1주차 과제

1. **머신러닝을 어떻게 정의할 수 있나요?**

   1) 아서 사무엘의 정의

   명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구분야 (1959)

   2) 톰 미첼의 정의

   어떤 작업 T에 대한 컴퓨터 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로 학습한 것이다. (1997)

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

   1) 기존 솔루션으로 많은 수동 조정과 규칙이 필요한 문제: 하나의 머신러닝 모델이 코드를 간단하게 만들고 전통적인 방법보다 수행능력 향상

   2) 전통적인 방식으로 해결 방법이 없는 복잡한 문제

   3) 유동적인 환경: 머신러닝 시스템은 새로운 데이터에 적응

   4) 복잡한 문제와 대량의 데이터에서 통찰 얻기

3. **레이블된 훈련 세트란 무엇인가요?**

   훈련 데이터에 레이블이라는 답을 포함하고 있는 데이터 세트

4. **가장 널리 사용되는 지도 학습 작업 두 가지는 무엇인가요?**

   1) 분류: 특성을 사용하여 데이터를 분류하는 문제

   2) 회귀: 특성을 사용하여 타깃 수치를 예측하는 문제

5. **보편적인 비지도 학습 작업 네 가지는 무엇인가요?**

   1) 군집(clustering): 데이터를 비슷한 특징을 가진 몇 개의 그룹으로 나누는 것

   2) 시각화(visualization)와 차원 축소(dimensionality reduction): 다차원 특성 데이터셋 -> 2D or 3D, 상관관계 여러 특성을 하나로 합쳐 데이터 특성 수 줄이기, 시각화를 위해선 데이터 특성 2가지로 축소

   3) 이상치 탐지(anomaly detection)와 특이치 탐지(novelty detection): 이상치 탐지는 정상 샘플을 이용하여 훈련 후 입력 샘플의 정상여부 판단, 특이치 탐지는 clean 훈련 세트 활용, 훈련 데이터와 다른 데이터 감지

   4) 연관 규칙 학습(association rule learning): 데이터 간의 흥미로운 관계 찾기

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

   강화 학습: 에이전트가 환경을 관찰하여 행동을 실행하고 그 결과로 보상 혹은 벌점, 시간이 지나면서 가장 큰 보상을 얻기 위해 정책(주어진 상황에서 에이전트가 어떤 행동을 선택해야 할지 정의)이라고 부르는 최상의 전략 학습

7. **고객을 여러 그룹으로 분할하려면 어떤 알고리즘을 사용해야 하나요?**

   준지도 학습: 적은 수의 샘플에 레이블을 적용, 비지도 학습을 통해 군집을 분류한 후 샘플들을 활용해 지도 학습에 활용

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

   지도 학습 중 분류: 특성을 사용하여 데이터를 분류하는 문제

9. **온라인 학습 시스템이 무엇인가요?**

   온라인 학습(online learning): 적은 양의 데이터(미니배치, mini-batch)를 사용해 점진적으로 훈련, 나쁜 데이터가 주입되는 경우 시스템 성능이 점진적으로 하락, 지속적인 시스템 모니터링 필요

10. **외부 메모리 학습이 무엇인가요?**

    외부 메모리 학습(out-of-core learning): 빅데이터 분석 시 데이터 양이 지나치게 커서 컴퓨터의 메모리로 감당되지 않는 경우 사용되는 온라인 학습 알고리즘, 데이터 일부를 읽어 들여 머신러닝 알고리즘이 학습 후 전체 데이터가 모두 적용될 때까지 일부를 학습하는 과정 반복

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

    사례 기반 학습: 샘플을 기억하는 것이 훈련의 전부, 예측을 위해 기존 샘플과의 유사도 측정

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

    모델 파라미터: 데이터를 통해 모델이 직접 학습하는 값

    학습 알고리즘의 하이퍼파라미터: 알고리즘 학습 모델을 정의하는데 사용되는 파라미터로 훈련 과정에 변하는 파라미터가 아님, 하이퍼파라미터를 조절하면서 가장 좋은 성능의 모델 선정

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

    모델 기반 학습: 모델을 미리 지정한 후 훈련 세트를 사용하여 모델을 훈련, 훈련된 모델을 사용해 새로운 데이터에 대한 예측 실행

모델 기반 알고리즘은 성공을 위해 학습 알고리즘이 비용 함수를 최소화하는 모델 파라미터를 찾은 후 새로운 데이터를 모델에 적용해 예측

14. **머신러닝의 주요 도전 과제는 무엇인가요?**

    1) 충분하지 않은 양의 훈련 데이터
    2) 대표성 없는 훈련 데이터: 샘플링 잡음, 샘플링 편향
    3) 낮은 품질의 데이터: 이상치 샘플, 특성 누락
    4) 관련이 없는 특성: 특성 공학(특성 선택, 특성 추출)
    5) 훈련 데이터 과대적합: 훈련 세트에 특화되어 일반화 성능이 하락
    ~~6) 훈련 데이터 과소적합: 모델 단순해서 훈련 세트 학습 X~~

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

    훈련 데이터 과대적합 문제

    1) 검증: 훈련 세트(80%), 테스트 세트(20%)로 구분하여 모델 훈련에 훈련 세트, 모델 성능평가에 테스트 세트 활용
    2) 하이퍼파라미터 튜닝과 모델 선택
    3) 교차 검증: 여러 개의 검증 세트를 사용한 반복적인 예비표본 검증 적용 기법

16. **테스트 세트가 무엇이고 왜 사용해야 하나요?**

    훈련된 모델의 성능 평가에 이용되는 데이터 세트, 일반화 오차를 측정

17. **검증 세트의 목적은 무엇인가요?**

    검증 세트: 훈련 세트의 일부로 만들어진 데이터셋, 다양한 하이퍼파라미터 값을 후보 모델 평가용으로 예비표본을 검증세트로 활용하는 기법

18. **테스트 세트를 사용해 하이퍼파라미터를 튜닝하면 어떤 문제가 생기나요?**

    모델이 테스트 세트를 훈련해 일반화 오차를 측정하기 어려워지는 문제 발생