출석수업 과제물(평가결과물) 표지(온라인제출용)

**교과목명 : 머신러닝**

**학 번 : 202034-153746**

**성 명 : 이동열**

**강 의 실 : 서울남부학습센터 온라인(ZOOM8, ZOOM9, ZOOM19)**

**연 락 처 : 010-5264-5565**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1 – (1)

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

1 - (2)

지도학습은 모델 학습에 입력 데이터와 출력 데이터를 한 쌍으로 제공해 두 데이터간의 관계를 학습하는 방식이다. 전체적인 흐름은 데이터 세트를 통해 모델을 학습시키고 충분한 학습이 되었는지 중간에 모델을 테스트 한다. 이때, 모델이 오답을 내는 경우 추가적인 데이터 세트를 제공해 지속적으로 학습을 시키면서 개선해나간다. 지도학습의 지도는 Supervisor를 의미하는데 말 그대로 출력값을 제공함으로써 학습의 방향을 알려주고 이끌어준다는 의미를 가지고 있다. 이러한 지도학습 방식은 분류 및 회귀 문제를 해결하는데 적합한다. 예를 들어 스팸 메일을 분류한다던가, 이미지 속에 객체를 구분한다던가, 학력 및 경력에 따른 임금 예측 등이 있다.

지도학습이 출력 데이터를 통해 학습의 방향성을 제시한 방식이라면 비지도 학습은 출력 데이터 없이 데이터의 구조 및 패턴을 찾아가는 학습 방법이다. 비지도 학습은 데이터의 특징을 분석하고 그룹화하는 과정을 통해 새롭게 특징을 추출해낸다. 이러한 비지 학습 방식은 군집화 문제 해결에 적합하다.

지도학습의 경우 데이터를 쌍으로 제공해야 하므로 데이터 수집에 큰 비용이 발생한다. 때문에 지도 학습 방식과 비지도 학습 방식을 함께 사용하는 학습 방식이 있는데 이를 준지도 학습이라고 한다. 준지도 학습은 일부 데이터에만 출력 데이터가 제공되고 나머지 데이터는 출력 데이터가 없이 제공된다. 이는 데이터가 부족할 때 유용하게 사용된다.

1 – (3)

학습 모델이 학습 데이터에는 높은 성능을 보이지만 새로운 데이터에 대해서는 낮은 성능을 보일 때 이를 과다적합되었다고 한다. 모델을 학습할 때 학습에 사용된 현재 데이터와 이후 데이터들은 같은 모집단에서 추출한 표본으로 볼 수 있기 때문에 데이터의 오차값이 크지 않을 것으로 예상할 수 있다. 때문에 이러한 학습 데이터에 모델을 과하게 최적화 시켜버리면 학습 데이터에 대해서는 높은 성능을 보이지만 새로운 데이터에 대해서는 낮은 성능을 보이게된다.

과다적합이 발생하는 이유는 학습 데이터가 적은 상태에서 제한된 데이터를 너무 오래 학습하거나 모델 자체가 너무 복잡해 데이터 세트 내에서 잡음이 발생하고 관련없는 정보를 학습하는 등의 이유가 있다.

과다적합을 해결하려면 여러 방법이 있는데, 먼저 모델이 잡음을 학습하기 전에 학습을 조기 종료시키는 방법이 있다. 이 방법은 너무 일찍 조기 종료시킬 경우 오히려 과소적합이 발생할 수 있기 때문에 최적점을 잘 찾아야한다. 그리고 더 많은 데이터를 학습시켜서 모델의 정확도를 높이는 방법도 있다. 이때 학습 데이터는 관련성이 높고 정제된 데이터가 효과적인데 그렇지 않은 데이터를 사용할 경우 모델의 복잡성이 증가해 과적합이 발생할 수 있다. 만약 모델이 너무 복잡해서 과적합이 발생했다면 모델에서 중복되거나 관련성이 없는 데이터를 제거할 수 있다.

1 – (4)

데이터들을 비슷한 특성을 가진 것들끼리 나누어서 그룹화하는 것을 군집화라고 한다. 데이터 그룹을 나누는 기준은 유사도와 거리가 있는데 유사도가 높고 거리값이 적을 수록 비슷한 특성을 가지고 있다고 판단해 그룹으로 묶는다. 이는 분류와 비슷해보이지만 분류의 경우 출력값을 같이 제공해 미리 그룹화가 가능한 지도학습이고 군집화는 데이터를 분석해서 특성을 파악해 그룹화하는 비지도 학습이라는 차이가 있다.

군집화의 종류는 계층적 군집과 분할적 군집으로 나뉘는데, 계층적 군집은 군집을 계층으로 분리해 군집의 데이터 구조나 군집의 종류를 확인할 수 있고 분할적 군집은 각 데이터를 하나의 군집에만 포함시켜 여러 군집을 만들어내 데이터의 패턴이나 구조를 파악할 수 있다.

이렇게 군집화를 하는 이유는 복잡한 대량의 데이터의 패턴이나 구조를 이해하고 이를 요약 및 시각화하여 해석하기 쉽게 처리하기 위함인데, 이렇게 군집화를 하게 되며 데이터 군집들 사이에서 이상치를 탐지하거나 데이터 특성과 관계를 발견할 수 있게 된다.

실제로 군집화를 활용하고 있는 분야를 보면 고객 데이터를 기반으로 고객을 분류해 관리하는 마케팅 분야나 맞춤 정보를 제공하기 위한 개인화 서비스, 불량품 구분을 위한 이상치 탐지와 같은 사례가 있다.

2 - (1)

베이즈 분류기를 사용해 판별함수 값을 구하면 가 속하게 될 클래스를 판별할 수 있다.

및 는 아래 베이즈 정리 수식으로 구할 수 있다.

=C 일 때, 사전 확률을 구하면 다음과 같다. 사전확률은 사건 에 대한 발생확률을 구하면 된다. 예를 들어 총 8개의 사건중 사건 가 3개라면 해당 사건이 발생할 확률은 가 된다. 이렇게 도 구하면 다음과 같이 된다.

위 값을 베이즈 정리에 대입하면 아래와 같이 값을 도출해낼 수 있다.

사전 확률을 구했다면 사후확률을 구해야한다. 사후확률 는 를 선택했을 때 사건 가 발생할 확률로, 클래스가 총 4개이고 클래스에 포함되어 있는 C가 1개 이므로 가 된다. 같은 방식으로 를 구하면 가 된다.

즉, 이므로 일 때, 가 된다.

2 - (2)

K-최근접이웃 분류기를 사용하면 주어진 점으로부터 가장 가까운 K개의 데이터를 선택해 빈도수가 많은 클래스를 찾아 주어진 점을 해당 클래스에 할당할 수 있다.

먼저, 주어진 점과 각 데이터의 거리를 구하면 다음과 같다.

(3,2) -> (2,1)

(3,2) -> (1,3)

(3,2) -> (0,-1)

(3,2) -> (3,3)

(3,2) -> (-1,1)

(3,2) -> (0,0)

(3,2) -> (2,5)

(3,2) -> (-1,2)

(3,2) -> (1,5)

(3,2) -> (4,5)

이때, K=5이므로 아래와 같이 가장 거리가 짧은 5개의 데이터를 선별한다. 하지만 거리가 동일한 데이터 (0,0), (2,5), (4,5)로 인해 6개의 데이터가 선택되므로 K를 7로 늘려 추가 선별한다.

B(3,3), A(2,1), B(1,3), A(0,0), A(2,5), B(4,5), A(1,5)

선별된 데이터 중 클래스A에 속한 데이터가 더 많으므로 주어진 점(3,2)은 클래스A에 속하게 된다.