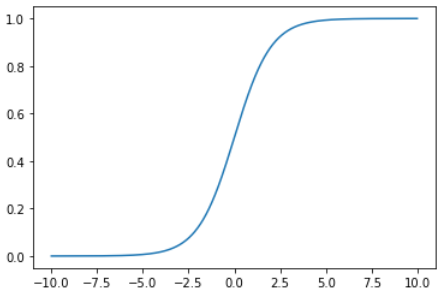
로지스틱 회귀(Logistic)에 대하여 알아보자.

1학기 마지막 프로젝트로써 선형 회귀(Linear regression)으로 진행한 프로젝트의 데이터 성능 평가가 상당히 낮았다. 이를 보완하기 위해서 더 많은 머신러닝 주제에 대한 공부가 필요하다고 판단하였고 로지스틱 회귀(Logistic)를 선택하게되었다.

로지스틱 회귀는 연속 형 데이터의 입력값으로 이산형(자료값의 속성이나 그룹을 나타내는 자료 (남녀 등)) 출력값을 가지는 회귀 분석의 한 종류이다.

입력된 데이터의 특징을 이용하여 0 과 1 사이의 출력값을 가지기 때문에 S커브 형태의 함수로 표현된다 때문에 분야에 따라 시그모이드(sigmoid) 함수로도 불리기도 한다.



승산(Odds)는 임의의 사건 A가 발생하지 않을 확률 대비 일어날 확률의 비율을 뜻하는 개념이다.

만약 P(A)가 1에 가까울수록 승산은 치솟을 것이다. 또한 P(A)가 0이라면 0이 된다.

승산(odds)에 로그를 씌우게 된다면 로그승산(로짓, logit)이라고 하는데 이를 종속변수로 하여 선형회귀모형을 만드는 것을 로지스틱 회귀분석이라고 한다.

승산비(odds ratio)는 특정 요인B의 여부에 따른 사건 발생 확률을 비교할 때 사용되는 척도이다

만약 3개 이상의 범주를 가진 데이터라면 다항 로지스틱 회귀(multinomial logistic regression)을 이용할 수 있다. 다항 로지스틱 회귀에서는 A, B, C 3개의 범주와 독립변수가 있을 때 2개의 유사로짓(pseudo logit)식을 만들 수 있다.

[링크]

타이타닉 데이터를 sklearn의 로지스틱 회귀 클래스를 이용하여 처리해보았다.



데이터 성능 평가결과 약 80%정도의 정확도가 측정되었다.

1학기 프로젝트의 정확도보다는 휠씬 높은 수치였지만 이산형 출력값을 가지는 로지스틱 회귀의 특성상 아쉬움이 남았다.

서울 iCT이노베이션스퀘어에서 진행한 수업에서 배웠던 시그모이드(sigmoid)를 활성화 함수로 사용한 분석 방법이라는 것이 인상 깊었다.