5장 회귀

5.1 회귀 소개

- 회귀 분석이란
    - 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 통계학 기법
    - 여러개의 독립변수와 한개의 종속변수 간의 상관 관계를 모델링 하는 기법을 통칭
    - 머신러닝 회귀 예측의 핵심
        - 주어진 피처와 결정값 데이터 기반 학습을 통해 최적의 회귀 계수를 찾아내는 것
        - 회귀 계수가 중요한 만큼 회귀 계수에 따라 선형, 비선형으로 나뉘고/ 독립벼수 개수에 따라 단일, 다중 회귀로 나뉜다
            - 대표적인 선형회귀
                - 일반 선형회귀
                - 릿지
                - 라쏘
                - 알라스틱넷
                - 로지스틱회귀

5.2 단순 선형 회귀를 통한 회귀 이해

- 단순 선형 회귀란
    - 독립변수, 종속 변수가 하나인 선형 회귀
    - 최적의 단순 선형 회귀 모델을 만든다는 것은 실제 값과 회귀 모델 차이에 따른 오류 즉, 남은 오류(잔차)를 최소화 하는 것
    - 오류 합 계산시에는 mean absolute error나 rss(residual sum of square)를 사용
    - 미분 등의 계산을 편리하게 하기 위해 rss 반식으로 오류 합을 구함
        - Error^2 = RSS

5.3 경사 하강법

- 점진적으로 반복적인 계산을 통해 회귀 계수를 업데이트하면서 오류값이 최소가 되는 회귀계수를 구하는 방식
    - 2차 함수의 최저점은 미분값인 1차 함수의 기울기가 가장 최소일때

5.4 사이킷런 LinearRegression을 이용한 보스턴 주택 가격 예측

- 사이킷런의 linear_models: 다양한 선형 기반 회귀를 클래스로 구현해 제공
- LinearRegression 클래스를 사용해서 보스턴 주택 가격을 옉측
    - LinearRegression 클래스는 예측값과 실제 값의 RSS를 최소화해 OLS 추정 방식으로 구현한 클래스
- 회귀의 평가
    - 실제값과 예측값의 차이를 그냥 더하면 오류가 상쇄됨
    - 이 때문에 여러 평가 지표가 있다
        - MAE
        - MSE
        - RMSE
        - R^2

5.5 다항 회귀

- ~5.3까지의 회귀는 1차 방정식(직선)으로 표현한 회귀
- 독립 변수의 방정식이 2차, 3차의 다항식으로 표현되면 다항 회귀라고 함(비선형 회귀와 다름)
- 회귀를 나누는 기준은 회귀 계수의 선형/비선형 여부, 독립변수의 선형/비선형 여부가 아님
- 사이킷런에서 다항 회귀에 대한 클래스는 명시적으로 제공하지 않아 여러 다른 클래스를 사용해, 선형 회귀함수에 넣는 방식으로 사용해야 함

5.6 규제 선형 모델 - 릿지, 라쏘, 엘라스틱 넷

- 회귀 모델은 과적합, 과소적합이 쉬워 이를 개선하기 위해 규제해야함
- 과적함을 개선하기 위한 비용함수 목표
    - 비용함수목표 = $Min(RSS(W)+alpha*||W||)$
    - 규제는 크게 L2 방식과 L1방직으로 구분
        - L2 규제는 대표적으로 릿지가 있음
            - W의 제곱에 대해 패널티를 부여하는 방식
            - 릿지 클래스의 주요 생성 파라미터는 alpha이고 이는 릿지 회귀의 alphaL2 규제 계수에 해당
        - L1규제는 대표적으로 라쏘가 있고
            - W의 절댓값에 대해 패널티를 부여
            - 라쏘 클래스는 L2 구제가 회귀 계수의 크기를 감소시키는데 반해, L1규제는 불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거한다
        - 엘라스틱 회귀는 L2 규제와L1규제를 결합한 회귀이다

5.7 로지스틱 회귀

- 로지스틱 회귀는 선형 회귀 방식을 분류에 적용한 알고리즘
    - 분류에 사용된다
    - 회귀가 선형/비선형인지는 가중치 변수에 따른다
    - 선형 회귀와 다른점은 학습을 통해 선형 함수의 회귀 최적선을 찾는 것이 아니라 시그모이드 함수 최적선을 찾고 시그모이드 함수의 반호나 값을 확률로 간주해 확률에 따라 분류를 결정

5.8 회귀 트리

- 크리 기반의 회귀를 말함
    - 함수로 봤을 때 X 값의 균일도를 반영한 지니 계수를 따라 분할하여 규칙을 만들고 노드를 만들 수 있음
    - 리프 노드 생성 기준에 부합하는 트리 분할이 완료됐다면 리프 노드에 소속된 데이터 값의 평균값을 구해서 최종적으로 리프 노드에 결정값으로 할당