5장 회귀

01 회귀 소개

- 머신러닝 회귀 예측의 핵심은 최적의 회구 lrPtn를 찾아내는 것
- 독립변수의 개수가 1개이면 단일회귀 / 여러개이면 다중 회귀
- 회구 계수의 결합이 선형이면 선형회귀/ 비선형이면 비선형회귀

02 단순 선형 회귀를 통한 회귀 이해

- 독립변수도 하나 종속변수도 하나인 선형 회귀
- 최적의 회귀 모델을 만든다는 것은 전체 데이터의 잔차(오류값)이 최소가 되는 모델을 만든다는 의미
- 일반적으로 계산을 편리하게 하기 위해 RSS 방식으로 오류 합을 구함

03 비용 최소화하기 – 경사 하강법

- 반복적인 계산을 통해 W 파라미터의 값을 업데이트하면서 오류 값이 최소가 되는 W 파라미터를 구하는 방식

- 경사 하강법의 프로세스

step 1 :w1,w2를 임의의 값으로 설정하고 첫 비용함수의 값으로 계산합니다.
step 2 :w1을 업데이트한 후 다시 비용함수의 값을 계산합니다
step 3: 비용함수가 감소하는 방향성으로 주어진 횟수만큼 step2를 반복하면서 w1과 w0를 계속 업데이트합니다.

04 사이킷런 LinearRegression을 이용한 보스턴 주택 가격 예측

회귀 평가 지표는 다음과 같다.

- MAE :실제 값과 예측값 차이를 절댓값으로 변환해 평균함
- MSE :Mean Squared Error(MSE)이며 실제 값과 예측값의 차이를 제곱해 평균함
- RMSE: 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 것
- R^2: 분산 기반으로 예측 성능 평가

05 다항 회귀와 과(대)적합/ 과소적합 이해

- 다항 회귀란 회귀가 독립변수의 단항식이 아닌 2차,3차 방정식과 같은 다항식으로 표현되는 것

- 다항식의 차수가 높아질수록 매우 복잡한 피처 간의 관계까지 모델링 가능
- 하지만 다항회귀의 차수를 높일수록 학습 데이터에만 너무 맞춘 학습이 이뤄져서 정작 테스트 데이터 환경에서는 오히려 예측 정확도가 떨어짐

편향-분산 트레이드오프

- 일반적으로 편향과 분산은 한쪽이 높으면 한쪽이 낮아지는 경향
- 즉 편향이 높으로 분산은 낮아지고(과소적합) 반대로 분산이 높으면 편향이 낮아짐(과적합)
- 편향이 너무 높으면 전체 오류가 높고 편향을 점점 낮추면 동시에 분산이 높아지고 전체 오류도 낮아짐
- 편향을 낮추고 분산을 높이면서 전체오류가 가장 낮아지는 '골디락스'지점을 통과하면서 분산을 지속적으로 높이면 전체 오류 값이 오히려 증가하면서 예측 성능이 다시 저하됨

06 규제 선형 모델 – 릿지, 라쏘, 엘라스틱넷

- 비용 함수는 학습 데이터의 잔차 오류 값을 최소로 하는 RSS 최소화 방법과 과적합을 방지하기 위해 회귀 계수 값이 커지지 않도록 하는 방법이 균형을 이루어야 함

릿지 회귀

사이킷런은 Ridge 클래스를 통해 릿지 회귀 구현

라쏘 회귀

W의 절댓값에 페널티를 부여하는 L1 규제를 선형 회귀에 적용한 것이 라쏘 회귀
L2 규제가 회귀 계수의 크기를 감소시키는 데 반해, L1 규제는 불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거

엘라스틱넷 회귀

L2 규제와 L1 규제를 결합한 회귀
엘라스틱넷은 라쏘 회귀가 서로 상관관계가 높은 피처들의 경우에 이들 중에서 중요 피처만

을 셀렉션하고 다른 피처들은 모두 회귀 계수를 0으로 만드는 성향이 강함

07 로지스틱 회귀

- 로지스틱 회귀는 선형 회귀 방식을 분류에 적용한 알고리즘
- 시그모이드 함수 최적선을 찾고 이 시그모이드 함수의 반환 값을 확률로 간주해 확률에 따라 분류를 결정
- 로지스틱 회귀는 가볍고 빠르지만 이진 분류 예측 성능도 뛰어남

08 회귀 트리

- 회귀를 위한 트리를 생성하고 이를 기반으로 회귀 예측
- 회귀 트리는 리프 노드에 속한 데이터 값의 평균값을 구해 회귀 예측값을 계산
- 선형 회귀가 직선으로 예측 회귀선을 표현하는 데 반해 회귀 트리의 경우 분할되는 데이터 지점에 따라 브랜치를 만들면서 계단 형태로 회귀선을 만듦.