로지스틱 회귀 분석

학부연구생 이현재

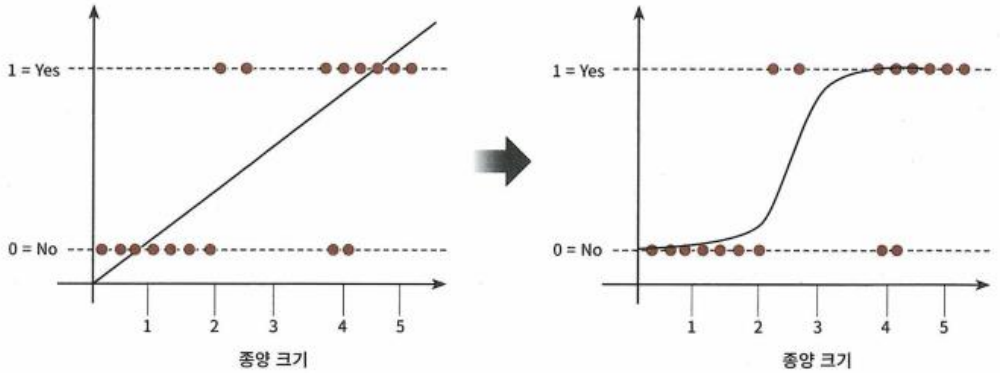
로지스틱 회귀 (Logistic Regression)

- 정의 : 선형 회귀 방식을 **분류**에 적용한 지도 학습 알고리즘

  → **시그모이드** 함수 최적선을 찾고 이 시그모이드 함수의 반환 값을 **확률**로 간주해 확률(0~1)

  에 따라 **분류**를 결정(**종속 변수가 범주형 데이터** 이며 **이진 분류**에 사용)

- 예시 : 종양의 크기에 따라 악성 종양인지 그렇지 않은지 회귀를 이용해 1과 0의 값으로 예측
  - 종양이 맞다면 Yes = 1이고 종양이 아니라면 No = 0의 값
  - 일반적인 선형 회귀(왼쪽 그림)은 0과 1을 잘 구분하지 못하지만 S자 커브 형태의 시그모이드 함수(오른쪽 그림)을 이용하면 더 정확하게 0과 1을 구분 가능.



$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}.$$

시그모이드 힘수

  - 일반적인 회귀(왼쪽 그림)에서 0과 1을 벗어나는 값을 가지는 것이 모순. 그래서 0부터 1까지의 값을 갖는 시그모이드 함수를 사용하여 분류

- 즉, 로지스틱 회귀는 이처럼 **선형 회귀 방식을 기반**으로 하되 **시그모이드** 함수를 이용해 **분류**를 수행하는 회귀

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

시그모이드 함수를 사용하는 이유

- 단순선형회귀분석에서는 $y = ax + b$를 이용하여 예측한다. 로지스틱 회귀에서는 선형 회귀를 기반으로 하기 때문에 y를 확률 P로 바꾸어 $P = ax + b$가 된다.

$$\underset{y = ax+b}{[-\infty, \infty]_{x} \Longleftrightarrow [-\infty, \infty]_{y}} \quad \Rightarrow \quad \underset{P = ax+b}{[-\infty, \infty]_{x} \Longleftrightarrow [0 \sim 1]_{y}}$$

→ 이렇게 식을 변환하기 위해 사용하는 것이 **odds**와 **log**

- Odds
  - 정의 : 실패에 비해 성공할 확률의 비

$$Odds = \frac{P(event\ occurring)}{P(event\ not\ occurring)} = \frac{p}{1-p}$$

  - p는 0에서 1사이 값을 가진다. p에 0을 대입하면 0/(1-0) = 0이고, p에 1을 대입하면 1/(1-1) = +∞이다. 즉, $\frac{p}{1-p}$는 x의 범위로 0부터 양의 무한대까지 값을 가지게 되므로 x의 도메인 조건에서 만족하지 못한다.

- Log Odds
  - 음의 무한대를 범위에 포함시키기 위해 자연로그를 취한다.(설명 참고)

$$Log\ Odds = \log_e \frac{p}{1-p} = ln \frac{p}{1-p}$$

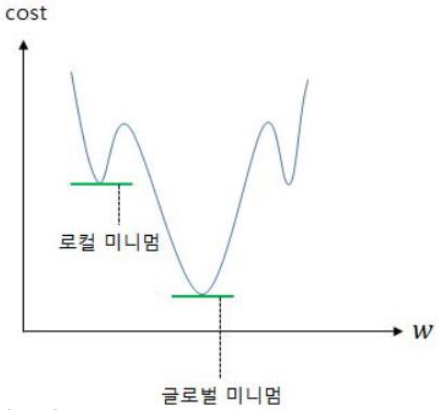
$\ln(p/1-p) = ax+b \Rightarrow p/1-p = e^{ax+b} \Rightarrow p = \frac{e^{ax+b}}{1+e^{ax+b}}$: Sigmoid.

- 시그모이드 함수를 사용하는 이유

  : 0 ~ 1인 확률을 표현하기 위하여 Log Odds를 적용하니 도출된 식이 시그모이드 함수. (시그모이드는 항상 0~1)
  시그모이드 함수를 사용하여 도출된 값이 임계점(=0.5, 조정가능)보다 높으면 1, 낮으면 0으로 분류

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

비용 함수 (Cost Function)



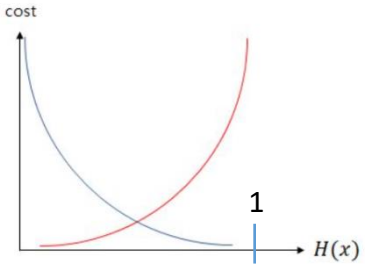
- 로지스틱 회귀에서 비용 함수로 평균 제곱 오차(RSS)를 사용하면 경사 하강법을 사용하였을 때 찾고자 하는 최소값이 아닌 잘못된 최소값에 빠질 가능성이 높다. 이를 전체 함수에 걸쳐 최솟값인 Global minimum이 아닌 특정 구역에서 최소값인 Local minimum에 도달했다고 한다.
- 그렇기 때문에 로지스틱 회귀에서 가중치 w를 최소로 만드는 적절한 새로운 비용 함수를 찾아야 한다.
- 앞서 확인했던 것처럼, 시그모이드 함수는 0과 1사이의 y값을 반환
  - 실제값이 0일 때, 예측값이 1에 가까워지면 오차가 커짐(반대면 오차가 작음)
  - 실제값이 1일 때, 예측값이 0에 가까워지면 오차가 커짐(반대면 오차가 작음)

$$\text{if } y = 1 \rightarrow \text{cost}\,(H(x), y) = -\log(H(x))$$

$$\text{if } y = 0 \rightarrow \text{cost}\,(H(x), y) = -\log(1 - H(x))$$

H(x) : 예측값, y : 실제값

  - y의 실제값이 1일 때 –log(H(x))을 사용하고 y=0일때, -log(1-H(x))를 사용



y=1일 때의 그래프는 파란색, y=0일 때의 그래프는 빨간색
실제값 1일 때, H(x)의 값이 1이면 오차는 0이므로 cost는 0, 반면 실제값(y)이 1인데도 H(x)가 0으로 수렴할수록 cost는 무한대가 된다. 이를 하나의 식으로 표현하면,

$$\text{cost}\,(H(x), y) = -[y log H(x) + (1 - y) log(1 - H(x))]$$ (크로스 엔트로피)

- 결과적으로, 로지스틱 회귀의 비용 함수는 다음과 같다. $J(w) = -\frac{1}{n}\sum_{i=1}^{n}[y^{(i)} log H(x^{(i)}) + (1 - y^{(i)}) log(1 - H(x^{(i)}))]$
- 이때 로지스틱 회귀에서 찾아낸 이 비용함수를 **크로스 엔트로피**(Cross Entropy)라고 한다. 가중치를 찾기 위해서 크로스 엔트로피의 평균을 취한 함수를 사용. 크로스 엔트로피 함수는 **소프트맥스** 회귀의 비용함수이기도 함

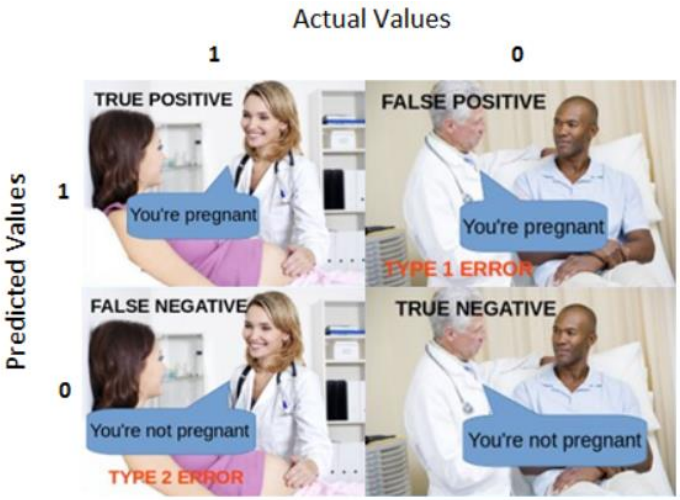
Prof. Hyerim Bae (hrbae@pusan.ac.kr)

분류 성능 평가 지표

- 오차 행렬
  - 정의 : 이진 분류의 예측 오류가 얼마인지와 더불어 어떠한 유형의 예측 오류가 발생하고 있는지를 함께 나타내는 지표

Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

- TN는 예측값을 Negative로 예측, 실제 값 역시 Negative
- FP는 예측값을 Positive로 예측, 실제 값은 Negative
- FN은 예측값을 Negative로 예측, 실제 값은 Positive
- TP는 예측값을 Positive로 예측, 실제 값 역시 Positive



- 정확도 : 예측 결과와 실제 값이 동일한 건수 / 전체 데이터 수 = (TN + TP) / (TN + FP + FN + TP)
- 정밀도 : 양성 예측도, 예측을 Positive로 한 데이터의 예측과 실제 값이 모두 Positive로 일치하는 비율

  = TP / (FP + TP) (스팸메일 여부를 판단하는 모델)
- 재현율 : 민감도, 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율

  = TP / (FN + TP) (암 판단 모델)
- F1 스코어 : 정밀도와 재현율을 결합한 지표, 정밀도와 재현율이 어느 한 쪽으로 치우치지 않을 수록 높은 값

$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \times \frac{precision \cdot recall}{precision + recall}$$

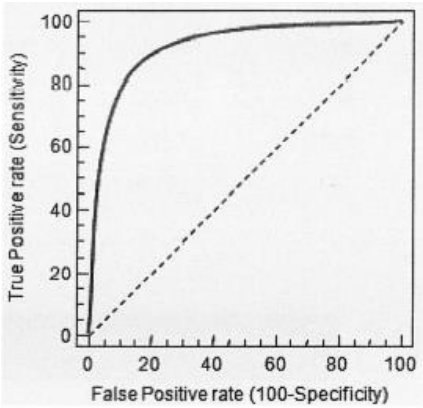
Prof. Hyerim Bae (hrbae@pusan.ac.kr)

분류 성능 평가 지표

- ROC 곡선 : 다양한 Threshold에 대한 이진분류 모델의 성능을 한 번에 표시한 것

  → FPR(False Positive Rate)이 변할 때의 TPR(True Positive Rate)이 어떻게 변하는지를 나타내는 곡선

  - 민감도(TPR) : 재현율, 실제값 Positive(양성)이 정확히 예측돼야 하는 수준(질병이 있는 사람을 질병이 있다고 판단)
  - 특이성(TNR) : 실제값 Negative(음성)이 정확히 예측돼야 하는 수준(질병이 없는 사람을 질병이 없다고 판단) = TN / (FP + TN)
  - FPR : Negative인데 Positive로 잘못 판단한 경우, = FP / (FP + TN)이므로 1 – TNR(특이성)과 같음



〈ROC 곡선 예시〉

어떻게 FPR을 0부터 1까지 변경할 수 있을까? => Threshold 변경
FPR을 0으로 만드는 방법 : Threshold를 1로 변경
-> Threshold가 높으니 모델이 Positive로 예측하지 않음
-> FP가 0이 되면서 FPR은 0
FPR을 1로 만드는 방법 : Threshold를 0으로 변경
-> Threshold가 낮으니 모델이 무조건 Positive로 예측
-> TN이 0이 되면서 FPR은 1

이렇게 임계값(Threshold)를 변경하면서 ROC 곡선을 그림

- AUC : ROC 곡선 아래 부분의 넓이, 일반적으로 1에 가까울수록 좋은 수치
  - AUC 수치가 커지려면 FPR이 작은 상태에서 얼마나 큰 TPR을 가질 수 있느냐가 관건

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

로지스틱 회귀 모델 코드 구현

https://github.com/LeeYunseol/Lab_study/blob/main/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1%ED%9A%8C%EA%B7%80%EB%AA%A8%EB%8D%B8%EA%B5%AC%ED%98%84.ipynb

참고 자료

- 파이썬 머신러닝 완벽 가이드
- Pytorch로 시작하는 딥러닝 입문
- 귀퉁이 서재 블로그(https://bkshin.tistory.com/)
- 딥러닝을 이용한 자연어 처리 입문