

로지스틱 회귀 (logistic regression)

로지스틱 함수의 이해

종속변수 Y 가 '범주형 변수' 일 경우. 선형회귀의 '정확성'을 따진다'는 가정을 취한
그래서 회귀식을 적용하되 좌식을 Y 가 범주에 해당할 확률이라고 두는 시작한다.

$$P(Y=1 | X=\vec{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \vec{\beta}^T \vec{x}$$

↳ 0~1

이 때 좌식은 확률이기 때문에 그 범위가 0~1 의 구간과 맞지 않는다.

일반적 범위가 일치할 수 있도록 식을 조정하면

$$\frac{P(Y=1 | X=\vec{x})}{1 - P(Y=1 | X=\vec{x})} = \vec{\beta}^T \vec{x}$$

↳ 0~∞


$$\log \frac{P(Y=1 | X=\vec{x})}{1 - P(Y=1 | X=\vec{x})} = \vec{\beta}^T \vec{x}$$

↳ -∞~∞

그렇다면 계수 추정값 β_m 의 계수가 a 라면, β_m 이 1증가하면 $\log \frac{P(x)}{1-P(x)}$ 는
 a 만큼 증가하게 된다.

$$\log \frac{P(x)}{1-P(x)} = a \Leftrightarrow \frac{P(x)}{1-P(x)} = e^a$$

$$\Leftrightarrow P(x) = \frac{1}{1 + e^{-a}}$$

회음함수의 형태는  로지스틱 함수의 형태를 띄어 로지스틱 회귀로 불린다.

계수 해석

$\frac{P(x)}{1-P(x)}$ 를 Odds 라고 부른다. 이는 양의 사건 A 가 $\frac{\text{발생할 확률}}{\text{발생하지 않을 확률}}$ 을 뜻한다. β_m 의 계수가 a 이면 β_m 이 1증가할 때 Odds는 e^a 만큼 증가하고,

$\text{Logit} = \log(\text{Odds})$ 는 a 만큼 증가한다.

$$\begin{cases} a \rightarrow +\infty & p \rightarrow 1 & \therefore \text{영향력 있음 (범주 1에 속할 확률)} \\ a \rightarrow 0 & p \rightarrow 0.5 & \therefore \text{영향력 없음} \\ a \rightarrow -\infty & p \rightarrow 0 & \therefore \text{영향력 있음 (범주 0에 속할 확률)} \end{cases}$$

\therefore 계수 a 의 절대값의 크기가 클수록 영향력 있는 변수라고 볼 수 있다.

반대로 0에 가까운 수를 영향력 없는 변수라고 할 수 있음.

이항 로지스틱 회귀의 선형성

$$P(Y=1 | X=\vec{x}) > P(Y=0 | X=\vec{x})$$

위식에서 전방항과 후방항으로 분류하면,

$$P(x) > 1 - P(x)$$

$$\log \frac{P(x)}{1-P(x)} > 0$$

$$\vec{\beta}^T \vec{x} > 0$$

즉, 선형경계는 $\vec{\beta}^T \vec{x} = 0$ 인 hyperplane 이 된다.