

Part.02
회귀분석

|로지스틱 회귀분석

FASTCAMPUS
ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택

I 로지스틱 회귀분석

■ 로지스틱 회귀란

- 로지스틱 회귀는 출력 변수를 직접 예측하는 것이 아니라, 두 개의 카테고리를 가지는 binary 형태의 출력 변수(‘성공’, ‘실패’ 또는 ‘예’, ‘아니오’)를 예측할 때 사용하는 회귀분석 방법임

$$p = P(\text{success} \mid k \text{ 개의 입력 변수}),$$

$$1 - p = P(\text{fail} \mid k \text{ 개의 입력 변수}).$$

- 로지스틱회귀에서는 k개의 입력 변수를 사용하여 성공 실패를 예측하기 위해 성공 확률 $p(X)$ 를 모델링 함

$$p(X) = P(\text{success} \mid X_1, \dots, X_k) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \text{random error}(\varepsilon)$$

- 방정식의 왼쪽의 범위는 $[0, 1]$ 이지만 오른쪽의 범위는 $[-\infty, +\infty]$ 이므로 다른 형태로 모델링 해야함

I 로지스틱 회귀분석

■ 로지스틱 함수(Logistic Function)

- 왼쪽항에 자연 로그를 취해줌으로써 $\ln(p(X))$ 는 $[-\infty, +\infty]$ 가 됨. 하지만 이를 만족하기 위해서는 $p(X)$ 가 $[0, +\infty]$ 의 범위이어야함

$$\ln(p(X)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

 $[-\infty, +\infty]$
 $[-\infty, +\infty]$

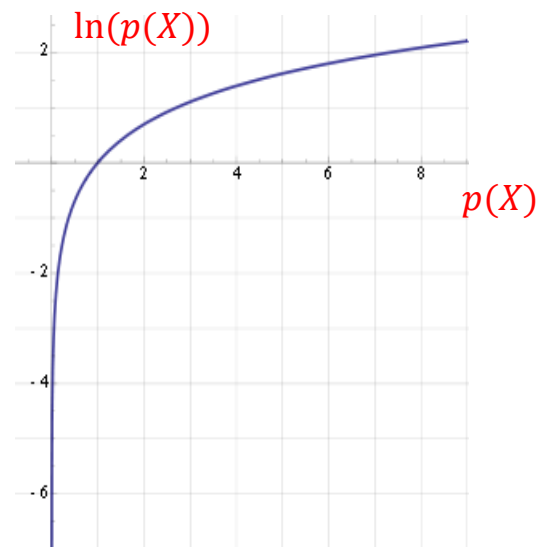
- 하지만, 확률 $p(X)$ 의 maximum값은 1이므로 $\ln(p(X))$ 가 $+\infty$ 값을 가질 수 없음. 따라서 왼쪽의 식을 다음과 같이 대체함

$$\text{logit} = \ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

도박에서
승률을 의미

- X 는 입력변수, Y 는 출력변수가 1이 될 확률일 때 식은 다음과 같이 정리할 수 있음

$$Y = p(X) = \frac{e^{\beta_0 + \beta_1 X + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X + \cdots + \beta_k X_k}}$$

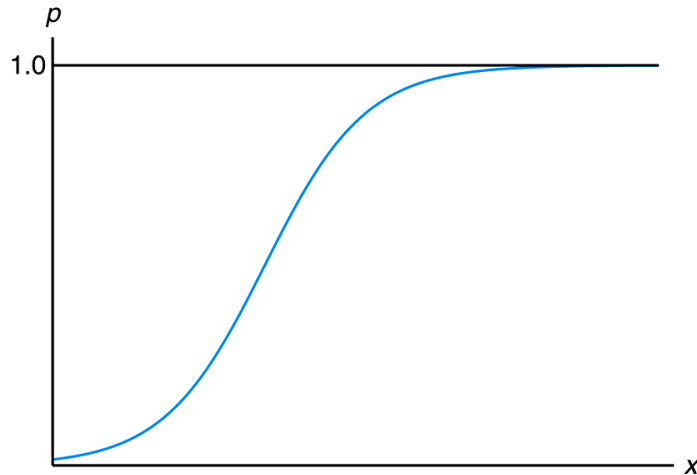


I 로지스틱 회귀분석

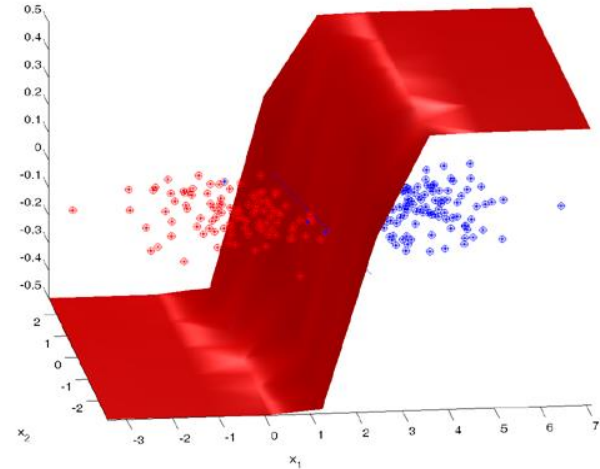
■ 로지스틱 함수(Logistic Function)

- 입력 변수가 1개일 때, 로지스틱 회귀식은 아래 왼쪽 그림과 같음
- 입력 변수가 2개일 때, 로지스틱 회귀식은 아래 오른쪽 그림과 같음

$$Y = p(X) = \frac{e^x}{1 + e^x}$$



출처: Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). *Probability and statistics for engineers and scientists* (Vol. 5). New York: Macmillan.

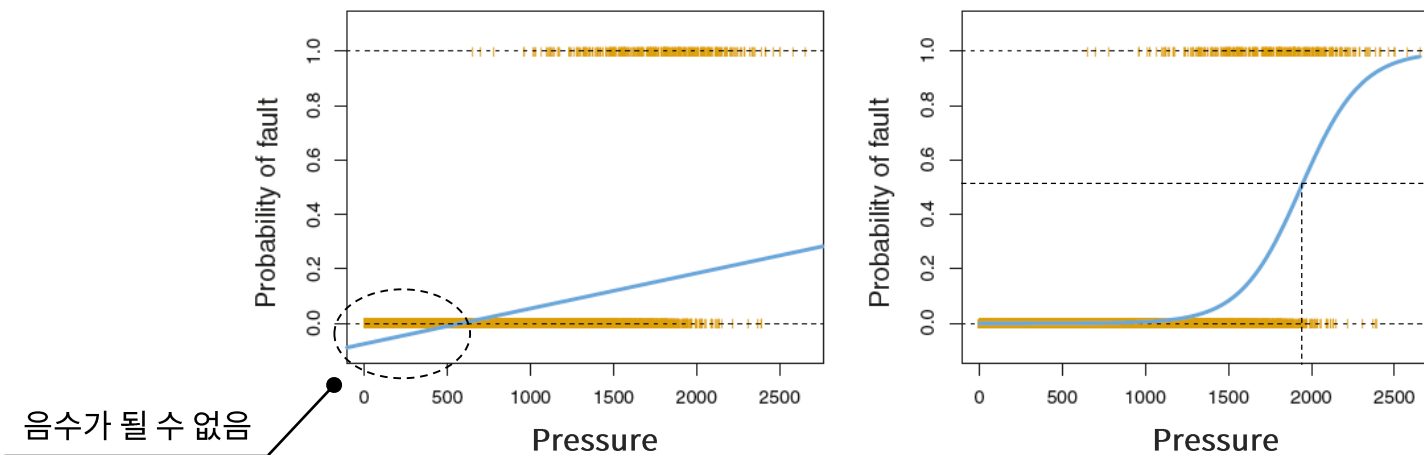


출처: <http://strijov.com/sources/demoDataGen.php>

I 로지스틱 회귀분석

■ 로지스틱 회귀 예제

- 다음의 예제는 반도체 공정에서 특정 변수 값을 가지고 정상이나 불량이나를 분류하는 문제임
- 이 때 Y는 '정상(0)', '불량(1)' 2가지 범주를 갖는 출력 변수이며, 입력 변수로는 반도체 공정의 특정변수(pressure)를 사용하여 예측한 예제임
 - 좌측 그림: 단순선형회귀를 이용했을 때, Pressure에 따른 불량 확률(불량 확률이 음수가 될 수 없음)
 - 우측 그림: 로지스틱회귀를 이용했을 때, Pressure에 따른 불량 확률
- Pressure 값이 약 2000 정도일 때 불량 확률이 0.5가 되므로 2000을 기준으로 불량을 결정할 수 있음



Part.02
회귀분석

로지스틱 회귀분석2 (회귀계수)

FASTCAMPUS
ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택