1주차 결과 보고서

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

1. 실험시간에 작성한 코드에 전체적인 진행 흐름을 \*\*diagram\*\*으로 그리고, 각 단계별로 어떤 기능을 구현했는지 서술하세요.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.머신 러닝을 위해 필요한 라이브러리(numpy, pandas, sklean 등)를 로드한다.

라이브러리 연결

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.sklearn에 있는 켈리포니아 집값에 있는 데이터를 가지고 온다. fetch\_california\_housing이라는 함수를 사용한다. data에는 DataFrame 형식으로, description에는 텍스트 형식으로 가지고 왔다.

데이터 로드

텍스트, 폰트, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.가지고 온 데이터의 크기를 출력하고 seabon에 있는 기능을 이용하여 히스토그램으로 데이터를 시각화한다.

데이터 시각화

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

변수 X와 y에 각각 Feature 데이터와 Target 값을 넣어준다.

데이터 분리

(훈련용, 테스트용)

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.train\_test\_split 함수를 이용하여 학습용 데이터와 테스트용 데이터로 분리해준다. 이때 학습과 테스트의 데이터 비율은 8:2로 하였다.

모델 선언 및 학습

sklearn에서 LinearRegression모델을 선언해준다. 선언한 모델에 fit 함수로 학습용 데이터를 학습시킨다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

predict로 모델의 예측값을 받아 모델의 예측값과 실제값의 오차의 제곱을 mean\_squared\_error으로 구하고 그것의 제곱근을 구해 RMSE를 알아낸다.

예측 값 확인 및 평가

2. Linear Regression 모델이 California housing dataset의 변수들을 어떻게 모델링했는지, 수식으로 표현하세요.

Linear Regression 모델은 을 따른다.

이때

= 예측한 집값

= 번째 Feature 데이터

= 번째 Feature 데이터에 대한 가중치

= 보정치

이다.

가중치()에 따라서 해당하는 Feature 데이터()가 집값()에 미치는 영향이 달라진다.

coef\_로 학습에 사용된 가중치()를 알 수 있고 intercept\_를 이용하여 보정치()를 알 수 있다. 학습에 사용된 Feature 데이터는 전체 9개의 데이터 중에서 MedHouseVal를 제외한 8개이다. 따라서 가중치() 또한 8개가 나와야 한다.

modle.coef\_과 modle.intercept \_를 실행하면 다음과 같이 나온다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

가중치() 8개와 보정치() 1개가 나온 것을 확인할 수 있다. Linear Regression 모델은 결과적으로 modle.coef\_으로 나온 가중치()들과 제공되는 데이터()들의 곱의 합 더하기 보정치()의 결과값이 예측치로 나오는 것이다. modle.coef\_의 값을 순서대로라고 하고 modle.intercept\_의 값을 라 한다. 이를 수식으로 표현하면 “”이다.

가중치의 값이 가장 큰 것은 임으로 Feature 데이터 중 의 영향이 가장 크다고 볼 수 있다. 가중치의 값이 가장 작은 것은 임으로 Feature 데이터 중 의 영향이 가장 작다고 볼 수 있다.

3. 결과값으로 나온 \*\*0.7456\*\*의 `Test RMSE`값이 적절한 값인지 생각해보고 의견을 서술하세요.

결과값으로 나온 Test RMSE값 0.7456은 적절하다고 볼 수 없다. MedHouseVal 범위가 0.15~5.0이라고 하면 약 15%의 오차를 보이는 것이다. 또한 0.7456을 금액으로 환산한다면 MedHouseVal은 100,000 USD단위임으로 평균적으로 약74,560 USD의 오차가 발생함을 의미한다. 결코 작지 않은 오차임으로 더 나은 방법을 찾아야 한다고 생각한다.

다만 머신 러닝에 사용한 모델이 선형구조를 가지는 Linear Regression것을 고려하였을 때 단순한 선형구조가 아니라 여러 비선형적 특성이 있는 집값이라는 데이터를 예측할 때 다소 오차가 발생함은 이해해야 한다고 생각한다.

더 나은 결과를 얻고 싶다면 범위를 벗어나는 이상치를 제거하여 모델을 학습시킨다면 보다 더 나은 성능을 기대할 수 있을 것이다. 또는 Random Forest모델과 같은 비선형적 데이터를 반영할 수 있는 모델을 선택하여 학습시킨다면 더 낮은 RMSE값을 얻을 수도 있을 것이다.