<머신러닝 톺아보기> 1주차 과제

박민서

1. 머신러닝이란 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구분야이다.
2. 기존 솔루션으로 많은 수동 조정과 규칙이 필요한 문제, 전통적인 방식으로 해결 방법이 없는 복잡한 문제, 새로운 데이터에 적응할 수 있는 머신러닝 시스템의 특징을 활용할 수 있는 유동적인 환경, 복잡한 문제와 대량의 데이터에서 통찰을 얻어야 하는 분야 등에서 머신러닝은 두각을 나타낸다.
3. 레이블된 훈련 세트란 각 샘플에 대해 원하는 정답을 담고 있는 훈련 세트다.
4. 가장 널리 활용되는 대표적인 지도 학습으로는 분류와 회귀가 있다. 분류는 특성을 사용하여 데이터를 분류하는 문제이고, 회귀는 특성을 사용해 타깃 수치를 예측하는 문제를 의미한다.
5. 대표적인 비지도 학습으로는 군집, 시각화와 차원 축소, 이상치 탐지와 특이치 탐지, 연관 규칙 학습 등이 있다. 군집은 데이터를 비슷한 특징을 가진 몇 개의 그룹으로 나누는 과정이다. 시각화와 차원 축소는 다차원 특성을 가진 데이터셋을 2D 또는 3D로 표현하는 과정이다. 이상치 탐지와 특이치 탐지는 각각 정상 샘플을 이용하여 훈련 후 입력 샘플의 정상여부를 판단하는 과정, 전혀 오염되지 않은 훈련 세트를 활용하여 훈련 세트에 포함된 데이터와 다른 데이터를 감지하는 과정이다. 연관 규칙 학습은 데이터 간의 흥미로운 관계를 찾는 것을 말한다.
6. 사전 정보가 없는 여러 지형에서 로봇을 걸어가게 하려면 강화 학습을 사용할 수 있다. 강화 학습이란 에이전트가 환경을 관찰하여 행동을 실행하고 그 결과로 보상 혹은 벌점을 부여하는 시스템을 말하는데, 이 과정에서 가장 큰 보상을 얻기 위해 정책이라고 부르는 최상의 전략을 학습하게 된다. 보행로봇은 이 과정을 거쳐 에이전트가 원하는 결과인 사전 정보가 없는 여러 지형에서 걸어갈 수 있는 능력을 갖추게 될 것이다.
7. 고객을 여러 그룹으로 분할하려면 데이터를 비슷한 특징을 가진 몇 개의 그룹으로 나누는 군집 방법을 이용할 수 있다.
8. 스팸 감지는 대표적인 지도 학습의 문제라고 볼 수 있다. 이 경우 지도학습 중 분류를 사용하는데 소속 정보, 특정 단어 포함 여부 등 특성을 사용하여 타깃인 스팸을 잡아내는 방식을 사용한다.
9. 온라인 학습 시스템이란 적은 양의 테이터(미니배치)를 사용해 점진적으로 훈련시키는 방법을 의미한다. 다만, 나쁜 데이터가 주입되는 경우 시스템 성능이 점진적으

로 떨어질 수 있기에 지속적인 시스템 모니터링이 필요하다.

10. 외부 메모리 학습은 메인 메모리에 들어갈 수 없는 아주 큰 데이터셋을 학습하는 시스템에서 사용될 수 있다. 알고리즘이 데이터 일부를 읽어 들이고 훈련 단계를 수행하는데 전체 데이터에 모두 적용될 때까지 이 과정을 반복한다.
11. 예측을 하기 위해 유사도 측정에 의존하는 학습 알고리즘은 사례 기반 학습이다. 사례 기반 학습은 새로운 샘플이 주어지면 유사도 측정을 사용해 학습된 샘플 중 가장 비슷한 것을 찾아 예측으로 사용하는 방법입니다.
12. 모델 파라미터는 모델을 사용하기 전에 정의해야 하는 파라미터를 말하며, 모델이 최상의 성능을 내도록 하는 값을 말한다. 모델 훈련은 훈련 데이터에 가장 잘 맞는 모델 파라미터를 찾기 위해 알고리즘을 실행하는 것을 의미한다. 학습 알고리즘의 하이퍼파라미터는 학습 알고리즘으로부터 영향을 받지 않으며, 훈련 전에 미리 지정되고, 훈련하는 동안에는 상수로 남아있는 것을 말한다.
13. 모델 기반 학습은 모델을 미리 지정한 후 훈련세트를 사용하여 모델을 훈련시키는 방식을 말한다. 이후 훈련된 모델을 사용해 새로운 데이터에 대한 예측을 실행한다. 성공을 위해 모델 기반 알고리즘이 사용하는 일반적인 전략은 선형회귀 알고리즘이다. 훈련 데이터에서 시스템의 예측이 얼마나 나쁜지 측정하고 이러한 비용 함수를 최소화하는 방향으로 시스템을 훈련시킨다. 예측을 만들려면 학습 알고리즘이 찾은 파라미터를 사용하는 모델의 예측 함수에 새로운 샘플의 특성을 주입시킨다.
14. 머신러닝의 주요 도전 과제는 다음과 같다. 첫째, 충분하지 않은 양의 훈련 데이터이다. 간단한 문제라도 수천 개의 데이터가 필요하다. 데이터가 부족하면 알고리즘 성능 향상이 어려울 수 있다. 둘째, 대표성 없는 훈련 데이터이다. 샘플링 잡음은 우연에 의해 대표성이 없는 데이터이고, 샘플링 편향은 표본 추출 방법이 잘못된 대표성이 없는 데이터를 말한다. 셋째, 낮은 품질의 데이터이다. 이 경우, 이상치 샘플이라면 고치거나 무시하고 만약 특성이 누락되었다면 해당 특성을 제외, 해당 샘플을 제외, 누락된 값을 채우기, 해당 특성을 넣은 경우와 뺀 경우 각기 모델을 훈련시켜 해결할 수 있다. 넷째, 관련이 없는 특성이다. 이는 특성 공학에서 해결할 수 있는데 풀려는 문제에 관련이 높은 특성을 찾는다. 대표적으로 준비되어 있는 특성 중 가장 유용한 특성을 찾는 특성 선택과 특성을 조합하여 새로운 특성을 만드는 특성 추출을 들 수 있다. 다섯째, 훈련 데이터 과대적합을 들 수 있다. 과대적합은 훈련 세트에 특화되어 일반화 성능이 떨어지는 현상을 의미하는데, 이는 여러 규제를 적용해 해결할 수 있다. 여섯째, 훈련 데이터 과소적합을 들 수 있다. 모델이 너무 단순해서 훈련 세트를 잘 학습하지 못하는 경우를 말한다. 이 경우 모델 파라미터가 더 많은 복잡한 모델을 사용하거나 특성 공학으로 더 좋은 특성을 찾기, 규제의 강도를 줄이기 등의 방법을 사용할 수 있다.
15. 모델이 훈련 데이터에서의 성능은 좋지만 새로운 샘플에서의 일반화 성능이 나쁘

다면 모델은 훈련 데이터 과대적합 현상이 나타났다고 볼 수 있다. 이 상황은 모델을 지나치게 복습하게 학습하여 훈련 데이터 셋에서는 모델 성능이 높지만 다른 데이터가 주어졌을 때에는 정확한 예측/분류를 못하는 경우이다. 이 경우 해결 방법으로는 크게 세 가지가 있다. 첫째, 데이터 양을 늘리면 된다. 과대적합의 경우 데이터 양이 적어서 해당 데이터의 특징패턴, 노이즈까지 학습해버리는 경우가 있다. 이 경우 데이터 양을 늘려야 모델은 일반적인 패턴을 학습하여 과대적합을 방지할 수 있다. 두 번째, 모델의 복잡도를 줄이는 방법이 있다. 이는 정규화 기법이라고도 하는데 대표적으로 L1 정규화와 L2 정규화가 있다. L1 정규화는 가중치의 절댓값 합에 패널티를 주고, L2 정규화는 가중치의 제곱합에 패널티를 주는 방식으로 둘 다 과도한 가중치를 억제하여 모델이 데이터에 지나치게 적응하지 않도록 하는 역할을 한다. 마지막으로, 교차 검증은 데이터를 여러 개의 부분으로 나누어 각각을 검증 세트로 사용하며, 나머지를 훈련세트로 사용하는 기법이다.

16. 훈련된 모델의 성능 평가는 테스트 세트를 활용하여 이루어진다. 흔히 전체 데이터 셋을 훈련 세트(80%)와 테스트 세트(20%)로 구분하는데 훈련 세트는 모델을 훈련시키는데 사용되고, 테스트 세트는 모델을 테스트하는데 사용된다.
17. 검증세트는 훈련세트의 일부로 만들어진 데이터셋으로 다양한 하이퍼파라미터 값을 후보 모델 평가용으로 예비표본을 검증세트로 활용하는 기법에서 사용된다.
18. 테스트 세트를 사용해 하이퍼파라미터를 테스트 세트에 과대적합이 될 위험이 있고 오차를 낙관적으로 측정하게 될 것이다.