WeTIE 머신러닝 톺아보기

1 주차 과제-김서연

1. 머신러닝을 어떻게 정의할 수 있나요?
데이터로부터 성능이 향상되는 시스템을 만드는 것

2. 머신러닝이 도움을 줄 수 있는 문제 유형 네 가지를 말해보세요.
-기존 솔루션으로 많은 수동 조정과 규칙이 필요한 문제
-전통적인 방식으로 해결 방법이 없는 복잡한 문제
-유동적인 환경에 속해있는 문제
-복잡한 문제와 대량의 데이터에서 통찰 얻기

3. 레이블된 훈련 세트란 무엇인가요?
특징 정보를 담는 훈련 데이터에 사용자가 원하는 정답 정보인 레이블을 포함하고 있는 데이터셋를 의미한다.

4. 가장 널리 사용되는 지도 학습 작업 두 가지는 무엇인가요?
분류, 회귀

5. 보편적인 비지도 학습 작업 네 가지는 무엇인가요?
군집, 시각화, 차원 축소, 연관 규칙 학습

6. 사전 정보가 없는 여러 지형에서 로봇을 걸어가게 하려면 어떤 종류의 머신러닝 알고리즘을 사용할 수 있나요?
강화학습 알고리즘, 강화학습 알고리즘은 사전 정보가 없는 상태에서 시작하여 스스로 최적의 알고리즘을 찾아내는 학습 방법이기 때문이다.

7. **고객을 여러 그룹으로 분할하려면 어떤 알고리즘을 사용해야 하나요?**
1) 고객의 그룹이 사전에 정의되어 있는 경우-지도 학습의 분류 알고리즘
2) 고객의 그룹이 사전에 정의되어 있지 않은 경우-비지도 학습의 군집 알고리즘

8. 스팸 감지의 문제는 지도 학습과 비지도 학습 중 어떤 문제로 볼 수 있나요?
지도 학습, 기존에 스팸인지 아닌지를 알 수 있는 레이블 된 데이터셋을 이용해서 학습하기 때문이다.

9. 온라인 학습 시스템이 무엇인가요?
데이터를 순차적으로 한 개 또는 미니 배치를 이용해서 학습하는 방법

10. 외부 메모리 학습이 무엇인가요?

컴퓨터의 메모리에 들어갈 수 없는 아주 큰 데이터셋을 학습할 때, 한 번에 모두 메모리에 넣는 것이 아니라 점진적으로 넣으면서 학습하는 과정

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

12. 모델 파라미터와 학습 알고리즘의 하이퍼파라미터 사이에는 어떤 차이가 있나요?
모델 파라미터는 모델이 학습을 하는 과정에서 점차 최적화되는 파라미터인 반면, 하이퍼파라미터는 모델이 학습을 시작하기 이전에 사람이 직접 입력해주어야 하는 값이라는 점에서 차이가 있다.

13. 모델 기반 알고리즘이 찾는 것은 무엇인가요? 성공을 위해 이 알고리즘이 사용하는 가장 일반적인 전략은 무엇인가요? 예측은 어떻게 만드나요?
-목표: 최적화된 모델 파라미터를 찾는 것
-성공을 위해 사용하는 일반적인 전략: 데이터를 바탕으로 예측을 진행하고, 손실함수와 비용함수를 최소화하는 방향으로 모델 파라미터를 업데이트하고 다시 그 과정을 반복해가며 모델 파라미터를 최적화해나간다. 이때 자주 사용되는 알고리즘 중 하나는 경사하강법이다.
-예측을 만드는 법: 모델이 학습을 마친 후에 새로운 데이터를 입력하여 예측값을 출력한다.

14. 머신러닝의 주요 도전 과제는 무엇인가요?
부족한 데이터, 낮은 데이터 품질, 대표성 없는 데이터, 무의미한 특성, 과소적합, 과대적합

15. 모델이 훈련 데이터에서의 성능은 좋지만 새로운 샘플에서의 일반화 성능이 나쁘다면 어떤 문제가 있는 건가요? 가능한 해결책 세 가지는 무엇인가요?
-문제: 모델이 훈련 데이터에 과대적합되었을 가능성이 높다
-해결책: 데이터를 더 많이 모으기, 데이터를 더 잘 정제하기, 모델을 단순화하기

16. 테스트 세트가 무엇이고 왜 사용해야 하나요?
테스트 세트는 모델이 학습을 마친 후에 마지막으로 모델에 돌려보는 데이터 세트이다. 실제로 현실에서 모델이 사용되기 전에 처음 사용해 보는 데이터에 대해 발생하는 오차를 의미하는 일반화 오차를 추정하기 위해 사용한다.

17. 검증 세트의 목적은 무엇인가요?
모델이 훈련 데이터에 과대적합 되지 않았는지 검증하는 용도로 쓰이고, 하이퍼파라미터를 결정할 때 사용된다.

18. 테스트 세트를 사용해 하이퍼파라미터를 튜닝하면 어떤 문제가 생기나요?

모델이 테스트 데이터에 과대적합할 가능성이 커지고, 이로 인해 신뢰할 수 있는 일반화 오차 값을 구하기 어려워진다.