Introduction to Machine Learning with Python

A Guide for Data Scientists

Andreas Muller & Sarah Guido / O’REILLY

1장. Introduction

선형모델(Linear Model)

선형 모델은 입력 변수들의 선형 함수를 사용하여 예측을 하는 기법.

회귀분석을 위한 선형 모델

회귀 선형 모델에서 일반적인 예측 공식은 다음과 같다.

y = w0\*x0 + …… + wp\*xp + b

x : feature. p는 피쳐 개수.

w & b : 학습된 모델의 파라미터

y : 모델을 통해 나온 예측값

피쳐가 1개인 데이터셋의 경우

y = w\*x + b

w : 기울기, b : y 절편

피쳐가 여러 개인 경우에 각각의 w는 각 피쳐의 축에 따른 기울기. (다차원)

1차원 데이터셋인 wave 데이터를 학습한 결과.

w[0] 값에 의해 기울기가 약 0.4

회귀선이 y축을 지나는 절편값이 0보다 낮은 -0.03이고 plot을 통해 확인할 수 있다.

회귀선형모델은 2개의 피쳐에 대한 평면에 1개의 피쳐에 대한 회귀선으로 표현되며, 피쳐가 여러 개인 경우에는 그 개수만큼 다차원 공간에 초평면으로 표현된다.

직선을 이용한 예측을 하는 것은 매우 제한적인 결과를 낸다. 데이터의 상세한 특성들을 잃게 된다. 하지만, 어떤 경우 비현실적이긴 하지만 목표변수 y는 피쳐들의 선형 결합에

1차원 데이터에 대한 선형 모델은 어느 정도 편향된 관점을 준다. 하지만 많은 피쳐를 가진 데이터셋에 있어서 선형 모델은 매우 강력하다.

회귀분석을 위한 선형 모델은 여러 가지가 있다. 이 모델들은 트레이닝 데이터로부터 모델의 파라미터인 w와 b를 어떻게 학습하는지, 어떻게 모델의 복잡도를 조정하는지 등에 있어서 차이가 있다.

선형 회귀(Linear regression)

정규방정식(ordinary least squares)으로 알려진 선형 회귀는 가장 단순하고 고전적인 선형 모델이다. 선형 회귀는 트레이닝셋을 통해 실제 목표값인 y와 예측값 사이의 평균제곱오차(MSE, mean square error)를 최소화하는 방식으로 파라미터 w와 b를 찾아낸다. MSE는 실제 값과 예측값 간의 차를 제곱한 후 더한 것이다.

scikit-learn에서 선형 회귀는 아무런 파라미터가 없다. 이것이 장점이기도 하지만, 모델의 복잡도를 조정하는 방법이 없다는 단점도 된다.

기울기 파라미터인 w는 가중치(weight) 또는 coefficents로 불리며, coef\_ 속성에 저장되며, y절편(intercept)인 b는 intercept\_ 속성에 저장된다.

intercept\_ 값은 항상 단일 부동소수 값이며,