ML\_Test 1차

KDT 6기 이동건

1. 기계학습 프로세스에 대해 간략하게 설명하세요

* 기계학습이란 기존의 데이터들을 이용하여 학습을 시켜 미래에 대해 예측하는 것을 말한다
* 데이터 수집 – 데이터 전처리 – EDA - 모델 생성 – 모델학습 – 모델성능평가

1. 기계학습에서 사용되는 데이터셋에 대해 설명하세요.

* 학습용데이터, 검증용데이터, 테스트용 데이터로 나누고

학습용데이터는 모델을 학습시킬 때 사용하는 데이터셋

검증용데이터는 모델의 학습이 잘되었는지 확인하기위한 데이터셋

테스트용 데이터는 실제로 모델을 테스트할 때 사용하는 데이터셋이다.

* 범주형 데이터는 주로 분류, 수치형 데이터는 주로 회귀에서 사용한다.

1. 지도학습의 특징을 설명하세요.

* 지도학습은 정답이 있는 정답지를 학습시키는 방식이다.

그래서 문제와 결과를 함께 준다음 문제만 주었을 때 결과를 예측하는 방식이

지도 학습이라고 할 수있다.

* 종류로는 회귀, 분류가 있다.

1. 데이터 전처리 중 처음 데이터 수집 후 데이터 정제하는 과정에 진행되는 작업을 설명하세요.
2. 파일을 열어 기본적인 데이터의 형태를 파악
3. 데이터 타입확인 - object라면 인코딩을 통해 숫자로 변환
4. 데이터의 결측치 확인 – 결측치를 평균,최대,최소,최빈,중간값등으로 채우거나 삭제
5. 데이터의 이상치 확인 – 박스플롯이나, 4분위수를 이용해 이상치를 확인하고 데이터에 큰 오류가 생길 것 같은 이상치들은 삭제
6. 시각화를 통해 데이터의 분포를 파악
7. 피쳐와 타겟을 설정
8. 모델 개념

머신러닝의 모델이란 학습을 통해 알고리즘이 적용된 프로그램을 말하며

이 모델을 통하여 새로운 입력에 대해 결과를 예측할 수 있다.

1. 모델의 결과 값이 수치를 예측해주는 기계학습은 무엇인가?

* 회귀

입력된 데이터들의 관계를 기울기와 절편으로 식을 만들어 입력 값에 대한 결과를 예측해주는 학습방법

1. 모델 파라미터는 무엇인가요?

개발자가 설정할수 없는것으로 회귀분석에서 사용하는 결정계수와 같은 말한다. 데이터들을 통해 모델 내부에서 결정되는 값을 말한다.

개발자가 바꿀수 있는것들은 하이퍼파라미터라고 한다.

1. KNN 알고리즘에 대해 설명해주세요

KNN은 인접한 데이터들과의 거리를 측정하여 가까운 K개의 이웃을 기준으로

* 분류일 경우 K개의 점들중 속한 그룹이 많은 값을 나타내고
* 회귀일 경우 K개의 값들의 평균으로 나타낸다.

K가 클수록 범위가 넓어져 최적의 K를 찾는 것이 중요하다.

학습데이터의 범위를 벗어난 아주큰 데이터나, 작은 데이터는 정확한 예측이 안된다는 단점이 있다.

1. 하이퍼 파라미터는 무엇인가요?

하이퍼파라미터는 개발자가 모델에서 설정 할 수 있는 값으로 KNN의 K와

가중치등 이있다.

1. Regression의 성능평가 지표들을 설명하세요.

MSE는 평균제곱 오차로 실제값과 예측값의 제곱의 평균이다. 낮은값일수록 좋다

RMSE는 MSE에 제곱근을 씌운것이다

R2 score는 결정계수로 값이 1에 가까울수록 모델이 정확하다는것이다.

1. LinearRegression 알고리즘에 대해 설명

회귀분석은 독립변수와 종속변수에 대한 관계를 알아보는것인데 그중

선형회귀알고리즘은 훈련데이터에 대해 가장 잘 맞는 선형 방정식을 찾는것이다.

선형회귀를 하기 위해서는 연속형 변수를 사용하여야 한다.

1. Classification의 성능평가 지표들을 설명

F1 score는 0~1의 값을 가지며 1에 가까울수록 좋다.

1. 피쳐공학에 대해 설명하세요.

피쳐공학은 기존데이터를 학습시키기 용이하게 바꿔주는 것을 말한다.

차원축소, 특징 생성등이 있다.

1. LogisticRegression 알고리즘에 대해 설명

로지스틱 회귀는 2진 분류 모델로 데이터가 어떤 범주에 속할지 0과 1사이의 값으로 예측하고 그 확률에 따라 더 높은 범주에 속하는 것으로 분류해지는 지도 학습 알고리즘이다.

1. 과대적합/과소적합에 대해 설명해 주세요.

* 과대적합: 과대적합은 훈련용 데이터에 과적합되어 테스트용 데이터에 적합하지않게 된경우를 말한다. 모델의 복잡도가 높을수록 과대적합이 일어날 경우가 높다.
* 과소적합: 모델이 데이터가 부족하여 훈련이 제대로 되지않아 규칙을 제대로 찾지 못한 경우를 말한다. 모델이 복잡도가 너무 낮거나, 데이터가 부족할 때 주로 생긴다.