Ⅱ. 일반 회귀분석 (Regression)

❹ Logistic

Regression의 확장

- **변수의 수가 너무 많으면? 가령, N < P 이면?**
  -> 고차원 회귀모형, 변수선택, 축소추정법
  ex) Ridge regression, Lasso regression, ElasticNet

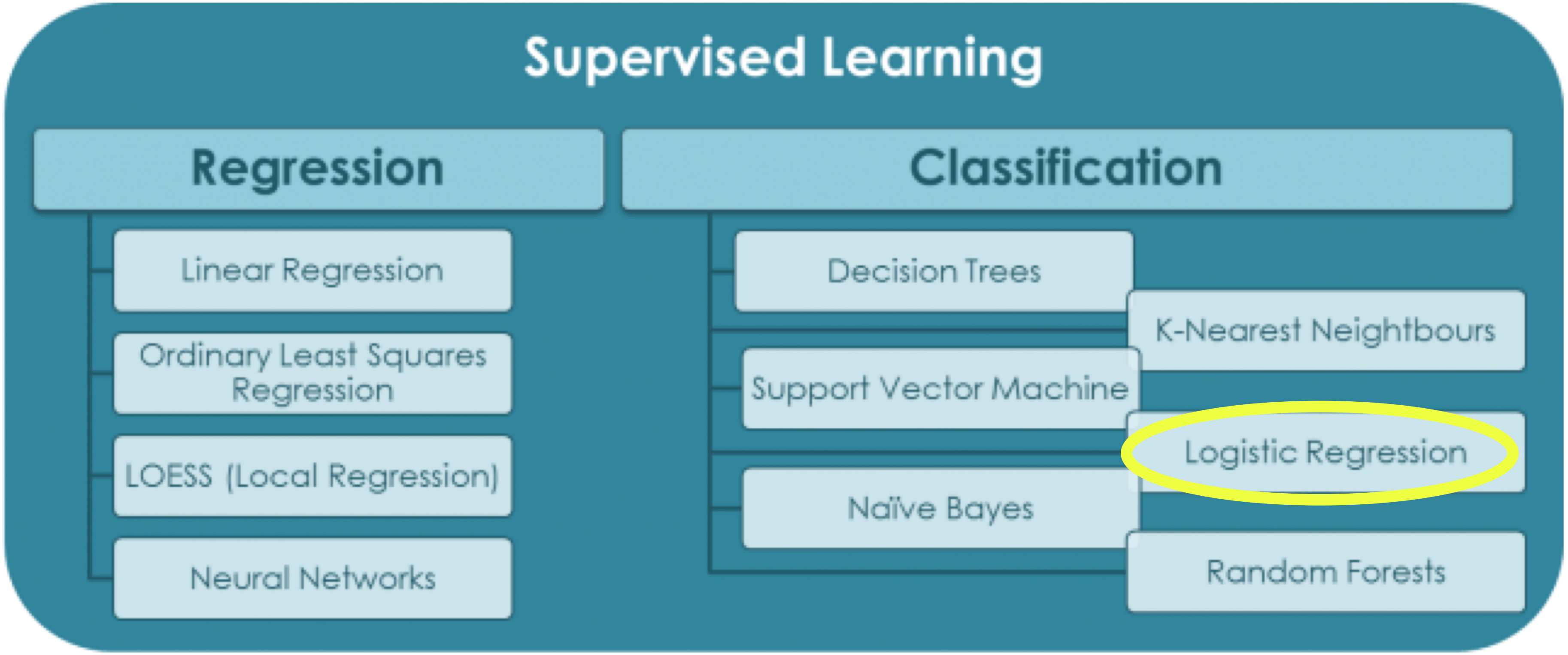
- **종속변수가 이산형은 아닌데... 연속형도 아니라면 ? 가산형(Count) 데이터라면?**
  -> 포아송 모형, 포아송 허들모형

- **종속변수가 연속형이긴 한데... 짤린 데이터라면 ? Ex) 비율데이터 등**
  -> 토빗 모형, Truncated model

- **종속변수가 이산형(Discrete)이면?**
  -> 로지스틱 회귀모형, 로짓모형, 프로빗 모형

- **종속변수가 이산형(Discrete)인데, 선택지가 여러 개면 ?**
  -> 다항 로짓모형, 혼합 로짓모형

- **종속변수가 이산형(Discrete)에 선택지가 여러 개인데, 각 소비자마다 계수가 다르면?**
  -> Heterogeneity 존재
  -> Mixture model, Gaussian model, Latent class(finite support) model


Supervised Learning
Regression
Classification
Linear Regression
Ordinary Least Squares Regression
LOESS (Local Regression)
Neural Networks
Decision Trees
K-Nearest Neightbours
Support Vector Machine
Logistic Regression
Naïve Bayes
Random Forests

로지스틱 회귀분석의 동기(Motivation)

<u>이 고객이 우리 회사 제품을 구매할 확률이 어떻게 될까?</u>

로지스틱 회귀모형(Logistic Regression)

- 로지스틱 회귀분석은 선형회귀분석과 마찬가지로 예측변수와 결과변수 사이의 관계를 특정하는 모형임

- 다만, 로지스틱 회귀분석은 결과변수인 Y가 연속형(Continuous)가 아니라 **범주형 (Categorical)인 경우의 문제**를 푸는 데 사용됨

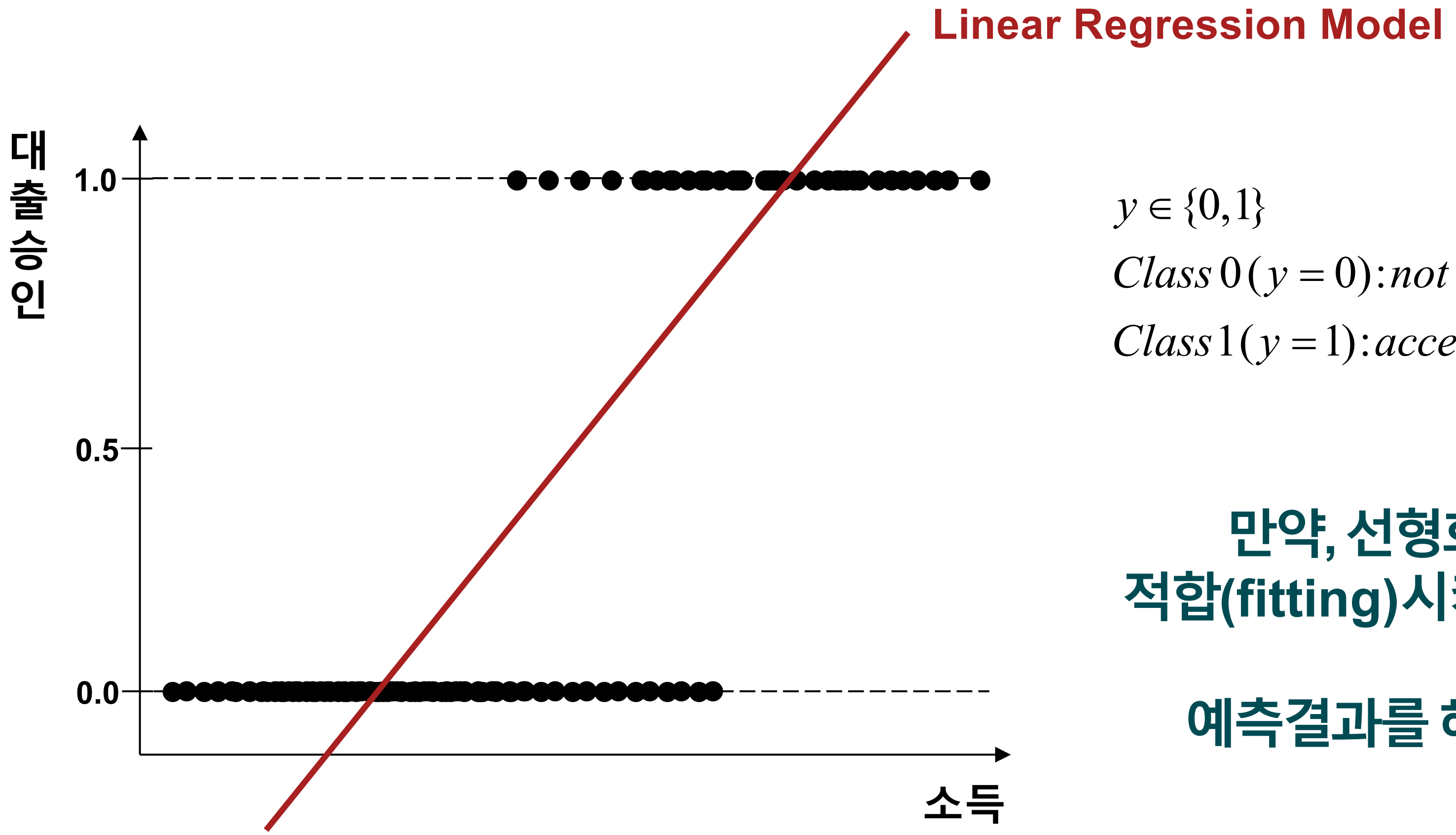
- 로지스틱 회귀모형은 사건이 발생할 '확률(Probability)'를 도출함으로써 **"발생" 또는 "미발생" 의 경우로 분류**할 수 있음.

- 로지스틱 회귀모형은 새로운 관측치가 어떤 클래스에 속할 지를 예측하기 위해 각 **클래스에 속할 성향(=확률)을 계산**해 해당 클래스로 분류하는 데 많이 활용됨. 따라서, 로지스틱 회귀모형은 분류(Classification) 문제에서 매우 다양하게 적용되고 있음

- 로지스틱 회귀모형에서 알아야 할 주요 개념은 **"오즈(Odds)"** 와 **"로짓(Logit)"**임

로지스틱 회귀모형(Logistic Regression) 원리

- 선형회귀분석이 연속형(Continuous) 종속변수를 예측한다면, 로지스틱 회귀분석은 **이산형(Discrete) 변수인 종속변수를 예측**함. 가령, E-mail이 스팸인지 아닌지 혹은 은행고객에게 대출을 할지 말지 등을 결정함



$$y \in \{0,1\}$$

$Class\,0\,(y=0): not\ accepted$ (대출 미승인)

$Class\,1\,(y=1): accecpted$ (대출 승인)

만약, 선형회귀모형으로 적합(fitting)시키면 어떻게 될까?

예측결과를 해석할 수 없음 !

로지스틱 회귀모형(Logistic Regression) 원리

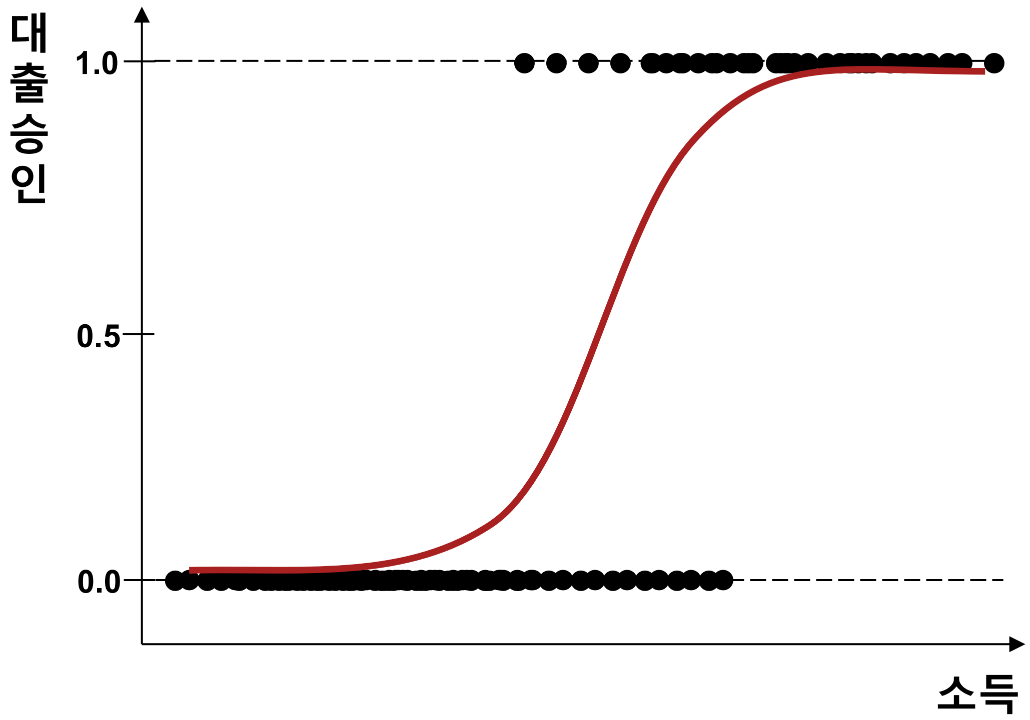
- 선택지가 0 또는 1인 경우, OLS를 이용한 선형회귀모형으로 추정해보자. 어떤 문제가 생기는가?
- 선형회귀모형으로 추정한 $\hat{y}$이 0 또는 1 의 값을 갖지 못하는 문제가 발생함

선형회귀모형으로 추정 : $\hat{\boldsymbol{y}} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_1 + \cdots + +\boldsymbol{\beta}_q \boldsymbol{x}_q$ **(X)**

로지스틱 모형 : $\mathbf{0} < \quad \hat{\boldsymbol{y}} = \boldsymbol{Pr}(\boldsymbol{y} = \mathbf{1} | \boldsymbol{x} = \dfrac{\boldsymbol{e}^{(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_1 + \cdots + +\boldsymbol{\beta}_q \boldsymbol{x}_q)}}{\mathbf{1} + \boldsymbol{e}^{(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_1 + \cdots + +\boldsymbol{\beta}_q \boldsymbol{x}_q)}} \quad < \mathbf{1}$ **(O)**

로지스틱 회귀모형(Logistic Regression) 원리

로지스틱 모형에서는 분석결과가 **"사건이 발생할 확률"**로 도출이 되도록 함수를 만들어 추정함. 따라서, 로지스틱 모형의 결과는 확률로 주어짐



Logistic Regression Model

$$0 < f(x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1}} < 1$$

Class 1(대출 승인)에 속할 확률과
Class 0(대출 미승인)에 속할 확률을 예측 !

예측값(Predicted value) ≥ 0.5 이면, Class 1에 분류
예측값(Predicted value) < 0.5 이면, Class 0에 분류

로지스틱 모형 유도

➢ 로지스틱 회귀모형을 오즈비(Odds ratio)와 로짓함수(Logit function)을 이용해 유도해보자.

$$\boldsymbol{Odds\,ratio} = \frac{\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})}{\mathbf{1}-\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})}$$

: 어떤 사건이 일어나지 않을 확률 대비 일어날 확률
Pr이 1에 가까울수록 오즈비는 무한대(∞), 0에 가까울수록 오즈비는 0에 가까워짐

양변에 Log

$$\boldsymbol{log}(\boldsymbol{odds}) = \boldsymbol{log}\left(\frac{\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})}{\mathbf{1}-\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})}\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{x}_1 + \cdots + +\boldsymbol{\beta}_q\boldsymbol{x}_q$$

: 이제, log(odds)값이 - ∞에서 + ∞값을
가지므로 선형회귀모형식을 가져올 수 있음

"Logit" 변환

양변에 Exp

$$\frac{\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})}{\mathbf{1}-\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x})} = \boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q}$$

$$\Leftrightarrow \quad \Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x}) = (\mathbf{1}-\Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x}))\boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q}$$

$$\Leftrightarrow \quad \Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x}) + \Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x}) * \boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q} = \boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q}$$

$$\Leftrightarrow \quad \Pr(\boldsymbol{y}=\mathbf{1}|\boldsymbol{x}) = \frac{\boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q}}{\mathbf{1}+\boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q}} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-(\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x}_1+\cdots+\boldsymbol{\beta}_q\boldsymbol{x}_q)}}$$

: Logistic Model

혼동행렬(Confusion Matrix)

혼동행렬(Confusion Matrix)는 지도학습 분류기법의 성능을 검증하기 위한 척도(Measure)로 혼동행렬을 통해 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 정밀도(Precision)을 측정할 수 있음

		실제 결과 (Actual)	
		참 (TRUE)	거짓 (FALSE)
분류예측 결과 (Predicted)	참 (TRUE)	TP (True Positive)	FP (False Positive)
	거짓 (FALSE)	FN (False Negative)	TN (True Negative)

“ Type I Error ”

“ Type II Error ”

Type I error vs. Type II error

실제 환자가 암 환자인데, 진료 결과 암 환자가 아니라고 분류하는 경우
VS
실제 환자가 암이 아닌데, 진료 결과 암 환자라고 분류하는 경우

		실제 결과 (Actual)	
		참 (TRUE)	거짓 (FALSE)
분류예측 결과 (Predicted)	참 (TRUE)	TP (True Positive)	**FP (False Positive)**
	거짓 (FALSE)	**FN (False Negative)**	TN (True Negative

“ Type I Error ”

“ Type II Error ”

분류모형 검증 지표(Index)

구분	정의	측정
정확도 (Accuracy)	전체 예측결과 중 올바르게 예측한 것의 비율	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ $Errorrate = \frac{FP+FN}{TP+TN+FP+FN} = 1 - Accuracy$
민감도 (Sensitivity)	실제로 참(True)인 것 중에서 참(True)으로 분류한 비율	$Sensitivity = \frac{TP}{TP+FN}$
특이도 (Specificity)	실제로 거짓(False)인 것 중에서 거짓(False)으로 분류한 비율	$Specificity = \frac{TN}{TN+FP}$
정밀도 (Precision)	참(True) 이라고 예측한 것 중에 실제로 참(True)인 비율	$Precision = \frac{TP}{TP+FP} = Positive\ Predict\ Value$