머신러닝 톺아보기 세션

2주차. 머신러닝 프로젝트 처음부터 끝까지

유선호

0 CONTEXT

1. 머신러닝과 데이터
2. 데이터 활용법 확인
3. 데이터 구하기
4. 데이터 탐색과 시각화
5. 데이터 준비: 정제와 전처리
6. 파이프라인
7. 모델 선택과 훈련
8. 모델 미세 조정
9. 최적 모델 저장과 활용

0 CONTEXT

주택 가격을 예측하는 다양한 **회귀 모델**의 훈련 과정을 이용하여 머신러닝 시스템의 전체 훈련 과정을 살펴본다.

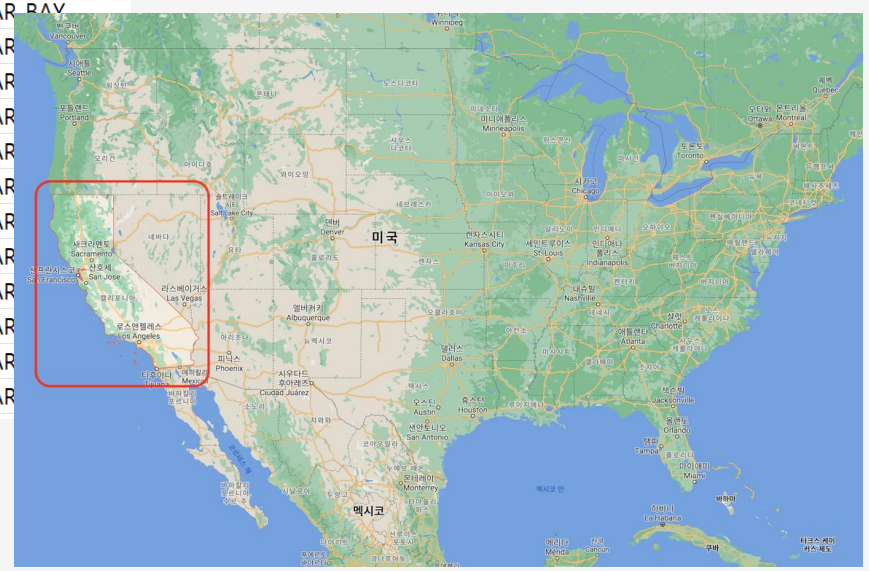
WeTIE

1 머신러닝과 데이터

머신러닝으로 문제를 해결하려면 먼저 관련된 데이터를 구하고

기초적인 정보를 확인한 후에 어떤 모델을 어떻게 훈련시킬 것인가를 판단한다.

	A	B	C	D	E	F	G	H	I	J
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
2	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
3	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
4	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
5	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
6	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
7	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
8	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
9	-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY
10	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR
11	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR
12	-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR
13	-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR
14	-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR
15	-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR
16	-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR
17	-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR
18	-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR
19	-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR
20	-122.26	37.84	50	2239	455	990	419	1.9911	158700	NEAR



1990년 미국 캘리포니아 주에서 수집한 주택가격 데이터

WeTIE

2 데이터 활용법 확인

1. 데이터 기초 정보 확인

데이터셋 크기와 특성

1990년도에 시행된 미국 캘리포니아 주의 20,640개의 구역별 주택가격 데이터

10개의 features : 경도, 위도, 주택 건물 중위연령, 총 방 수, 총 침실 수, 인구, 가구 수, 중위소득, 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.

다중 회귀이자, 단변량 회귀 모델이다.

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

*다중 회귀(multiple regression) : 구역별로 여러 특성을 주택 가격 예측에 사용

*단변량 회귀(univariate regression) : 구역별로 한 종류의 값만 예측

실습 자료 참고

실습 자료 참고

데이터 준비: 정제와 전처리

실습 자료 참고

실습 자료 참고

실습 자료 참고

8 모델 미세 조정

실습 자료 참고

9 최적 모델 저장 및 활용

실습 자료 참고

Q.E.D

2주차. 머신러닝 프로젝트 처음부터 끝까지