1주차 예비보고서

 전공: 컴퓨터공학과 학년: 2학년 학번: 20221519 이름: 구현모

1.

1-1

Machine Learning은 인공지능의 한 분야로 데이터를 학습하고 학습한 데이터를 바탕으로 패턴을 인식, 예측하는 기술이다. Machine Learning은 크게 Supervised Learning, Unsupervised Learning, Reinforcement, 3가지로 나누어진다. 이 중 Supervised Learning과 Unsupervised Learning에 대해 알아보자.

먼저 Supervised Learning은 input vector와 target value를 함께 사용해서 학습하는 방식이다. 정답을 기반으로 패턴을 학습하여 새로운 데이터의 정답을 예측하는 것이다. 장점에는 정확도가 높다는 것이 있다. 단점으로는 많은 데이터에 정답을 지정하는 데이터 라벨링의 과정의 비용이 많이 든다는 것이 있다. 분류와 회귀의 방식을 사용한다.

Unsupervised Learning은 input vector만을 사용하여 학습하는 방식이다. 학습된 input vector들의 공통점을 식별하여 입력된 데이터들의 패턴을 찾거나 군집화 시키는 것이다. 장점으로는 데이터 라벨링 비용이 들지 않는다는 점과 사람이 발견하지 못한 새로운 패턴들을 찾아낼 수 있다는 점이다. 단점으로는 정답이 제공되지 않기 때문에 정확도가 낮을 수 있다는 점이 있다. 군집화, 연관 규칙 학습 등의 방법으로 사용한다.

Supervised Learning, Unsupervised Learning의 가장 큰 차이점은 machine이 학습하는데 제공되는 데이터에 target value, 즉 정답이 함께 제공되는지의 여부이다. 이로 인하여 정답이 제공되는 Supervised Learning은 주어진 정답을 학습하여 새로운 데이터 예측하는 것을 목적으로 사용되고 Unsupervised Learning은 데이터 사이의 관계나 구조 등을 찾는 것을 목적으로 사용된다는 차이점이 생긴다.

1-2.

머신러닝 모델 학습 과정은 먼저 데이터를 수집하고 다음으로 수집된 데이터를 모델이 학습할 수 있도록 처리해주는 데이터 전처리를 해준다. 그 다음으로 모델이 제공된 데이터를 학습한다. 이후 테스트케이스를 이용하여 모델을 평가하는 것이다. 이때 데이터는 텍스트, 이미지, 센서 데이터 등 여러가지 형태일 수 있다.

이 중 모델 학습은 선형 회귀, 분류, 신경망 등의 알고리즘을 선택하여 모델이 입력 데이터와 결과 사이의 관계를 학습한다. 입력된 데이터로 예측 값을 만들고, 손실함수로 오차를 계산하며 이를 최적화하는 작업을 반복한다. 이 과정을 통해 데이터의 패턴을 점차 학습하여 학습하지 않은 새로운 데이터에도 적용할 수 있는 일반화된 모델을 만드는 것이다.

이러한 학습과정에서 과적합에 주의해야 한다. 또한 학습에 사용되는 데이터가 편향적이거나 잘못된 데이터라면 학습된 모델이 틀린 답을 내놓을 수 있다.

2.

 NumPy는 Numeric Python의 약자로 대규모 다차원 배열과 행렬 연산에 필요한 다양한 함수와 메소드를 제공하는 라이브러리이다. 과학 계산, 데이터 분석, 머신 러닝 등에서 빠르고 효율적인 수학적 연산을 수행하기 위해 만들어 졌다. Python의 기본 list보다 빠르게 연산할 수 있기 때문에 많은 양의 데이터를 다룰 때 이용한다.

NumPy의 특징으로는 다차원 배열, 메모리 효율성, 빠른 연산, 브로드캐스팅, 함수 제공이 있다. 다차원 배열은 같은 데이터 타입을 가진 요소들로 이루어진 다차원의 배열을 만들 수 있으면 알아서 벡터화 연산을 지원하기 때문에 따로 바꾸거나 반복문을 사용하는 등의 작업을 하지 않아도 빠르게 연산이 가능하게 해준다. 또한 같은 타입의 데이터를 사용하기 때문에 메모리 효율성이 높다. 내부 구조를 C언어로 구현하였기 때문에 python에 비하여 더 빠른 연산을 할 수 있다. 브로드캐스팅은 일정 조건을 만족할 때 서로 다른 크기를 가진 배열끼리의 연산을 가능하게 한다. 이를 통해 복잡한 코드를 간단하게 나타낼 수 있다. 통계, 선형대수 등의 여러가지 수학적 기능을 수행하는 함수들을 제공하여 사용자가 사용할 수 있도록 한다.

Pandas은 데이터를 조작하고 분석하는 것을 위해 만들어진 라이브러리이다. 엑셀이나 SQL과 비슷한 방법으로 데이터를 처리할 수 있도록 만들어졌다. 1차원, 2차원 데이터 구조를 지원한다.

Pandas의 특징으로는 데이터 조작, 입출력 지원, 인덱싱과 슬라이싱 기능, NumPy기반이라는 점이 있다. 데이터의 병합, 그룹화, 결측값 처리 등을 할 수 있도록 다양한 함수를 자원한다. Excel, SQL, JSON, CSV 등의 서로 다른 파일 형식에서 데이터를 입력 받을 수 있다. 강력한 인덱싱 능력을 가지고 있기 때문에 데이터를 쉽게 찾고 가공할 수 있으며 필터링이 쉽다. NumPy를 내부적으로 사용하기 때문에 NumPy의 장점인 빠른 수치연산이 가능하고 호완성이 뛰어나다.

Matplotlib는 2D 그래프와 플롯을 만드는 등 데이터를 시각적으로 표현해주는 것을 위해 만들어진 라이브러리이다. 데이터를 시각화 해주기 때문에 데이터에 대한 직관적인 이해를 가능하게 해주기 때문에 과학 연구, 데이터 분석에 주로 사용된다.

Matplotlib의 특징으로는 다양한 플롯의 그래프 지원, 커스터마이징, 객체 지향, Pandas와의 연동, 출력 옵션 제공이 있다. Matplotlib은 선 그래프, 막대 그래프, 산점도, 히스토그램, 파이 차트 등의 여러 그래프 방식을 지원하여 상황에 맞는 그래프를 골라 사용할 수 있다. 또한 색, 그래프의 여러 표시 등의 스타일을 세세하게 조정 가능하며 축 설정 등도 원하는 대로 조정할 수 있다. 간단한 경우 Matplotlib.pylot.plot()와 같은 함수를 이용할 수 있으며 세밀한 제어가 필요할 시 객체 지향 방식도 사용할 수 있도록 지원한다. Pandas DataFrame과 연동하여 데이터를 시각적으로 표현할 수 있어 호환성이 좋다. 데이터를 시각화하여 그래프를 화면에 표시할 수 있으며 PNG, PDF등의 파일 형태로 저장할 수 있다.

간단하게 정리하면 NumPy는 다차원 배열의 계산을 빠르고 효율적으로 가능하게 하고, Pandas는 데이터를 쉽게 조작하고 분석할 수 있도록 하며, Matplotlib는 데이터를 시각화해주는 기능을 한다.

3.

scikit-lear은 다양한 머신러닝 알고리즘을 지원하는 라이브러리로 약칭은 sklean이다. Supervised Learning과 Unsupervised Learning을 포함한 머신러닝 작업을 지원한다. 그 특징으로는 간단한 API(fit(), predict() 등)로 누구나 쉽게 사용할 수 있도록 되어 있다는 점, NumPy, Pandas DataFrame과 호환된다는 점이 있다. 또한 데이터 전처리, 모델평가 도구가 내장되어 있다는 점, 모델을 최적화하기 위해서 조정해야 하는 변수인 하이퍼파라미터를 조정해주는 도구가 내장되어 있다는 점이 있다. 마지막으로 오픈소스 라이브러리로 커뮤니티 지원으로 많은 정보들을 찾아볼 수 있다.

scikit-learn에서 제공하는 모델 중 대표적인 것은 선형 회귀, 의사결정나무, K-평균 군집화를 들 수 있다.

선형회귀(linear Regression)는 연속적인 target value를 예측하는 회귀 문제에 사용되면 input vector과 target value 사이의 선형 관계를 모델링 한다. 사용법은 먼저 그림1-1)과 같이 import 해준다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.그림1-1)

입력 데이터 X와 목표 변수 Y를 준비한다. 이때 X는 2차원 배열, Y는 1차원 배열로 한다. 그후 그림1-2)와 같이 모델을 생성하고 fit() 메서드를 이용하여 데이터를 학습시킨다. 텍스트, 폰트, 스크린샷, 타이포그래피이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.그림1-2)

다음으로 새로운 데이터에 대하여 predict() 메서드를 사용하여 예측값을 얻고 모델을 평가 한다. 이러하 과정으로 linear Regression 모델을 사용할 수 있다.

의사결정나무(Decision Tree)는 분류 문제에 사용되며 데이터를 기반으로 분류 규칙 학습하여 클래스 레이블을 예측한다.

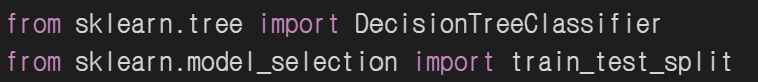
그림2-1)

그림2-1)과 같이 import해준다. Train\_test\_split()은 데이터 배열을 학습을 위한 데이터와과 테스트를 위한 데이터로 랜덤 분할시키는 동작을 한다.

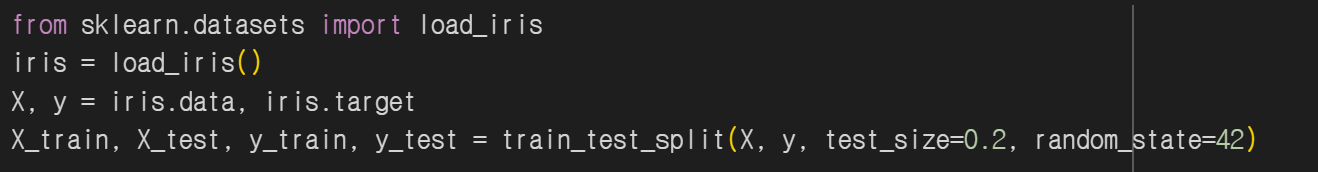
그림2-2)

그림2-2)는 sklean에서 제공하는 자료인 load\_iris를 import하여 데이터를 준비한 모습이다. 이와 같이 model에 학습시킬 데이터를 준비한다.

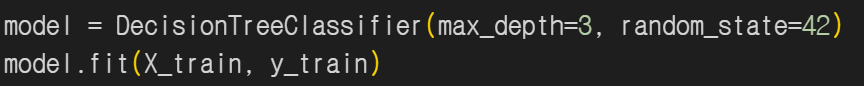
그림2-3)

그림2-3)과 같이 DecisionTreeClassifier()을 이용하여 의사결정나무 모델을 생성하고 fit()로 학습시킨다. 이후 linear Regression 방식과 같이 그림2-4)처럼 predict()를 이용하여 테스트 테스트 데이터를 이용하여 예측을 수행한다.

텍스트, 폰트, 그래픽, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.그림2-4)

이후 정확도를 확인하고 시각화 하는 등 모델을 평가하는 방식으로 Decision Tree 모델을 사용한다.

K-평균 군집화(K-Means Clustering)은 지도학습(Supervised Learning) 알고리즘인 앞의 두가지 방식과 달리 비지도학습(Unsupervised Learning) 알고리즘이다. 비슷한 특성을 가진 데이터들끼리 집단을 만들어 묶는데 사용된다.

그림3-1)과 같이 KMeans를 import하여 모델을 생성하고 준비된 데이터를 학습시킨다. 모델을 학습시킬 때 몇 개의 그룹(n\_clusters)으로 나눌 것인지를 미리 정해주어야 한다. 그림3-1)에서는 2개의 그룹으로 나누었다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다. 그림3-1)

학습된 모델을 이용하여 군집화가 잘 되었는지 cluster label을 확인한다. 이후 Matplotlib 라이브러리 등을 이용하여 결과를 시각화 하여 cluster이 잘 나누어 졌는지를 확인한다. 이러한 과정을 통하여 K-Means Clustering을 사용할 수 있다.

각 3가지 알고리즘을 정리하면 선형 회귀는 연속형 데이터를 예측하는 선형모델로 이루어져 있으며 회귀 문제에 알맞다. 의사결정나무는 분류 문제에 활용되며 분류의 규칙을 제공한다. K-평균 군집화는 Unsupervised Learning방식으로 데이터들을 같은 특성을 가진 집단으로 묶어줄 수 있다.

3가지 알고리즘 모두 scikit-lear 라이브러리에 있는 필요한 알고리즘을 import해준 뒤 모델을 생성한다. 데이터 준비하여 모델을 학습시키고 잘 학습되었는지 확인하는 비슷한 과정을 통해 사용된다.