세요.
기존 솔루션으로 많은 수동 조정과 규칙이 필요한 문제, 전통적인 방식으로 해결 방법이 없는 문제, 데이터 셋이 새롭게 들어오는 유동적인 환경의 문제, 복잡한 문제와 대량의 데이터에서 통찰을 얻어야 하는 문제에 머신러닝 도움을 줄 수 있다.

3. 레이블된 훈련 세트란 무엇인가요?
훈련 데이터(x)가 각각 정해진 답(y)을 포함하고 있는 데이터 셋을 의미한다.

4. 가장 널리 사용되는 지도 학습 작업 두 가지는 무엇인가요?
특성을 사용하여 데이터 분류하는 분류 작업과 특성을 사용해 타깃 수치를 예측하는 회귀 작업이 있다.

5. 보편적인 비지도 학습 작업 네 가지는 무엇인가요?
데이터를 비슷한 특징을 가진 그룹으로 나누는 군집과 데이터의 특성을 눈에 보이도록 시각화하는 것, 상관관계가 있는 여러 특성을 하나로 합쳐 데이터의 특성을 줄이는 차원 축소, 정상 샘플을 이용하여 입력 샘플의 정상 여부를 판단하는 이상치 탐지, 오염되지 않은 훈련 세트를 활용하여 다른 데이터를 감지하는 특이치 탐지가 있다.

6. 사전 정보가 없는 여러 지형에서 로봇을 걸아가게 하려면 어떤 종류의 머신러닝
알고리즘을 사용할 수 있나요?
강화학습을 사용해 지형의 정보를 로봇이 스스로 학습하고 가장 큰 보상을 얻을 수 있는 경로로 로봇이 걸어가게 할 수 있다.

7. 고객을 여러 그룹으로 분할하려면 어떤 알고리즘을 사용해야 하나요?
모델 기반 학습을 통해 새로운 데이터를 특성에 따라 분류할 수 있다.

8. 스팸 감지의 문제는 지도 학습과 비지도 학습 중 어떤 문제로 볼 수 있나요?
여러 데이터의 특성을 파악하여 스팸으로 의심되는 문자를 찾아내야 하고, 스팸의 유형은 계속해서 발전하기 때문에 비지도 학습 문제로 볼 수 있다.

9. 온라인 학습 시스템이 무엇인가요?

적은 양의 데이터를 사용해 점진적으로 새로운 데이터를 주입하며 훈련하는 학습 시스템 종류이다.

10. 외부 메모리 학습이 무엇인가요?
빅데이터를 분석할 때 이를 부분으로 나누어 학습을 반복하는 온라인 학습의 예시이다.

11. 예측을 하기 위해 유사도 측정에 의존하는 학습 알고리즘은 무엇인가요?
사례 기반 학습이며, 이는 예측을 위해 기존 샘플을 기억하는 것이 훈련의 전부인 학습 시스템이다.

12. 모델 파라미터와 학습 알고리즘의 하이퍼파라미터 사이에는 어떤 차이가 있나요?
모델 파라미터는 모델에서 데이터 사이의 관계식에 사용되는 계수로 학습 알고리즘의 비용함수가 최소화되도록 값이 설정되고, 하이퍼파라미터는 학습 알고리즘 자체를 정의하는 데 작용하는 계수로 모델이 새로운 샘플에 맞게 더 좋은 성능을 가지도록 설정된다. 하이퍼파라미터는 모델 파라미터와 달리 훈련과정에서 변하지 않는다.

13. 모델 기반 알고리즘이 찾는 것은 무엇인가요? 성공을 위해 이 알고리즘이 사용하는 가장 일반적인 전략은 무엇인가요? 예측은 어떻게 만드나요?
모델 기반 알고리즘은 데이터를 대표할 수 있는 관계식을 찾는 것이 목적이며, 이를 위해 대표적으로 선형 방정식을 찾는 선형 모델 학습을 활용한다. 새로운 데이터에 모델을 적용해 예측을 만든다.

14. 머신러닝의 주요 도전 과제는 무엇인가요?
충분하지 않은 훈련 데이터, 대표성이 없는 훈련 데이터, 낮은 품질의 데이터, 관련이 없는 특성이 있을 때 대처하는 방법을 찾는 것이 주요 도전 과제다.

15. 모델이 훈련 데이터에서의 성능은 좋지만 새로운 샘플에서의 일반화 성능이 나쁘다면 어떤 문제가 있는 건가요? 가능한 해결책 세 가지는 무엇인가요?
훈련 세트에 특화되어 일반화 성능이 떨어지는 과대적합 문제가 발생한 것이다. 이를 해결하기 위해 데이터를 전처리하여 일부 데이터만 사용하거나, 규제를 적용하거나, 데이터를 적당한 지점까지만 훈련하고 조기 종료하는 방법이 있다.

16. 테스트 세트가 무엇이고 왜 사용해야 하나요?
전체 데이터의 약 20%로 모델이 다 완성된 후 모델의 성능을 평가하기 위한 데이터 세트이다. 테스트 세트가 없다면 모델이 훈련 데이터에 과대적합되는 문제가 발생할 수 있기 때문에 이를 사용해야 한다.

17. 검증 세트의 목적은 무엇인가요?
다양한 하이퍼파라미터 값을 가진 후보 모델을 검증 세트 데이터로 평가함으로써 더욱 정확한 성능의 모델을 선택하기 위함이다.

18. 테스트 세트를 사용해 하이퍼파라미터를 튜닝하면 어떤 문제가 생기나요?
테스트 세트를 사용해 하이퍼파라미터를 튜닝하면 모델이 테스트 세트에 과대적합 되는 문제가 발생할 수 있다.