머신러닝 톺아보기 2주차 과제

<머신러닝 프로젝트 기획>

2025170938 박민서

(1)회귀 모델의 훈련 과정을 이용한 머신러닝 시스템의 전체 훈련 과정

:큰 그림=>데이터 구하기=>데이터 탐색&시각화=>데이터 준비=>모델 선택&훈련=>모델 조정=>솔루션 제시=>시스템 론칭, 모니터링, 유지 보수

(2)머신러닝과 데이터

:관련된 데이터를 구하고 기초적인 정보를 확인한 후에 어떤 모델을 어떻게 훈련시킬 것인가를 판단한다.

(3)데이터 활용법 확인

1. 데이터 기초 정보 확인

데이터셋 크기와 특성 ex)경도,위도,주택건물 중위연령,...,주택 중위가격, 해안 근접도

머신러닝 모델의 타깃:어떤 구역에 대해 주택 중위가격을 제외한 9개의 특성이 주어졌을 때, 해당 구역의 주택 중위가격을 target으로 예측하는 시스템

2. 훈련 모델 확인

구역별 주택 중위가격을 타깃으로 예측하는 시스템에 활용될 회귀 모델을 훈련시킨다.

지도 학습: 구역별 주택 중위가격을 타깃, 값 자체를 정확하게 예측해야 한다.

회귀: 주택 중위가격(연속형 데이터)를 예측한다.

다중회귀(구역별로 여러 특성을 주택 가격 예측에 사용)이자, 단변량 회귀(구역별로 한 종류의 값만 예측) 모델이다.

배치 학습:빠르게 변하는 데이터에 적응할 필요가 없으며, 데이터셋의 크기가 작기에 데이터셋 전체를 대상으로 훈련을 진행한다.

(4)데이터 구하기

:캘리포니아 주의 구역별 주택 중위가격을 예측하는 모델을 훈련시키려 한다. 이를 위해 먼저 데이터를 다운로드하고 적재한다.

(데이터 다운로드)

1. load_housing_data() 함수는 지정된 github repository에 압축파일로 저장되어 있는 캘리포니아 주택가격 데이터를 다운로드한 후에 pandas 데이터프레임으로 변환하여 반환한다. 따라서 housing 변수는 캘리포니아 주택가격 데이터를 담고 있는 데이터프레임을 가리킨다.

(5)데이터 탐색과 시각화

(데이터프레임과 데이터 탐색)

1.head(): 데이터프레임에 포함된 처음 5개의 샘플 확인

2.info(): 데이터셋 정보 요약

3.value_counts(): 범주형 특성 탐색

4.describe(): 수치형 특성 탐색

(훈련셋과 테스트셋): 훈련셋과 테스트셋을 구분하기 위해 train_test_split() 함수를 이용한다. 무작위로 데이터의 20%정도를 테스트셋으로 지정할 수 있다. 하지만 여기서는 소득 계층을 고려하면서 훈련셋과 테스트셋을 분류한다.

1.계층 샘플링: 계층 샘플링을 위해 대부분 구역의 중위소득이 1.5~6.0이라는 사실에 주목하고 소득 구간을 5개로 구분한다.

(데이터 시각화): 훈련셋만을 대상으로 탐색과 시각화를 적용한다. 먼저 훈련셋 원본을 그대로 두고 복사해서 사용한다.

(6)데이터 준비: 정제와 전처리

결측치 처리방법

방법1: 결측치 특성 포함 샘플 삭제

방법2: 결측치를 포함한 특성 삭제

방법3: 결측치를 해당 특성의 중앙값/평균값 등으로 대체

(7)파이프라인: 정제와 전처리 과정 전체를 아우르는 변환 파이프라인

1. 비율 변환기
2. 로그 변환기
3. 군집 변환기
4. 기본 변환기

(8)모델 미세 조정

(그리드 탐색): 그리드 탐색에 사용될 모델을 전처리와 함께 지정한다. 파이프라인에 포함된 전처리와 예측기에 사용되는 하이퍼파라미터 중에서 미세조정에 사용될 하이퍼파라미터가 가질 수 있는 값들의 리스트를 지정한다. 총 15개의 하이퍼파라미터 조합에 대해 모델을 지정한 다음에 매번 3-겹 교차 검증을 실행하기에 총 45번 훈련을 진행한다.

(랜덤 탐색): 무작위로 선택한 10개의 하이퍼파라미터 조합에 대해 3-겹 교차 검증을 진행하기에 총 30번의 훈련을 진행한다.

(9)최적 모델 저장 및 활용: 최적의 모델을 이름을 지정하여 저장한다. Joblib.load() 함수를 이용하여 저장된 모델을 불러올 수 있다. 다만, 모델 정의에 필요해서 사용자가 직접 정의한 함수, 클래스 등을 모두 함께 불러와야 한다. 불러온 모델을 이용하여 예측하려면 predict()를 이용한다.