10장 로지스틱 회귀분석

1. 서론

로지스틱 회귀 - 선형회귀의 개념을 종속변수 Y가 범주형인 경우로 확장한 것

↳ 예측변수의 값을 바탕으로 클래스가 알려져 있지 않은 관측치들을 분류

↳ 클래스가 알려져있는 예측변수들에 대해 서로 다른 클래스로 구분해주는 모델을 훈련시킬 수도 있음 (로그 파일링)

로지스틱 회귀는 다양한 분야에서 범주형 반응변수를 설명하거나 예측하기위해 구조화된 모델이 필요할 때 사용

다중선형회귀분석은 연속형 반응변수 Y의 값을 예측하는 것인 반면에 로지스틱 회귀분석은 분류가 목적

로지스틱 회귀는 2단계로 구성

① 각 클래스에 속하는 '성향' 혹은 '확률'을 추정

② 각 관측치의 클래스를 지정하기 위해 확률값에 대한 컷오프값을 적용

2. 로지스틱 회귀모델

로지스틱 회귀의 원리 : 종속변수 Y를 대신해서 log-odds(로짓) 이라고 부르는 Y의 함수를 사용

↳ log-odds을 예측변수들의 선형함수로 모형화. 일단 로짓이 예측되면 그로부터 확률 맵핑 가능

log-odds의 이해

① 클래스 1에 속할 확률 $p = P(Y=1)$을 구한다. 클래스 0에 속할 확률은 $1-p$가 된다.

0, 1 두 개의 값만 가질 수 있는 Y와 달리 p는 구간 [0,1]에 존재하는 모든 값을 가질 수 있다.

↳ 그러나 확률 p를 우개의 예측변수들에 대한 선형함수로 나타내면

$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$ → 우변이 구간 [0,1]에 들어가는 것을 보장 X

↳ 구간 [0,1]을 보장하게 위해 $p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q)}}$ 을 사용

↳ 로지스틱 반응함수

② 클래스 분류와 연관된 다른 척도인 odds(오즈)를 이해

클래스 1(Y=1)에 속할 오즈는 '클래스 0에 속할 확률에 대한 클래스 1에 속할 확률의 비'

$\text{Odds}(Y=1) = \frac{p}{1-p}$ → 병에 걸릴 확률이 0.5일 때, 병에 걸릴 오즈는 $\frac{0.5}{1-0.5} = 1$

↳ odds로 부터 확률 계산 가능 $p = \frac{\text{odds}}{1+\text{odds}}$

예측변수와 오즈 사이의 관계 — $\text{Odds}(Y=1) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q}$

↳ 예측변수 X_j가 한 단위 증가하면 다른 예측변수들이 모두 일정하다고 가정할 때 odds는 e^{β_j} 배를 증가

이 식에 자연로그를 취하면 $\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$

이 때 log(odds)로 log를 미분으로 하며 $-\infty$ ~ ∞ 사이의 값을 가질 수 있다.

↳ log값 = 0 이면 odds = 1 (확률 0.5)

반응변수와 예측변수의 관계를 나타내는 최종형태는 log가를 종속변수로 하여 여러개의 예측변수에 대한 선형함수로 모형화

3. 예제: 개인대출채환 수락

- 단일 예측변수를 갖는 모델

하나의 예측변수 X와 반응변수 Y의 관계를 직선으로 적합시킨 단순 선형회귀모델과 개념적으로 비슷.

ex) 예측변수로 '수입'만을 고려

$$P(\text{대출} = \text{Yes} \mid \text{Income} = X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}, \quad \text{Odds}(\text{대출} = \text{Yes}) = e^{\beta_0 + \beta_1 X}$$

로지스틱 회귀는 범주형 반응의 확률을 예측하게 해준다는 점에서 예측문제에 사용되기도 있지만 대개 분류문제에 사용 → 확률함수로서 표현 가능. 컷오프값 혹은 오즈 이용해서 분류 가능

- 데이터로부터 로지스틱 모델 추정: 모수 추정치의 계산

로지스틱 회귀에서 Y와 모수 β의 관계는 비선형 → 다중 선형회귀분석에서 최소제곱법 사용 X

↳ 최대 가능도법 (Maximum likelihood method) 사용하여 추정

↳ 우리가 가진 데이터를 얻을 가능성을 최대화하는 추정치를 찾는 방법 → 컴퓨터를 이용해 반복추정하는 과정 필요

〈최대 가능도법〉

최대가능도법은 추정치에 대한 좋은 점근적(대표본) 성질을 보장. 매우 일반적인 조건 하에서 다음을 만족

① 일치성 (consistent) - 표본크기가 증가함에 따라 추정량과 실제값이 다를 확률이 0에 수렴.

② 점근적 효율성 (asymptotically efficient) - 일치성을 만족하는 추정량 중에 가장 작은 분산을 갖는다.

③ 점근적 정규분포 (asymptotically normally distribution) - 표본크기가 클 때, 다중선형회귀모델분석과 유사한 방식으로 신뢰구간을 계산하고 통계적 검정을 수행할 수 있다

회귀계수 추정치를 계산하는 알고리즘은 선형회귀보다 덜 강건

그러나 데이터에서 반응변수로 0, 1을 갖는 관측치가 많을 때나 값이 0, 1 들 중 무엇과도 심하게 가깝지 않을 때, 로지스틱 회귀모델에서 회귀계수의 개수가 표본의 크기에 비해 작을 때 (10% 이하)는 일반적으로 회귀계수의 추정치 신뢰가능

선형 회귀와 마찬가지로 예측변수들 간의 강한 상관관계인 공선성 (collinearity)은 추정치 계산을 어렵게 만듦

- (프로파일링을 위한) 오즈관점에서의 결과 해석

데이터에 잘 맞는 로지스틱 모델은 서로 다른 예측변수들의 역할에 대한 유용한 정보 제공 가능

↳ 확률을 알아보는 것보다 오즈를 이용하는 것이 더 좋다.

$odds = e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}$ → 여기서 X_1이 한 단위 증가, 다른 변수값은 모두 고정이면

$$\frac{odds(X_1+1,\cdots,X_p)}{odds(X_1,\cdots,X_p)} = \frac{e^{\beta_0+\beta_1(X_1+1)+\cdots+\beta_pX_p}}{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}} = e^{\beta_1}$$

변수 X_1이 1단위 증가할 때 바뀌는 오즈의 양이 e^{β_1}

오즈로 결과를 보게 되면 X의 어떤 값에 대해서도 이런 해석이 가능하다는 것.

확률값은 $X: 3\to4$, $X: 30\to31$ 일 때 확률 p가 바뀌는 정도가 다르다.

4. 분류성능평가

가장 많이 사용되는 것은 정오분류표와 향상차트에 기반한 척도

로지스틱 회귀에서 정오분류표를 얻기 위해선 추정된 회귀식을 통해 클래스에 속할 성향을 예측.

그리고 클래스를 정하기 위해 컷오프값을 사용

↳ 확률은 $p = e^{\text{회귀식}} / 1+e^{\text{회귀식}}$ 의 관계식을 통해 얻을 수 있음.

- 변수선택

대체모델을 찾기

↳ 예측변수의 개수를 줄여 더 간단한 모델로 만들거나 예측변수들간의 상호작용을 고려하고

그로부터 파생되는 변수들을 고려하여 더 복잡한 모델로 만들거나

⇒ 대체모델의 선택은 성능에 근거 (검증 데이터에 대해서)

복잡한 모델보다는 간단한 모델

선형 회귀에서와 마찬가지로 단계적 선택, 전방선택, 후방소거 등 자동화된 변수선택 가능

6. 부록: 로지스틱 회귀 모자란점

- 선형 회귀가 범주형 반응변수에 대해 문제가 되는 이유

반응변수 Y를 연속형으로 간주하고 다중선형회귀 적용가능 → 선형확률모델

↳ 수치적 코드화 필요 ⇒ 이렇게 하면 몇 가지 이상한 점들이 존재

① 모델을 사용하여 Y값을 예측할 때, 반드시 0, 1이 나오지는 않는다

② 단차에 대한 히스토그램이나 확률플롯을 보면 결과변수(또는 단차)가 정규분포를 따른다는 가정 위배.

③ 모든 클래스에서 Y의 분산이 일정하다는 가정 성립 X. Y는 이항분포를 따르기 때문에 분산은 $np(1-p)$

④ 이 경우 분류 모델(특히 Y=1이 분류)의 경우에 부적합

①의 경우, [illegible] 부적합.

②,③의 경우 표준오차를 이용한 통계적 추론을 사용한 프로파일링에 문제.

⇒ 범주형 반응변수의 분류에 로지스틱 회귀를 쓰는 이유

· 설명력 평가

분석의 목적이 프로파일링인 경우 → 새로운 데이터의 분류보다 기존 데이터의 설명에 더 관심

↳ 모델이 데이터를 얼마나 잘 적합했는지를 알아보는 척도가 필요

◦ 전체적인 적합강도

모델의 전체적인 설명력 평가 필요

클래스 간 차이를 설명하는데 예측변수가 필요한가? ⇒ 이탈도 D

이탈도 D는 전체적인 적합도를 측정하는 통계량 – 최소제곱법에서 오차제곱합 개념과 유사

↳ 모델의 이탈도를 예측변수가 없는 나이브모델과 비교 ⇒ 통계적으로 의미가 있다면 예측변수 있는게 좋다.

· 단일 예측변수의 영향

로지스틱 회귀의 출력 : 각 예측변수 X_i에 대한 회귀계수 b_i와 표준편차가 있는 회귀계수표 제공

↳ p-값은 예측변수 X_i의 통계적 유의성을 나타내며, 낮은 p-값은 예측변수와 반응변수 간에 통계적으로

유의미한 관련이 있음을 나타내고, 이러한 관계가 우연이 아님을 나타냄

<중요한 사항들>

① 통계적 유의미 = 실질적 유의미는 아니다. → 예측변수의 영향력이 큰 것을 의미

② 모든 예측변수의 스케일이 동일하지 않는 한 계수의 크기나 오즈의 크기 비교는 무의미

↳ 각 계수에 예측변수의 값이 곱해지므로 계수들만 비교하는 건 의미가 없다

③ 통계적으로 유의미한 예측변수는 평균적으로 예측변수의 한 단위 증가가 결과에 미치는 어떤

특정 영향과 관련있다는 것을 의미. → 예측력을 나타내는 것은 아님

↳ 모델링이나 프로파일링에서는 통계적 유의성이 매우 중요하지만 분류에서는 이차적인 문제

Confusion matrix나 lift chart에 근거해서 변수선택 필요

· 두 개 이상의 클래스에 대한 로지스틱 회귀

클래스가 m일 때, m개의 확률합은 1이기 때문에 m-1개의 확률만 추정하면 됨

· 순서형 클래스

클래스에 의미있는 순서가 존재 ┬ 클래스의 수가 5이상이면 연속형으로 취급하여 다중선형회귀 가능

└ $3 \le m \le 5$ 일 때 로지스틱 회귀의 확장은?

비례오즈(혹은 누적로짓) 클래스 3개 1= 매수 2= 보유 3= 매도라 할 때,

$P(Y \le 1)$ - 매수를 추천할 확률 $P(Y \le 2)$ - 매수/보유를 추천할 확률

$P(Y=1) = P(Y \le 1)$, $P(Y=2) = P(Y \le 2) - P(Y \le 1)$ $P(Y=3) = 1 - P(Y \le 2)$

⇒ $\text{logit}(Y=1) = \log \frac{P(Y \le 1)}{1 - P(Y \le 1)}$ $\text{logit}(Y \le 2) = \log \frac{P(Y \le 2)}{1 - P(Y \le 2)}$

↳ 이런 식으로 각 클래스에 속할 확률 추정 가능

· 명목형 클래스

순서는 없고 단순히 서로 다른 클래스일 때 → m개가 있다면 m-1개의 확률만 추정

$$\text{logit}(A) = \log \frac{P(Y=A)}{P(Y=C)} = \alpha_0 + \alpha_1 X$$

$$\text{logit}(B) = \log \frac{P(Y=B)}{P(Y=C)} = \beta_0 + \beta_1 X$$

$\therefore P(Y=C) = 1 - (P(Y=A) + P(Y=B))$