머신러닝 회귀분석 요약

1. 회귀의 이해: 회귀의 가장 종요한 내용은 회귀 계수이다. 회귀계수가 선형,비선형에 따라 선형회귀와 비선형회귀로 분류한다. 독립변수의 개수가 한 개인지 여러 개인지에 따라 단일회귀, 다중회귀로 나누어진다. 데이터가 정형데이터이면 선형회귀가 더 효과적이다. 선형회귀는 as\*b1+a2\*b2 이런 식으로 결합이 되어있는 것임. 비선형은 a1/(b1+b2)이런식임

분류와 회귀의 차이점은 분류는 결과값이 category값으로 이산값이고 회귀는 숫자값으로 연속값임

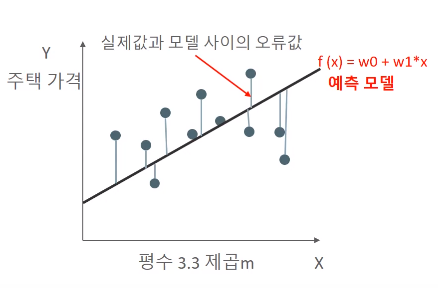
일반 선형회귀는 예측값과 실제값의 RSS(Residual Sum of Squares)를 최소화할 수 있도록 회귀계수를 최적화하며 규제를 적용하지 않음

릿지:L2규제 추가

라쏘: L1 규제 적용

엘라스틱넷: L2,L1같이 적용

로지스틱회귀: 이산값을 분류해서 분류에 사용되는 선형 모델이다.

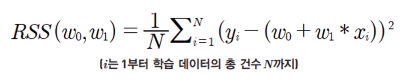


예측 모델의 평가는 실제값과 모델 사이의 오류값이 작아야함

그렇기에 최적의 회귀 모델을 만든다는 것은 바로 전체 데이터의 오류값 합이 최소가 되는 모델을 만든다는 의미로 동시에 오류 값 합이 최소가 될 수 있는 최적의 회귀 계수를 찾는다는 의미

Rss는 오류값의 제곱의 합을 더해서 계산한다.

Rss는 회귀식의 독립변수 x, 종속변수 Y 가 중심변수가 아니라 w변수(회귀 계수)가 중심 변수임을 인지하는 것이 가장 중요함



Rss는 비용함수라 부른다. 머신러닝 회귀 알고리즘은 반환값을 감소시키고 최종적으로는 더 이상 감소하지 않는 최소의 오류값을 구함. 손실함수라고도 한다

비용 최소화하기- 경사하강법(GRADIENT DESCENT)

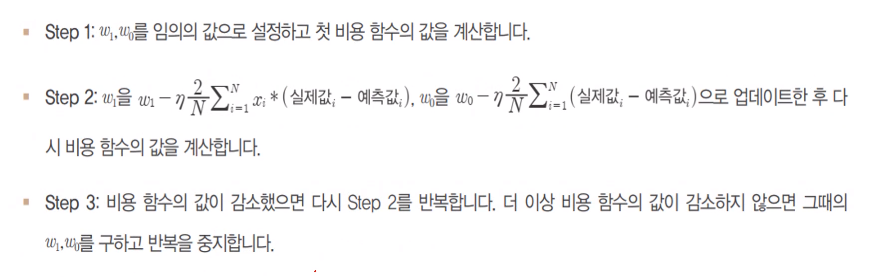
W파라미터의 개수가 적다면 고차원 방정식으로 비용함수가 최소가 되는 W 변숫값을 도출하지만 개수가 많으면 해결하기 어렵기에 경사하강법을 사용한다.

점진적으로 반복적인 계산을 통해 W파라미터의 값을 업데이트하면서 오류값이 최소가 되는 W파라미터를 구한다. 미분을 통하여 오류가 작아지는 방향으로 보정한다.

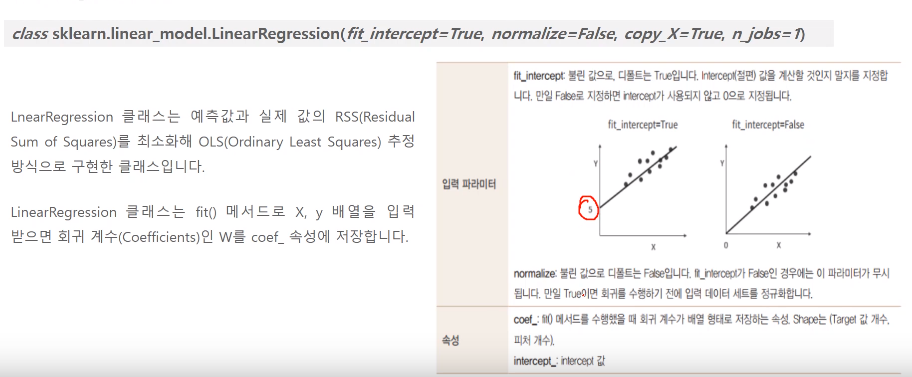
미분은 증가 또는 감소의 방향성을 나타내기 때문에 최초 W에서부터 미분을 적용한 뒤 이 미분 값이 계속 감소하는 방향으로 순차적으로 W를 업데이트 함,.

편미분을 통해 경사하강법을 진행하는데 결과값이 너무 클 수 있기 때문에 보정계수 N을 곱한다. 이를 학습률이라 하며 0~1사이에 있다.

경사 하강법 수행 프로세스



사이킷런 linear regression클래스



선형회귀는 입력 피처의 독립성에 많은 영향을 받아 피처간의 상관관계가 매우 높으면 분산이 매우 커져 오류에 민감해짐. 이를 다중 공선성 문제라고 한다. 일반적으로 상관관계가 높은 피처가 많으면 중요한 피처만 남기거나 규제함

회귀 평가지표



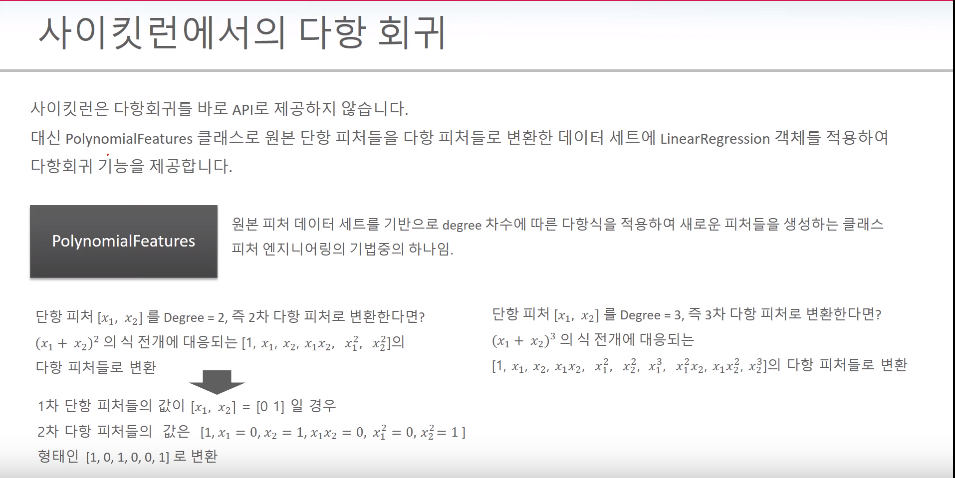
로그를 씌우는 경우들은 데이터의 값이 커지면 오류값도 많아지기에 로그를 씌워서 조정해줌

사이킷런 scoring 함수에 회귀 평가 적용시 유의사항

MAE의 사이킷런 scoring 파라미터는 음수값이다. 그 이유는 회귀 분석에서 오류값이 작을 수록 좋은 모델이기 때문에 음수를 곱해서 작은 오류값이 더 큰 숫자로 인식하게 해준다.

다항회귀(polynomial regression)

다항회귀는 선형회귀로 선형/비선형 회귀를 나누는 기준은 회귀 계수가 선형/비선형인지에 따른 것이고 독립변수의 선형/비선형 여부와는 무관하다.



단항 피처들을 다항 피처로 변경하고 linear regression 객체로 학습한다.

사이킷런에서는 일반적으로 pipeline 클래스를 이용하여 피처 변환과 linear regression 학습/예측을 결합하여 다항회귀를 구현한다.

다항회귀는 overfitting이 쉽다

Bias, variance가 낮은 모델이 좋다.

편향-분산 트레이드 오프

편향이 높고 분산이 낮아지면 과소적합이 되고 분산이 높고 편향이 낮아지면 과대적합이 된다.

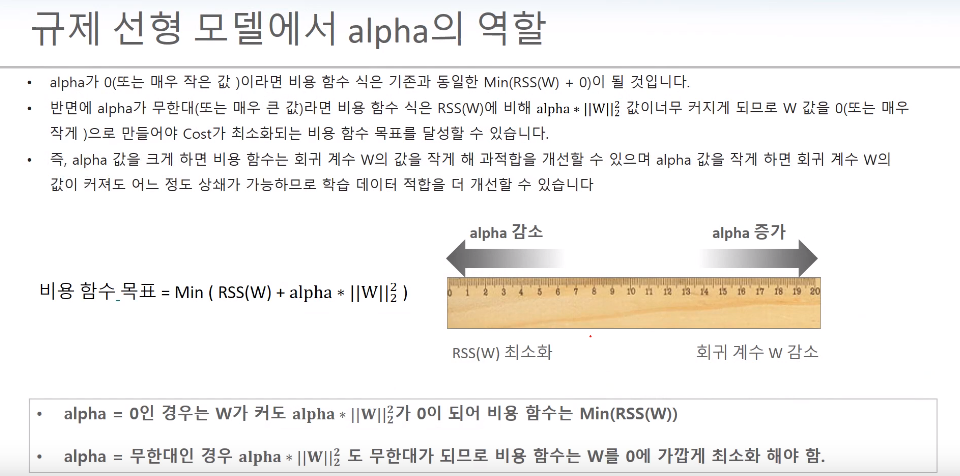
규제 선형 회귀

비용함수에 ALPHA값으로 페널티를 부여해 회귀 계수 값의 크기를 감소시켜 과적합을 개선하는 방식을 규제라고 부른다.

규제는 L2,L1방식으로 구분되며 L2 규제는 alpha\* |w|^2과 같이 w의 제곱에 대해 페널티를 부여하는 방식을 말한다. L2 규제를 적용한 회귀는 릿지 회귀

라쏘 회귀는 l1규제를 적용한 회귀. W의 절대값에 대해 페널티를 부여함 그래서 회귀계수 값을 0으로 변환한다.

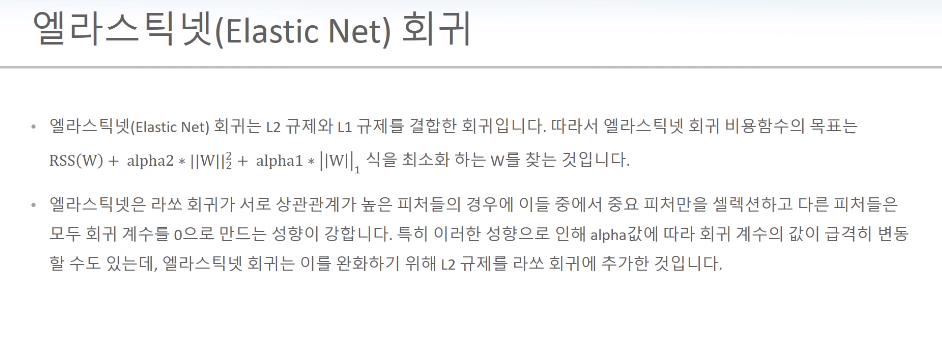
Elastic Net은 l1 l2규제를 모두 사용한 모델이다. 주로 피처가 많은 데이터 세트에서 적용되며 l1 규제로 피처의 개수를 줄이고 l2 규제로 계수 값의 크기를 조정한다.



라쏘 회귀

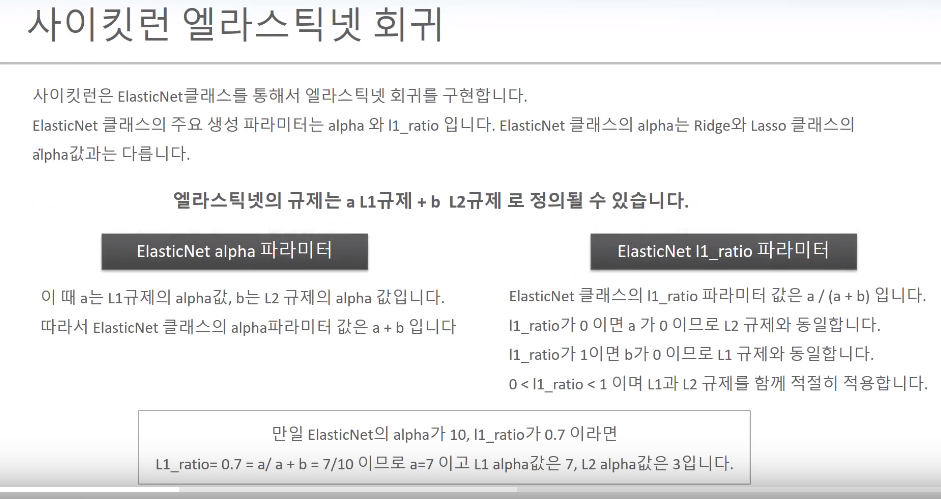
L1 규제는 불필요한 회귀 계수를 급격하게 줄여 0으로 만들고 제거하기 때문에 피처 셀렉션의 특징을 가진다.

엘라스틱 회귀



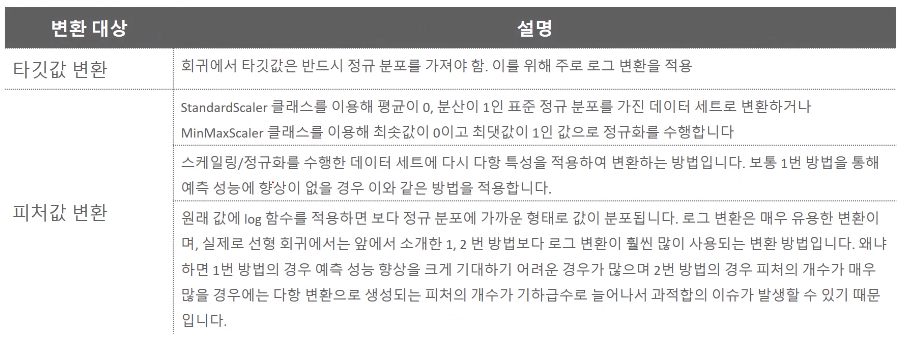
Elastic Net클래스로 구현한다.

주요 파라미터: alpha, l1\_ratio가 있다.



선형회귀 모델은 피처값과 타깃값의 분포가 정규 분포형태를 이루는 것을 가장 선호한다

타깃값을 변화시킬 때 반드시!! 정규 분포를 가져야 하기에 주로 로그 변환을 적용한다.

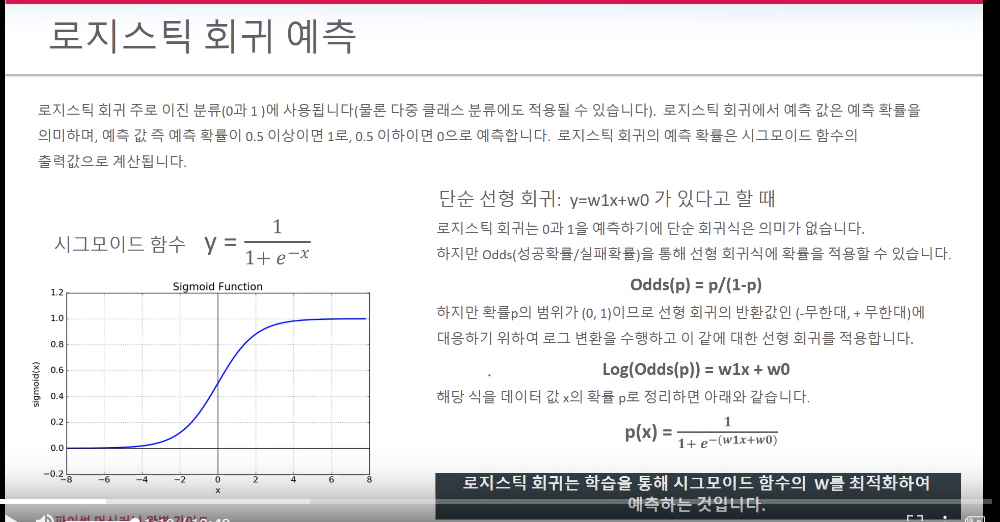


선형회귀의 데이터 인코딩은 원-핫 인코딩 사용

로지스틱 회귀는 분류에 선형 회귀 방식을 적용한 알고리즘이다.

로지스틱 회귀는 시그모이드 함수 최적선을 찾고 이 함수의 반환값을 확률로 간주해 분류를 결정함

로지스틱 회귀는 주로 이진 분류에 사용되며 로지스틱 회귀의 예측값은 예측 확률을 의미한다.



회귀 트리

