organized_02

Hyeonho Lee

2018년 10월 7일

Contents

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

4 모수의 추정(최대 우도 추정)

1. 모수 : β
2. 자료 : $(y_1, x_1), ..., (y_n, x_n)$
3. 최대 우도 추정량(Maximum likelihood estimator) $\hat{\beta}$
   1) 우도 함수를 최대로 하는 모수값
4. 우도함수
   1) $L(\beta) = \prod_{i=1}^{n} F(x^T\beta)^y \times (1-F(x^T\beta))^{1-y_i}, \ where F(x) = \frac{exp(x)}{1+exp(x)}$
   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) $\beta_1 > 0$: x가 증가하면 $P(Y=1|X=x)$가 증가한다.
   2) $\beta_1 < 0$: x가 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) = 42배 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 방법이다.
5. 일반적으로 case-control sampling는 A그룹과 B그룹에 대해서 y에 영향을 미치는지에 대해서 알고 싶은데, y가 너무 작은 나머지 sampling과정에서 y가 뽑히지 않은 확률이 매우 높다.
    1) 두 그룹 분류문제의 경우 첫 번째 그룹에서 n1명의 자료를 임의추출법을 사용하여 추출하고 두 번째 그룹에서 n2명의 자료를 임의 추출한다. 그리고 두 자료를 모두 사용하여 분석한다.
    2) case-control sampling으로 추출된 자료는 원자료와 그 분포가 매우 상이하여 분석 결과의 해석에 매우 조심하여야 한다.
    3) P(x)를 원자료에서 입력변수 x가 주어진 경우의 자료가 두 번째 그룹에 속할 확률이라 하자. (1-P(x)는 입력변수 x가 주어진 경우의 자료가 첫 번째 그룹에 속할 확률이다.)
    4) Q(x)는 사후추출법으로 추출된 자료에서 입력변수 x가 주어진 경우의 자료가 두 번재 그룹에 속할 확률이라 하자.(1-Q(x)는 입력변수 x가 주어진 경우의 자료가 첫 번째 그룹에 속할 확률이다.)
    5) 일반적으로 P(x)와 Q(x)는 매우 다르다. case-control sampling으로 추출된 자료를 이용하면 확률 Q(x)를 추정하게 된다. 하지만 관심사는 원자료에서의 확률 P(x)이다. 원자료에서 두 그룹의 분포를 아는 경우에는 이 정보를 이용하여 Q(x)로 부터 P(x)를 구할 수 있다. 많은 분류문제에서는 P(x)의 정확한 값보다는 두 개의 주어진 입력변수 x1과 x2에 대하여 P(x1)과 P(x2)의 대소를 비교하는 것을 목적으로 한다.
    6) 원자료의 정보가 없는 경우에도, Q(x)를 이용하여 P(x)의 대소를 알 수 있다. Q(x)와 P(x)사이에 다음의 관계가 성립된다. 주어진 두 개의 입력변수 x1과 x2에 대하여 만약 Q(x1)<Q(x2)이면 P(x1)<P(x2)를 만족한다.
    7) 로지스틱 회귀모형과 case-control sampling는 가능하다.

모형선택

모형평가

기계학습 방법론 : 고차원회귀모형

기계학습 방법론 : 의사결정나무 및 앙상블