Part. 02

회귀분석

| 로지스틱 회귀분석2 (회귀계수)

FASTCAMPUS ONLINE

머신러닝과 데이터분석 A-Z

강사. 이경택

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

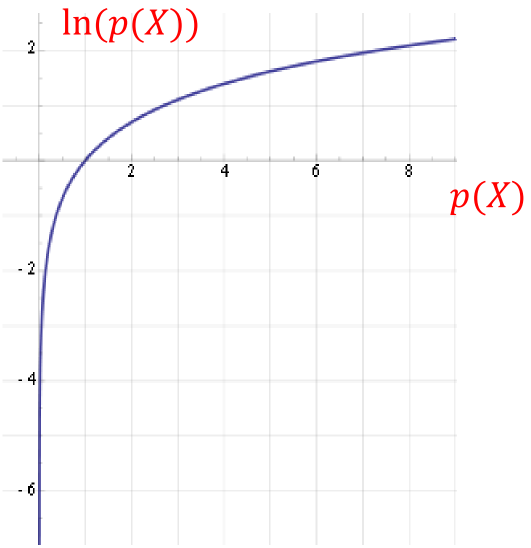
$$ln(p(X)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

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

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

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

도박에서
승률을 의미



- X는 입력변수, Y는 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 회귀계수 추정

- 단순(다중)선형회귀의 최소제곱법을 사용하는 것이 아닌 최대우도법(maximum likelihood)를 사용
- Likelihood function은 아래와 같고, 이를 최대화하는 β_0, β_1를 추정
- 베르누이 확률분포(0 또는 1의 값을 가지는 확률 변수의 확률 분포)를 이용하여 추정

$$Maximize_{\beta_0,\beta_1} \quad l(\beta_0, \beta_1) = \prod_{i=1}^{N} \theta_i(x_i)^{y_i}(1-(\theta_i))^{1-y_i}$$

- 위를 실제 사례(앞의 그림)에 적용하면 아래의 표와 같은 결과가 도출
- $\hat{\beta}_0$과 $\hat{\beta}_1$ 모두 유의하였고, Pressure가 1 증가할 때마다 logit이 0.0055 증가

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
Pressure	0.0055	0.0002	24.9	<0.0001

I 로지스틱 회귀분석

▪ 로지스틱 회귀계수 추정

$$Maximize_{\beta_0,\beta_1} \quad l(\beta_0,\beta_1) = \prod_{i=1}^{N} \theta_i(x_i)^{y_i}(1-(\theta_i))^{1-y_i}$$

$$\log(\, l(\beta_0,\beta_1)) = \sum_{i=1}^{N} y_i \log \theta_i(x_i) + (1-y_i)\log(1-\theta_i(x_i)))$$

$$\log(\, l(\beta_0,\beta_1)) = \sum_{i=1}^{N} (y_i \log \frac{1}{1+\exp(-w^T x_i)} + (1-y_i)\log(\frac{\exp(-w^T x_i)}{1+\exp(-w^T x_i)}))$$

$$\frac{\partial \log(\, l(\beta_0,\beta_1))}{\partial w} = \sum_{i=1}^{N} (y_i \frac{1}{\theta_i(x_i;w)} + (1-y_i)\ \frac{1}{1-\theta_i(x_i;w)})\frac{\partial\theta}{\partial w}$$

$$= \sum_{i=1}^{N} (y_i(1-\theta_i(x_i;w)) + (1-y_i)\,\theta_i(x_i;w))\, x_i$$

$$= \sum_{i=1}^{N} (y_i - \theta_i(x_i;w))\ \ x_i$$

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = w^T x$$

$$\theta(x) = \frac{1}{1+\exp(-w^T x)}$$

$$\frac{\partial\theta}{\partial w} = \theta(1-\theta)x$$

I 로지스틱 회귀분석

▪ 다중 로지스틱회귀 예제

- 단순선형회귀와 마찬가지로 로지스틱회귀도 입력 변수가 여러 종류일 때로 확장이 가능
- 입력 변수가 하나일 때와 마찬가지로 최우추정법을 이용하면 회귀계수의 추정이 가능

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- 웨이퍼가 불량일 확률에 영향을 주는 요인이 RF_impedance의 특정 summary 변수와 CL2 Flow 특정 summary 변수가 추가되었을 때 다중 로지스틱회귀를 적용하면 아래 표와 같은 결과가 도출
- RF_impedance와 Pressure가 유의한 입력 변수였으며, RF_impedance는 값이 높아질수록 불량일 확률(실제로는 logit)이 낮다는 결과가 도출

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
Pressure	0.0057	0.0002	24.74	<0.0001
CL2 Flow	0.0030	0.0082	0.37	0.7115
RF_impedance	−0.0468	0.0172	−2.74	0.0062

I 로지스틱 회귀분석

▪ 다중 로지스틱회귀 예제

- Logit으로 해석하는 방법과 odds로 해석하는 방법이 존재

RF_impedance는 값이 높아질수록 불량일 확률(실제로는 logit)이 낮다는 결과가 도출

Logit : RF_impedance가 1단위 증가할때 불량일 logit이 -0.0468단위 증가한다.

Odds : RF_impedance가 1단위 증가할때 불량일 확률이 0.954배(exp(-0.0468)) 증가한다

$$odds = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

도박에서
승률을 의미

$$logit = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

회귀식을 선형으로
변환하는 함수

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
Pressure	0.0057	0.0002	24.74	<0.0001
CL2 Flow	0.0030	0.0082	0.37	0.7115
RF_impedance	−0.0468	0.0172	−2.74	0.0062

Part. 02

회귀분석

| 회귀계수 축소법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택