

# organized\_02

Hyeonho Lee

2018년 10월 7일

## Contents

타겟마케팅	2
로지스틱회귀모형 . . . . .	2
모형선택 . . . . .	3
모형평가 . . . . .	3
기계학습 방법론 : 고차원회귀모형 . . . . .	3
기계학습 방법론 : 의사결정나무 및 앙상블 . . . . .	3

# 타겟마케팅

## 로지스틱회귀모형

### 1 로지스틱 회귀모형

1. 출력변수가 범주형 변수인 경우, 분류문제에 사용하는 대표적인 회귀모형
2. 범주가 2가지인 경우를 고려하자
  - 1)  $Y = 1$  : 출력변수가 첫 번째 범주에 속할 경우
  - 2)  $Y = 0$  : 출력변수가 두 번째 범주에 속할 경우
3. 목적 : 입력변수와 범주형인 출력변수간의 관계를 잘 표현할 수 있는 모형 구축

### 2 모형

1.  $P(Y = 1|X) = F(X^T \beta)$
2.  $F(x)$  는 연속이고 증가하며 0과 1사이에서 값을 갖는 경우

### 3 여러 가지 $F(x)$

1. 로지스틱 모형 :  $F(x) = \exp(x)/(1 + \exp(x))$
2. 공배르츠 모형 :  $F(x) = \exp(-\exp(x))$
3. 프로빗 모형 :  $F(x)$ 가 표준정규분포의 분포함수 (distribution function)
4. 이중 로지스틱 모형이 계산의 편의성으로 가장 널리 사용된다!!!
5. 로지스틱 회귀모형
  - 1)  $P(Y = 1|X = x) = \frac{\exp(x^T \beta)}{1 + \exp(x^T \beta)}, ((\beta = (\beta_1, \beta_2, \dots, \beta_p)^T)$
  - 2) i.e  $\log\left(\frac{p(Y=1|X=x)}{p(Y=0|X=x)}\right) = x^T \beta$

### 4 모수의 추정(최대 우도 추정)

1. 모수 :  $\beta$
2. 자료 :  $(y_1, x_1), \dots, (y_n, x_n)$
3. 최대 우도 추정량(Maximum likelihood estimator)  $\hat{\beta}$ 
  - 1) 우도 함수를 최대화 하는 모수값
4. 우도함수
  - 1)  $L(\beta) = \prod_{i=1}^n F(x^T \beta)^{y_i} \times (1 - F(x^T \beta))^{1-y_i}, \text{ where } F(x) = \frac{\exp(x)}{1 + \exp(x)}$
  - 2) 우도함수의 최대화는 수치적 방법을(numerical method)를 사용하여 구한다.  
예시) Newton-raphson method

### 5 예측 및 모형의 해석

1. 예측
  - 1)  $\hat{P}(Y = 1|X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \times x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \times x)}$
  - 2)  $\hat{P}(Y = 1|X = x) > 0.5$  이면 1로 예측
  - 3)  $\hat{P}(Y = 1|X = x) < 0.5$  이면 0으로 예측
2. 해석
  - 1)  $\beta_1 > 0$  :  $x$ 가 증가하면  $P(Y = 1|X = x)$ 가 증가한다.
  - 2)  $\beta_1 < 0$  :  $x$ 가 증가하면  $P(Y = 1|X = x)$ 가 감소한다.

## 6 회귀계수와 오즈비

### 1. 오즈(odds)

$$P(Y = 1|x)/P(Y = 0|x)$$

### 2. 오즈비(odds ratio)

$$\frac{P(Y = 1|x + 1)P(Y = 0|x)}{P(Y = 0|x + 1)P(Y = 1|x)}$$

### 3. 성질

$$\text{오즈비} = \exp(\beta)$$

## 7 오즈비의 의미

1. X가 한 단위 증가 할 때 y=1일 확률과 y=0일 확률의 비가 증가하는 양

2. 예시

1) x는 소득이고 y는 어떤 상품에 대한 구입여부(1=구입, 0=미구입)

2) b=3.72

3) 소득이 한 단위 증가하면 물품을 구매하지 않을 확률에 대한 구매할 확률의 비(오즈비)가  $\exp(3.72) = 42$ 배 증가함을 의미한다.

## 8 불균형 자료 분석방법

1. 많은 분류문제에서 모집단에서 두 그룹의 크기가 현저히 다른 경우가 종종 발생한다. 예제) 부도예측, FDS(Fraud Detection System), 이탈방지, 암진단

2. 모집단이 불균형이 된 경우, 임의추출법으로 자료를 구성하면, 작은 그룹의 자료의 수가 매우 작을 수 있어서 분석에 많은 문제가 생김(ex. 파워가 너무 작다)

3. 이런 경우에는 임의추출법을 사용하지 않고 흔히 case-control sampling을 사용

4. case-control sampling은 역학에서 주로 사용되는 방법임.

## 모형선택

## 모형평가

기계학습 방법론 : 고차원회귀모형

기계학습 방법론 : 의사결정나무 및 앙상블