핸즈온머신러닝
챕터2 - 머신러닝 프로젝트 처음부터 끝까지

22.04.06

목차

진행 단계

1. 큰 그림을 봅니다.
2. 데이터를 구합니다.
3. 데이터로부터 통찰을 얻기 위해 탐색하고 시각화합니다.
4. 머신러닝 알고리즘을 위해 데이터를 준비합니다.
5. 모델을 선택하고 훈련시킵니다.
6. 모델을 상세하게 조정합니다.
7. 솔루션을 제시합니다.
8. 시스템을 론칭하고 모니터링하고 유지 보수합니다.

2.1 실제 데이터로 작업하기

- Dataset: California Housing Prices
- https://www.kaggle.com/datasets/camnugent/california-housing-prices

2.2 큰 그림 보기

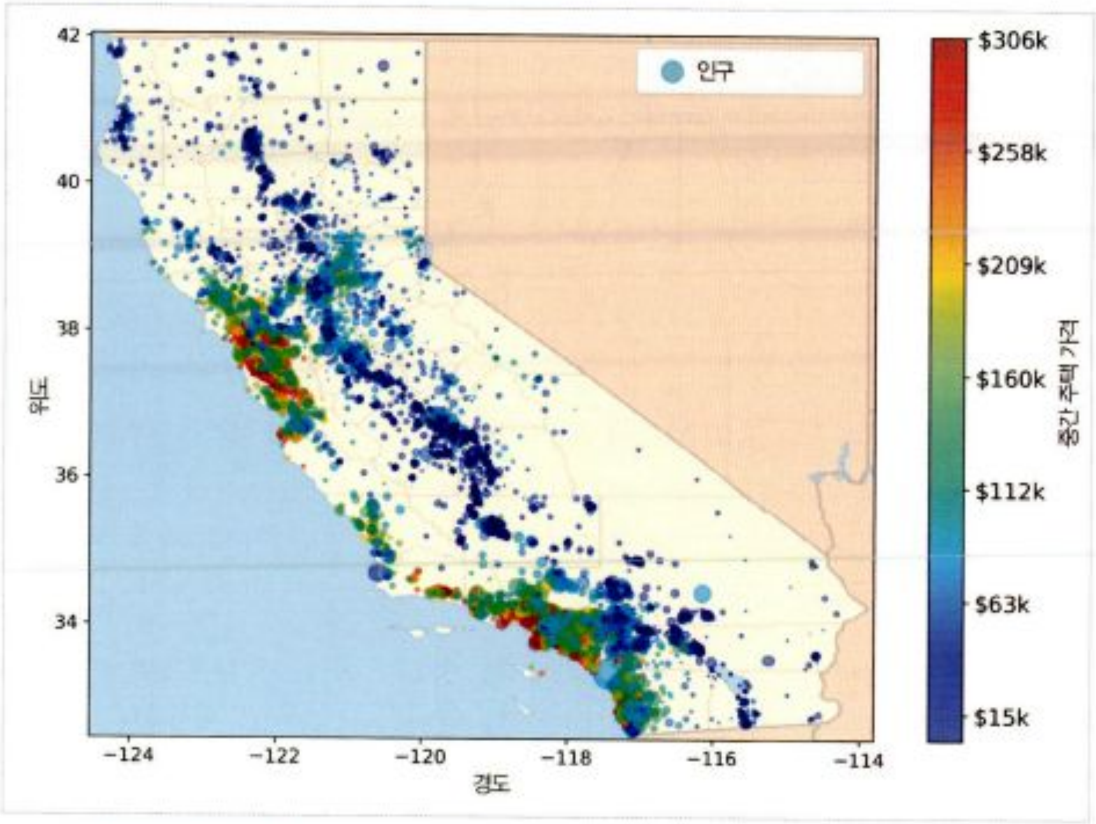

그림 2-1 캘리포니아 주택 가격

2.2 큰 그림 보기

캘리포니아 인구조사 데이터를 사용해 캘리포니아의 주택 가격 모델을 만드는 것

- block group
- population
- median income
- median housing price

파이프라인

데이터 파이프라인(pipeline)

- 데이터 처리 컴포넌트(component)들이 연속되어 있는 것
- 컴포넌트들은 비동기적 동작/독립적
- 일정 시간 후 파이프라인의 다음 컴포넌트가 그 데이터를 추출해 자신의 출력 결과를 만듦

2.2.1 문제 정의

목적이 무엇인가?

- 이 모델을 사용해 어떻게 이익을 얻으려고 하는지
- 문제를 어떻게 구성할지
- 어떤 알고리즘을 선택할지
- 모델 평가에 어떤 성능 지표를 사용할지
- 모델 튜닝에 얼마나 노력을 투여할지

2.2.1 문제 정의

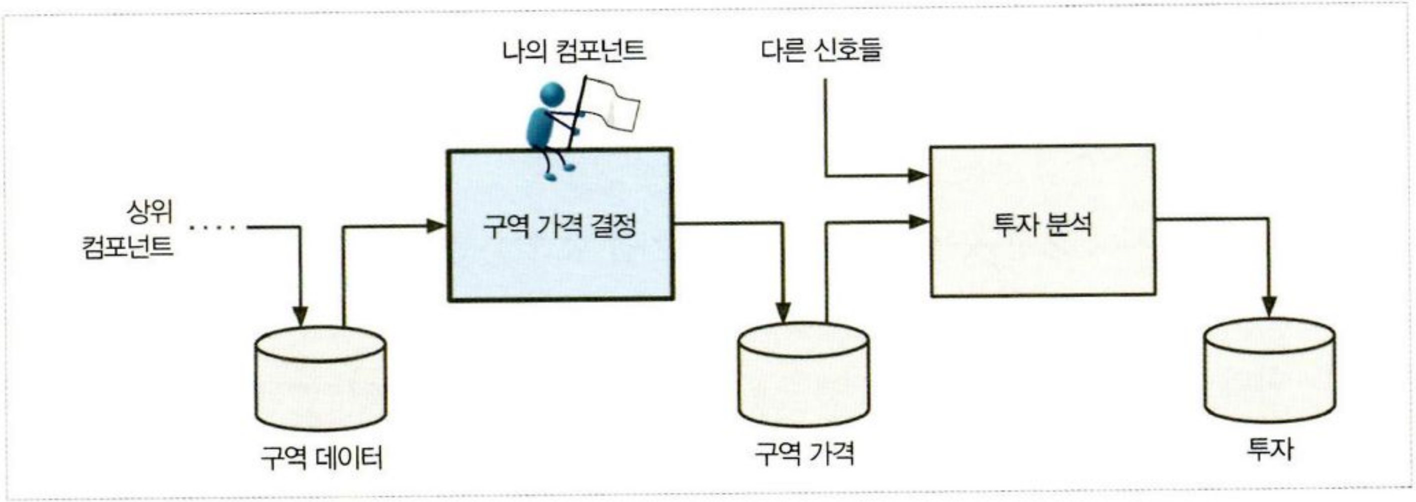

그림 2-2 부동산 투자를 위한 머신러닝 파이프라인

2.2.1 문제 정의

- 훈련 샘플에 레이블이 있음 → 지도 학습
- 값을 예측해야 함 → 회귀
- 예측에 사용할 특성이 여러 개 → 다중 회귀(multiple regression)
- 구역마다 여러 값을 예측 → 다변량 회귀(multivariate regression)
- 데이터에 연속적인 흐름이 없고, 메모리에 들어갈만큼 작으므로

  일반적인 배치 학습이 적절함

2.2.2 성능 측정 지표 선택

- 회귀 문제의 전형적인 성능 지표: 평균 제곱근 오차

$$root\ mean\ square\ error^{RMSE}$$

- 오차가 커질수록 이 값의 더 커지므로 예측에 얼마나 오류가 있는지 알 수 있음

$$RMSE(X,h) = \sqrt{\frac{1}{m}\Sigma_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

$$RMSE(X,h) = \sqrt{\frac{1}{m}\Sigma_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

m 데이터셋에 있는 샘플 수

$x^{(i)}$ 데이터셋에 있는 i번째 샘플의 전체 특성값 벡터

$y^{(i)}$ 해당 레이블(해당 샘플의 기대 출력값)

X 데이터셋에 있는 모든 샘플의 모든 특성값을 포함하는 행렬

(샘플이 하나의 행이어서 i번째 행은 $(X^{(i)})^T$ 로 표기

h 예측함수이며 가설(hypothesis)라고 한다.

$$RMSE(X,h) = \sqrt{\frac{1}{m}\Sigma_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

- 가설 h를 사용하여 일련의 샘플을 평가하는 비용 함수
- 시스템이 하나의 샘플 특성 벡터 $x^{(i)}$ 를 받으면

  그 샘플에 대한 예측값 $\hat{y}^{(i)} = h(x^{(i)})$ 를 출력함

평균 절대 오차(mean absolute error)

$$MAE(X,h) = \frac{1}{m}\Sigma_{i=1}^{m}|h(x^{(i)}) - y^{(i)}|$$

- 예측값의 벡터와 타깃값의 벡터 사이의 거리를 재는 방법
- 거리 측정에는 norm 계산이 가능
- RMSE는 유클리디안 노름(Euclidean norm)에 해당 $(l_2, ||\cdot||_2, ||\cdot||)$
- 절댓값의 합을 계산하는 것은 l_1 노름에 해당하며 $||\cdot||_1$ 로 표기(맨해튼 노름)

평균 절대 오차(mean absolute error)

$$MAE(X,h) = \frac{1}{m}\Sigma_{i=1}^{m}|h(x^{(i)}) - y^{(i)}|$$

- 일반적으로 원소가 n개인 벡터 v의 l_k 노름은 아래와 같이 정의

$$||v||_k = (|v_0|^k + |v_1|^k + \cdots + |v_n|^k)^{\frac{1}{k}}$$

- 노름의 지수가 클수록 큰 값의 원소에 치우치며 작은 값은 무시됨

  그래서 RMSE가 MAE보다 조금 더 이상치에 민감함