**5장. 회귀**

1. 회귀 소개

회귀: 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 통계학 기법

여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법을 통칭함

머신러닝 회귀 예측

핵심: 주어진 피처와 결정 값 데이터 기반에서 학습을 통해 최적의 회귀 계수를 찾아내는 것

|  |  |
| --- | --- |
| 독립변수 개수 | 회귀 계수의 결합 |
| 1개: 단일 회귀 | 선형: 선형 회귀 |
| 여러 개: 다중 회귀 | 비선형: 비선형 회귀 |

회귀는 예측값이 연속형 숫자 값

선형 회귀: 실제 값과 예측값의 차이를 최소화하는 직선형 회귀선을 최적화하는 방식

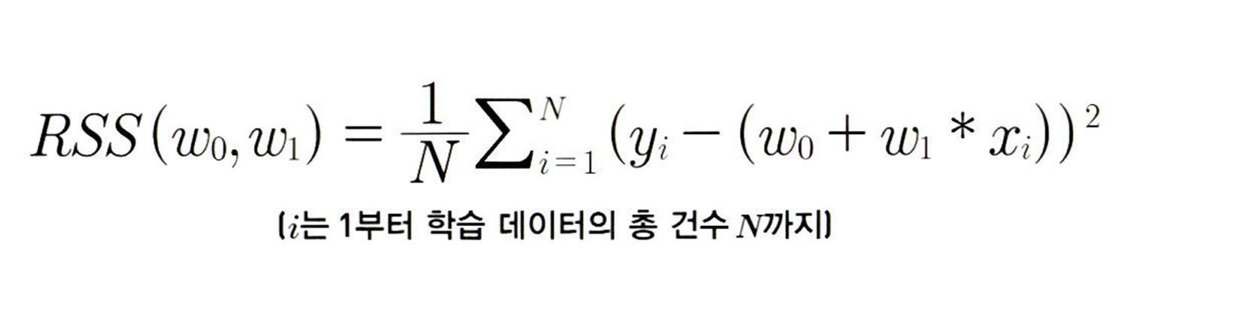
* 일반 선형 회귀: 예측값과 실제 값의 RSS를 최소화할 수 있도록 회귀 계수를 최적화하며, 규제를 적용하지 않은 모델
* 릿지: 선형 회귀에 L2 규제를 추가한 회귀 모델  
  L2 규제: 상대적으로 큰 회귀 계수 값의 예측 영향도를 감소시키 위해서 회귀 계수값을 더 적게 만드는 규제 모델
* 라쏘: 선형 회귀에 l1 규제를 적용한 방식  
  L1 규제: 예측 영향력이 작은 피처의 회귀 계수를 0으로 만들어 회귀 예측 시 피처가 선택되지 않게 하는 것/피처 선택 기능
* 엘라스틱넷: L2, L1 규제를 함께 결합한 모델. 피처가 많은 데이터 세트에서 적용  
  L1 > 피처의 개수를 줄이고 L2 > 계수 값의 크기를 조정
* 로지스틱 회귀: 분류에 사용되는 선형 모델

매우 강력한 분류 알고리즘으로 뛰어난 예측 성능을 보임

1. 단순 선형 회귀를 통한 회귀 이해

단순 선형 회귀: 독립변수 1개, 종속변수 1개인 선형 회귀

* 특정 기울기와 절편을 가진 1차 함수식으로 모델링 가능
* 예측값:
* 잔차: 실제 값과 회귀 모델의 차이에 따른 오류 값(남은 오류)
* 최적의 회귀 모델 > 잔차의 합이 최소가 되는 모델을 만들며 오류 값의 합이 최소가 되는 최적의 회귀 계수를 찾는 것
* 오류값의 합을 구하는 방식
  + Mean Absolute Error: 오류값의 절댓값을 합하는 방식
  + Residual Sum of Square(RSS): 오류값의 제곱을 구해서 더하는 방식
  + 일반적으로 RSS를 많이 이용

RSS는 정규화 된 식으로 표현된다

회귀에서 RSS는 비용(Cost)이며 w변수(회귀 계수)로 구성되는 RSS를 비용 함수라고 함

머신러닝에서는 데이터 학습을 통해 비용 함수의 반환 값을 감소시키며 가장 최소의 오류 값을 구함

1. 비용 최소화하기 – 경사 하강법 소개

경사 하강법: 점진적으로 반복적인 계산을 통해 W 파라미터 값을 업데이트하면서 오류 값이 최소가 되는 W 파라미터를 구하는 방식

반복적으로 W 값을 보정하면서 오류 값이 최소가 될 때의 W 값을 구한다.

경사 하강법 in 파이썬

모든 학습 데이터에 대해 반복적으로 비용함수 최소화를 위한 값을 업데이트 > 수행 시간이 매우 오래 걸림

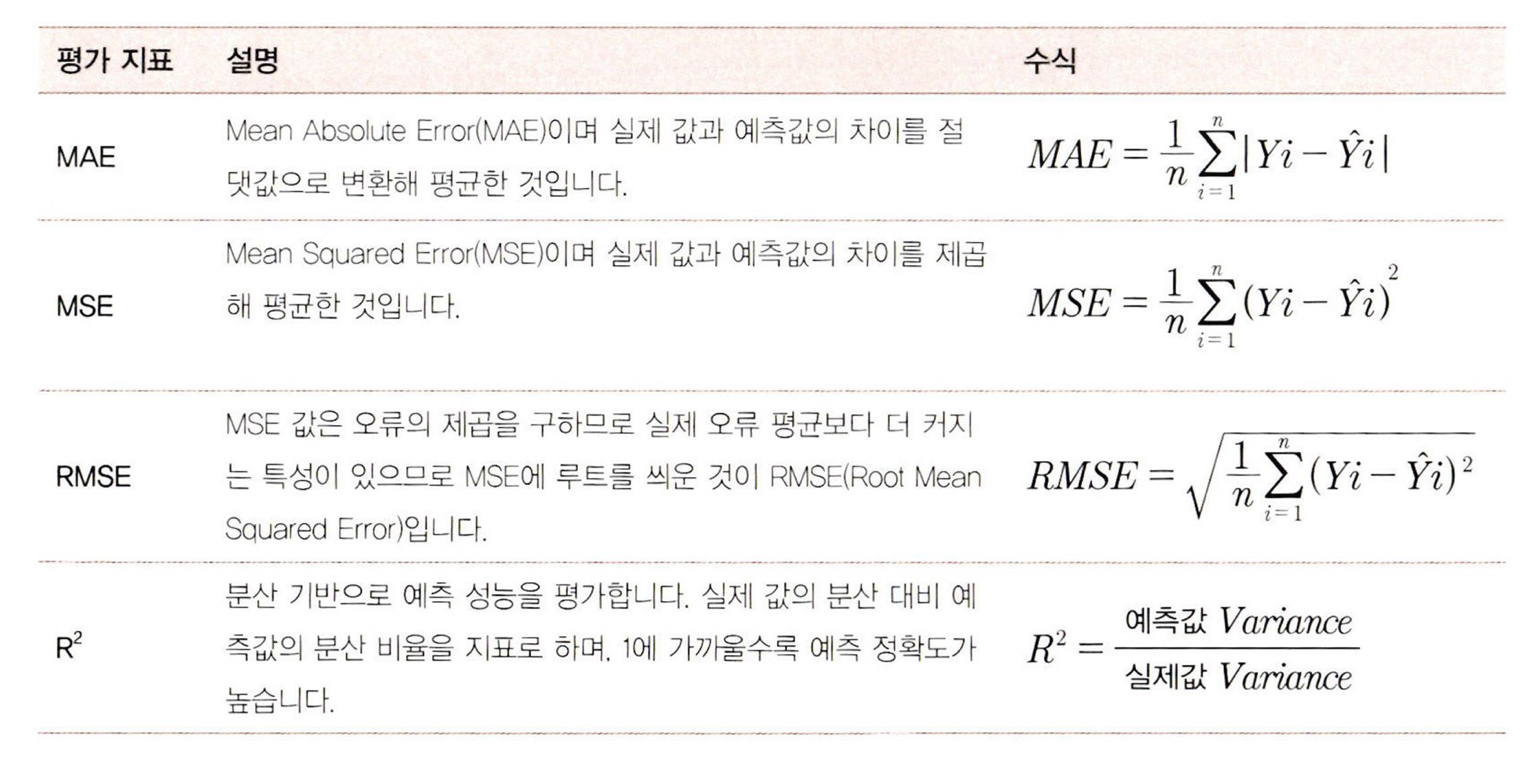
: 확률적 경사 하강법 이용

확률적 경사 하강법: 일부 데이터만 이용해 w가 업데이트 되는 값을 계산

1. 사이킷런 LinearRegressiond을 이용한 보스턴 주택 가격 예측

LinearRegression 클래스: 예측값과 실제 값의 RSS를 최소화해 OLS 추정 방식으로 구현한 클래스

fit( ) 메서드로 X, y 배열을 입력 받아 회귀 계수인 W를 coef\_ 속성에 저장

회귀 평가 지표

regplot( ) API: X, Y 축 값의 산점도와 함께 선형 회귀 직선을 그려줌

matplotlib.suplots( ): 여러 개의 그래프를 한 번에 표현하기 위해 자주 사용  
 인자) nclos: 열 방향으로 위치할 그래프 개수  
 nrows: 행 방향으로 위치할 그래프 개수

1. 다항 회귀와 과적합/과소적합 이해

다항 회귀: 회귀가 독립변수의 단항식이 아닌 2차, 3차 방정식과 같은 다항식으로 표현되는 것(선형 회귀)

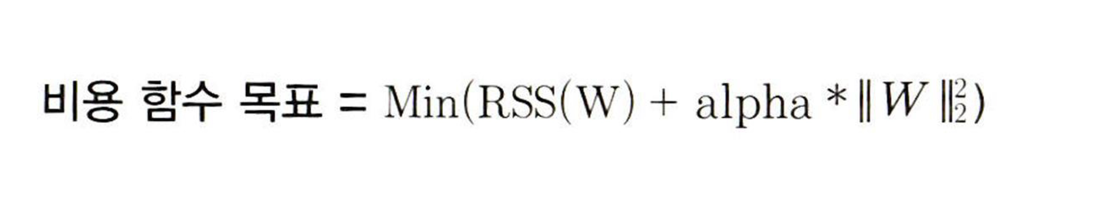
차수를 높일 수록 학습 데이터에만 맞춘 학습이 일어나 테스트 데이터 환경에서 예측 정확도가 떨어질 수 있다. 즉, 차수가 높아질 수록 과적합 문제 발생

편향-분산 트레이드오프

편향이 높으면 분산은 낮아지고(과소적합) 반대로 분산이 높으면 편향이 낮아진다(과적합)

편향과 분산이 서로 트레이드오프를 이루며 오류 cost 값이 최대로 낮아지는 모델을 구축하는 것이 가장 효율적인 머신러닝 예측 모델 만드는 방법

1. 규제 선형 모델 – 릿지, 라쏘, 엘라스틱넷

회귀 계수의 크기를 제어해 과적합을 개선하기 위해 비용함수의 목표를

위의 식을 최소화하는 것을 변경한다. 이때 alpha를 0에서부터 지속적으로 값을 증가시키면서 회귀 계수 값의 크기를 감소시킬 수 있는데 이와 같이 비용 함수에 alpha 값으로 페널티를 부여해 회귀 계수 값의 크기를 감소시켜 과적합을 개선하는 방식을 규제라고 함.

릿지 회귀

L2 규제를 적용한 회귀

주요 생성 파라미터: alpha

학습 결과: alpha 값이 증가하면서 회귀 계수가 지속적으로 작아지고 잇음 관찰 가능. 회귀 계수를 0으로 만들지는 않음.

라쏘 회귀

L1 규제를 적용한 회귀

불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거

적절한 피처만 회귀에 포함시키는 피처 선택의 특성을 가짐

학습 결과: alpha 값이 증가함에 따라 일부 피처의 회귀 계수는 0으로 바뀌기도 함. 회귀계수가 0인 피처는 회귀 식에서 제외되어 피처 선택의 효과를 얻을 수 있음.

엘라스틱넷 회귀

L2 규제와 L1 규제를 결합한 회귀

식을 최소화하는 W를 찾는 것

이 목표

두 규제가 결합되었기 때문에 수행시간이 상대적으로 오래 걸림.

07. 로지스틱 회귀

선형 회귀 방식을 분류에 적용한 알고리즘

시그모이드 함수 최적선을 찾고 이 함수의 반환값을 확률로 간주해 확률에 따라 분류 결정

시그모이드 함수:

08. 회귀 트리

회귀를 위한 트리를 생성하고 이를 기반으로 회귀 예측을 하는 것

리프 노드에 속한 데이터 값의 평균값을 구해 회귀 예측값을 계산