4.1

분류 문제 예시

1. 응급실 환자의 증상들이 3가지 질병 중 어느 것과 연관이 있는지
2. 온라인 뱅킹에서 현재 거래가 사기인지 사기 아닌지
3. 어떤 돌연변이가 질병을 유발하는지

분류에서는 (x\_1, y\_1),...,(x\_n, y\_n) 이런 형태로 데이터 구성. 그냥 지도 학습.

4.2

뇌졸중, 약물과다 복용, 간질 발작 중에서 하나 예측하려면

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

이런식으로 만들어 놓고.

선형 회귀로 접근하면.

순서가 있는 것을 전제로 해서

2가 1이랑 3 사이에 있는 것처럼 해석될 위험 존재. 이렇게 되면 집단 간의 차이가 일정하다는 건데, 이 순서랑 간격이 이 문제에서 의미가 있나? 뭐 암튼 그럼.

폰트, 텍스트, 화이트, 디자인이(가) 표시된 사진

자동 생성된 설명

이렇게 이진 변수로 한다고 해도, Y^ > 0.5 식으로 임계점 두고 할 수도 있는데,

Pr(drug overdose | X) 이런식으로 근사치로 계산 됨. 근데 예측값 [0,1] 범위값 넘어가면? 답 없음.

요약하면

1. 두개 이상 클래스 가진거면 처리 불가능
2. 애초에 의미있는 추정을 못함.

그래서 이런 문제는 분류 방식을 적용하는게 적합.

4.3

|  |  |  |  |
| --- | --- | --- | --- |
| Default(카드값 연체여부) | Balance(월 카드 잔액) | income | student |
| No | 729.5 | 44361 | Yes |
| Yes | 817.2 | 31723 | No |
| No | 1073.2 | 33321 | No |

불편해서 대충 데이터 그려봄.

Pr(default = Yes|balance) 🡺 p(balance) 줄이면 이렇게 됨.

p(balance)>0.5 여기서 임계값은 0.5로 뒀는데, 이 임계값은 충분히 바꿔서 구성 가능. 0.1이든 0.6이든

4.3.1

선형회귀 문제점.

선형회귀에서는 p(x)를 모델링 할 때, p(X) = B\_0 + B\_1X 형태로.

확률은 0~1 사이값 나와야 하는데, p(x) < 0, p(x) >1 예측하는 것과 같은 문제 발생.

그래서 이런 문제에서는 쓰면 안됨.

로지스틱 회귀 모델.

텍스트, 폰트, 화이트, 친필이(가) 표시된 사진

자동 생성된 설명

이 식이 0~1 값으로 제한.

폰트, 화이트, 도표, 라인이(가) 표시된 사진

자동 생성된 설명 맨날 보는 이놈이랑 동일.

텍스트, 스크린샷, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명 그래프 형태는 이리 되고,

Odds(승산), log odds(로그 승산)

Odds 🡺 p / (1-p) # 사건이 일어날지 안 일어날지.

Ex) 사건 일어날 확률 0.2면, 0.2/0.8 = 0.25 🡺 사건 발생 확률, 안 일어날 확률보다 4배 낮음.

사건 일어날 확률 0.9면 9 🡺 일어날 확률 9배 높음.

Odds 값 1보다 크면 사건 일어날 확률 더 높음

1보다 작으면 사건 안 일어날 확률 더 높음 의미.

텍스트, 폰트, 화이트, 서예이(가) 표시된 사진

자동 생성된 설명

위의 시그모이드 함수 갖고 계산하면 이거 유도 됨.

여기다 로그 취하면

폰트, 텍스트, 화이트, 친필이(가) 표시된 사진

자동 생성된 설명

결국 선형 회귀에서 파생 되었다, 선형 회귀랑 같은 해석이 가능하다. 뭐 그런 의미.

4.3.2

최대우도 추정법(MLE)로

P^(x\_i)가 실제 Label 값이랑 최대한 일치하도록 B\_o, B\_1을 추정.

연체한 사람한테는 1에 가까운, 연체 안한 사람은 0에 가까운 확률값 내도록 B\_0, B\_1값들 조정.

폰트, 타이포그래피, 텍스트, 화이트이(가) 표시된 사진

자동 생성된 설명

이게 우도 함수

연체한 사람은 p(x\_i), 연체 안한 사람은 (1-p(x\_i)) 곱해서 각 값 확률값들 최대화 되게.

|  |  |  |  |
| --- | --- | --- | --- |
| Default(카드값 연체여부) | Balance(월 카드 잔액) | income | student |
| No | 729.5 | 44361 | Yes |
| Yes | 817.2 | 31723 | No |
| No | 1073.2 | 33321 | No |

요로코롬 있다고 하면,

B\_1 = 0.0055의 의미는 balance 1 증가할 때마다, log odds 0.0055 만큼 증가함을 의미.

즉, balance 높을수록 연체 확률 증가함 의미.

Z-statistic 값 크면 B\_1이 0이 아닐 가능성이 매우 높음.

P-value도 작으면 Label이랑 어떤 관계가 존재한다. 상관관계가 존재 한다. 그런 의미.

4.3.3

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명



뭐 잔액 1000원이다. 하면 x에 1000 들어가고, 쭉쭉 계산.

최종 값 보면 많이 낮음. 남은 돈 1000원이면 연체 가능성 매우 낮음 의미.

잔액 2000원이다 하면 값 0.586으로 올라가는데, 이는 연체 가능성 높아짐 의미.

Student 칼럼도 들어가는데, 이건 질적 변수.

그럼 더미 변수 만들어서 yes 면 1, no면 0으로 표시.

Table 4.2 껄로 대입해서 보면,

학생인 경우

학생 아닌 경우

결론은 학생이 좀 더 연체할 확률이 높다 의미.

4.3.4

폰트, 친필, 텍스트, 서예이(가) 표시된 사진

자동 생성된 설명

다중 로지스틱 회귀

아까까지는 단순 로지스틱 회귀. 변수 하나만 고려.

이거는 여러 개 고려 가능.

|  |  |  |  |
| --- | --- | --- | --- |
| Default(카드값 연체여부) | Balance(월 카드 잔액) | income | student |
| No | 729.5 | 44361 | Yes |
| Yes | 817.2 | 31723 | No |
| No | 1073.2 | 33321 | No |

결국에는 Balance, income, student 다 들어갈 수 있다는 거.

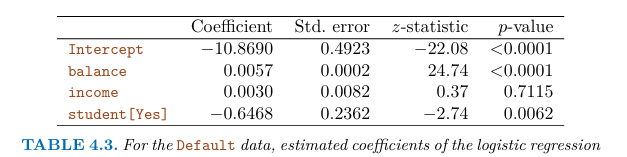


표 봐보면 balance랑 student[yes]랑 같이 p-value 값 작아서 상관관계 존재.

다중 로지스틱 회귀로 오게되면 한 변수에 따라 다른 변수에 영향 끼쳐서 값 다르게 나올 수도 있는데 이걸 confoundling(혼란) 현상이라 함.

Ex) 잔액 똑같은데, 학생이 학생 아닌 사람보다 더 낮은 연체 확률.

* 변수 하나만 쓸때랑 변수 여러 개 쓸때랑 효과 다르게 나타나는 것을 confounding 현상

 학생 경우

학생 아닌 경우

balance랑 income 똑같으면 학생이 연체비율 더 낮음. 뭐 암튼 변수 많아지면 달라질 수도 있다는 의미.

4.3.5

다중 클래스 🡺 Multinomial Logistic Regression으로.

클래스 3개 이상일 때,

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

이걸로 확률값 계산.

텍스트, 폰트, 라인, 화이트이(가) 표시된 사진

자동 생성된 설명

Log odds 봐보면, 선형적으로 나타나는게 보임.

기준 하나 잡아서 하라 했는데, 확률 결과값은 동일.

폰트, 화이트, 도표, 라인이(가) 표시된 사진

자동 생성된 설명결국 이거랑 똑같아서.

이게 softmax랑 동일. Multinomial logistic == softmax