Data Science

[Data Science] 데이터 사이언스 개념 - 2.머신러닝의 기본

2021. 1. 8. 14:38 수정 삭제 공개

2.머신러닝의 기본

1.머신러닝이란

컴퓨터 과학에서 컴퓨터를 적극적으로 이용한 통계학을 의미
머신러닝에서는 문제 설정을 학습 시나리오라고 부른다
학습 시나리오에는 크게 지도학습과 비지도 학습이 있다.

지도학습

- 입력 데이터와 출력 데이터가 세트로 되어 있는 데이터를 다루는것
ex)주택 가격을 특징량을 이용해 휘귀모델로 예측하는 문제, 회귀, 분류, 랭킹

비지도 학습

- 입력 데이터만 주어진 상황
ex) 클러스터링, 차원삭감, 행렬보완, 다양체학습

준지도 학습

- 양쪽 측면을 포함, 일부 데이터에는 출력 데이터가 있지만 나머지는 출력 데이터가 없는 상황
ex) sns 등에서 수집된 우호 관계 네트워크와 일부 성별이 판명된 데이터 세트가 있는 경우

2.지도 학습

기본적인 회귀 문제에서 살펴보면
주택가격을 예측하고 싶다고 할때 우리에게 주어진 특정량(설명변수)를 통해 주택가격을 예측하는 것을 목표로 한다.
이걸 수학적으로 풀어낸다면 아래와 같은 수식이 생긴다.

Y = f(X) + 오차항
지도 학습의 목적은 이 f를 추정하는 것이다.

f를 추정하는 이유는 두가지이다.
첫번째로는 f를 추정함으로써 **예측을 할 수 있게 되기 때문**이다. 주어진 데이터(학습 데이터)를 이용해 추정한 추정함수를 바탕으로 주택가격에 대해 예측을 할 수 있게 된다.
두번째로는 **데이터 해석**이다. 만약 f가 선형회귀 처럼 아주 이해하기 쉬운 모델이라면 상관관계를 분석하기도 쉽다.
이런식의 상관관계를 검출하는데 사용할 수 있다.

3.훈련오차 / 테스트 오차

머신러닝에서 중요해지는 것이 일반화 성능이다.
추정 시에는 사용하지 않았던 데이터로 측정한 성능이 일반화 성능이다.
즉 학습을 하는데 사용하지 않았던 데이터로 모델의 성능을 확인하는 것이 일반화 성능이다.
그래서 머신러닝에서는 데이터를 학습 데이터와 시험 데이터로 구별하는 것이 일반적이다.
보통은 데이터를 가지고 랜덤하게 원하는 비율로 나누어 사용한다.

4.모수적 모델과 비 모수적 모델

모수적 모델

- 수식을 이용해, 명시적으로 함수를 정의한 모델
ex) 선형회귀 모델

장점
비모수적 모델보다 안정적으로 적합하는데 필요한 데이터양이 비교적 적다.
모델을 추정하기가 쉽다.
해석가능성이 높다(과학 이론 등에 기반해서 모델을 정한 경우는 계수 자체에 핵석할 수 있는 의미가 부여되는 일이 많다)
단점
모델 가정이 나쁘면, 체계적으로 예측을 벗어나게 된다.

비모수적 모델

- 함수형에 대해 명시적인 가정을 두지 않는 모델

장점
데이터에 맞추는 형태로 모델을 구성하므로 실제 모델에 가까울 가능성이 높다는 보증이 있다.
단점
모수적 모델보다 안정적으로 모델을 추정하는 데 필요한 데이터양이 많아진다.
해석 가능성이 희생되기도 한다.
모델 추정이 어려운 경우가 많다.

5.추정법

예측이 목적이라면 시험 데이터의 손실(모델 예측과 데이터 사이의 괴리)을 최소화 하는 모델이 가장 좋은 모델이 된다.
평균제곱 오차(MSE)

중요한 점 2가지

1.학습 데이터와 시험 데이터가 같은 성질을 가져야 한다. (야구 선수 연봉 예측하는데 축구 선수 연봉으로 학습하는 경우)
2.과적합이 발생되지 않게 해야 한다.

'Data Science' 카테고리의 다른 글□

[Data Science] 데이터 사이언스 개념 - 6.분류문제□

[Data Science] 데이터 사이언스 개념 - 5.앙상블 학습□

[Data Science] 데이터 사이언스 개념 - 4.회귀 모델□

[Data Science] 데이터 사이언스 개념 - 3.과적합과 모델 선택□

[Data Science] 데이터 사이언스 개념 - 2.머신러닝의 기본□

[Data Science] 데이터 사이언스 개념 - 1.데이터 과학이란?□

머신러닝 모수적 모델 비모수적 모델 지도학습 추정법

나아무늘보

혼자 끄적끄적하는 블로그 입니다.