Basic Infromation
Of Machine Learning

정보융합학부
2018204009 김기수

Due Date : 2023.01.10

목차

- 기본적인 확률 개념
  - 베이즈 정리
  - 우도와 최대우도법

- 기본적인 성능 평가 방법
  - 분류와 회귀 문제
  - 분류 문제에서의 성능 평가
    - Confusion Matrix, Accuracy, Precision, Recall, F1-score
  - 회귀 문제에서의 성능 평가
    - MAE, MSE, RMSE, R^2

기본적인 확률의 개념

- 조건부 확률
  - 어떤 사건 A가 일어났다고 가정한 상태에서 사건 B가 일어날 확률을 의미함
    - 표본공간의 변화 : 전체(Ω) -> 사건 A
    - $\mathrm{P(B|A)} = \frac{P(A \cap B)}{P(A)}$와 같이 표현함.
  - 조건부 확률에서는 $\mathrm{P(A, B) = P(}$A)P(B|A)가 성립

- 의미
  - P(B|A)에서 A의 의미는 B의 확률을 계산하기 위해 주어진 문맥, 히스토리, 지식으로 해석이 가능
  - 주어진 지식 A가 B의 확률을 계산하는데 영향을 줄 수도 있고 그렇지 않을 수도 있음
    - 영향을 주는 경우 : A와 B는 독립 => P(B | A) = P(B), P(A, B) = P(A)P(B)를 만족
    - 영향을 주지 않는 경우 : A와 B는 종속이라고 함

기본적인 확률의 개념

- 베이즈 정리
  - 확률에 대한 관점 (주사위를 던져서 3의 배수가 나올 확률이 33.3%?)
    - 전통적인 관점 : 빈도주의 (300번 주사위를 던지면 100번은 3의 배수가 나온다.)
    - 베이지안 주의 관점 : 주사위를 던졌을 때 3의 배수가 나왔다는 주장의 신뢰도가 33%
  - 베이즈 정리의 핵심 개념 : 새로운 정보를 토대로 어떤 사건이 발생했다는 주장에 대한 신뢰도 갱신

- 베이즈 정리와 수식
  - 사전확률(Prior) : P(H), 사후확률(Posterior) : P(H | E)

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

기본적인 확률의 개념

- 베이즈 정리와 수식
  - 사전확률(Prior) : P(H), 사후확률(Posterior) : P(H | E)

$$\underbrace{\boxed{P(H|E)}}_{\text{사후확률}} = \frac{\overbrace{\boxed{P(E|H)}}^{\text{우도}}\underbrace{\boxed{P(H)}}_{\text{사전확률}}}{P(E)}$$

- P(H) : 어떤 사건이 발생했다는 주장에 관한 신뢰도(E가 관측되기 전)
- P(H | E) : 새로운 정보(E)를 알게 된 후 갱신된 신뢰도

기본적인 확률의 개념

- 우도(Likelihood)
  - 베이즈 정리에서 P(E | H)를 구할 수 있는가?에 관한 물음으로부터 출발

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

  - 확률(Probability) 과 우도(Likelihood)의 차이점
    - 확률 : 평균이 13, 표준편차가 4인 정규분포에서 5 < x < 10일 확률은?
    - 우도 : 5 < x < 10일때, 평균이 13이고 표준편차가 4인 우도는 얼마인가?
  - 즉, 쉽게 말해서 분포를 추측하는 것이다.
    - 확률 계산은 분포가 고정(파라미터가 고정), 우도 계산은 분포가 변동(파라미터가 변동)

기본적인 확률의 개념

- 최대우도법
  - 데이터를 관찰하면서 이 데이터가 추출되었을 것으로 생각이 되는 최적의 분포를 찾는 것을 의미함
    - 데이터가 1, 4, 5, 6, 9일 때, 평균이 5이고 표준편차가 1인 것과 평균이 4이고 표준편차가 0.8인 것 중 어느 것이 데이터를 더 잘 설명하는가?에 대한 답을 얻음

- 최대우도법 계산 방법
  - 각 사건은 독립이라고 가정

$$P(x|\theta) = \prod_{k=1}^{n} P(x_k|\theta)$$

$$L(\theta|x) = \log P(x|\theta) = \sum_{i=1}^{n} \log P(x_i|\theta)$$

$$\frac{\partial}{\partial\theta} L(\theta|x) = \frac{\partial}{\partial\theta} \log P(x|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log P(x_i|\theta) = 0$$

  - 만약 병뚜껑을 5번 던져서 겉면이 3번, 뒷면이 2번 나온 경우에 대하여 병뚜껑을 던질 때 겉면이 나오는 확률을 구하면, theta^3 * (1-theta)^2이 최대가 되는 theta를 찾으면 된다.(theta는 겉면이 나올 확률)

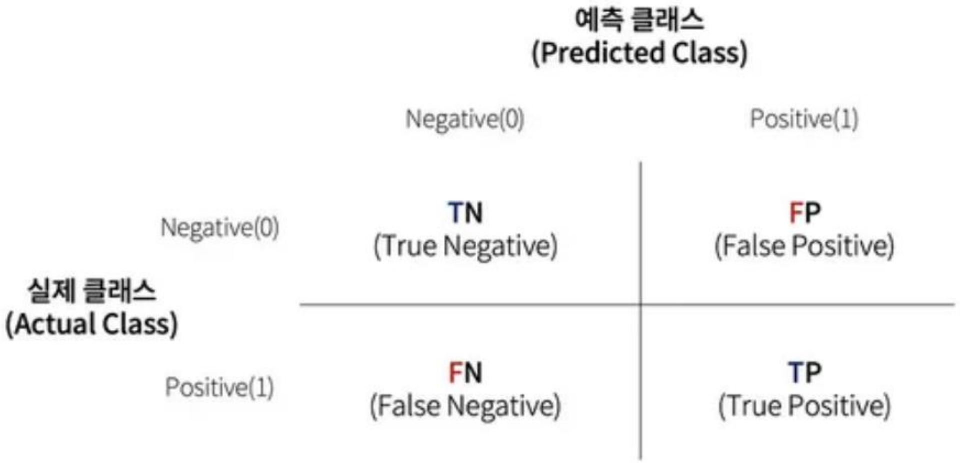
기본적인 성능 평가 방법

- 머신러닝 task
  - 분류 문제
    - 사람들의 정보를 바탕으로 심장병에 걸린 사람인 지 아닌 지를 예측
  - 회귀 문제
    - 날씨와 각종 정보를 바탕으로 특정 지역의 여행객 수 예측

- 분류 문제의 성능 평가
  - 데이터를 얼마나 잘 맞췄는 지에 따라 평가할 수 있음.
    - 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 스코어(F1-score) 등이 사용
  - 오차 행렬을 이용하여 성능 평가 지표(정확도, 정밀도, 재현율 등)를 구해볼 수 있음
- 회귀 문제의 성능 평가
  - 데이터를 얼마나 근사하게 맞췄는 지에 따라 평가할 수 있음
    - 평균 절대 오차(MAE), 평균 제곱 오차(MSE) 등이 사용 될 수 있음

분류 문제에서의 성능 평가(이진 분류 가정)

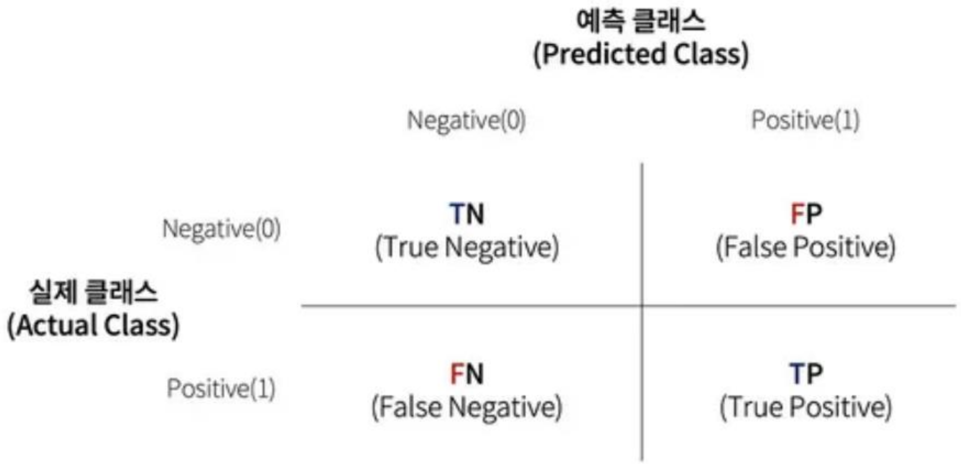
- 오차 행렬
  - 학습된 분류모델이 예측을 수행하면서 얼마나 헷갈리는 지 보여줌.
    - 단순히 정확도가 70%라는 것만 보여주면 어떤 부분을 더 잘 예측하는 지 판단할 수 없음.



- TN : 예측과 실제가 모두 Negative인 경우
- TP : 예측과 실제가 모두 Positive인 경우
- FN : 실제는 Positive인데 예측을 Negative로 잘못한 경우
- FP : 실제는 Negative인데 예측을 Positive로 잘못한 경우

분류 문제에서의 성능 평가(이진 분류 가정)

- 오차 행렬을 이용하여 성능 평가



- Accuracy(정확도) : (TN + TP) / 전체
- Precision(정밀도) : TP/(FP+TP)
- Recall(재현율) : TP/(FN+TP)
- F1-Score : 2 * Precision * Recall / (Precision + Recall)

분류 문제에서의 성능 평가(이진 분류 가정)

- 오차 행렬을 이용하여 성능 평가
  - 항상 정확도를 성능 평가의 기준으로 삼지는 않음
  - 만약, 심장병 분류를 예측한다고 할 때, 정확도를 성능 평가의 기준으로 삼으면 문제가 될 수 있음.
    - 심장병 분류에서의 목적은, 심장병에 걸린 사람을 사전에 예방
    - 만약, 모든 환자를 심장병에 걸리지 않았다고 생각을해서 정확도가 올라갔다면? 문제가 생김.
    - 이런 경우에는 재현율(Recall)을 이용함
  - 상황과 목적에 맞게 성능 평가 기준을 세우는 것이 매우 중요

실제 클래스 \ 예측 클래스	negative	positive
negative	예측 : Negative TN(405) 실제 : Negative	예측 : Positive FP(0) 실제 : Negative
positive	예측 : Negative FN(45) 실제 : Positive	예측 : Positive TP(0) 실제 : Positive

  - 이 경우에, 모든 환자를 음성(0)으로 예측해도 정확도가 90%로 높게 나옴. 반면, 양성인 사람들에 대해서는 하나도 맞추지 못함.

회귀 문제에서의 성능 평가

- 회귀 문제에서의 성능 평가
  - 예측과 실제가 얼마나 가깝냐를 기준으로 판단

- 성능 평가 지표(모든 지표에 대하여 y는 실제값, y^hat은 예측값)
  - MAE(Mean Absolute Error)

$$MAE = \frac{\sum |y - \hat{y}|}{n}$$

  - 오차의 절댓값을 모두 더한 후 평균을 낸 형태
    - 에러의 크기가 그대로 반영된다는 장점이 존재
    - 미분이 불가능하여 에러가 최소가 되는 지점을 찾기 어려움

회귀 문제에서의 성능 평가

- 성능 평가 지표(모든 지표에 대하여 y는 실제값, y^hat은 예측값)
  - MSE(Mean Squared Error)

$$MSE = \frac{\sum(y - \hat{y})^2}{n}$$

  - 오차의 제곱을 모두 더한 후 평균을 낸 형태
    - 에러에 제곱을 하기 때문에 에러가 크면 클수록 가중치가 높이 반영
    - 에러에 따른 손실이 기하 급수적으로 늘어나면 지표 자체가 커진다는 단점이 존재 => RMSE로 해결
  - RMSE(Root Mean Squared Error)
    - MSE에서 루트를 씌운 형태이며 일반적으로 많이 쓰이는 회귀모델 성능분석 지표
    - 에러에 따른 손실이 기하 급수적으로 늘어나는 상황에서 쓰기 매우 적합함

회귀 문제에서의 성능 평가

- 성능 평가 지표(모든 지표에 대하여 y는 실제값, y^hat은 예측값)
  - R2 Score(Coefficient of Determination)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

  - 앞 세 가지 지표(MAE, MSE, RMSE)는 에러에 대한 지표이기 때문에 작을수록 좋음
  - R2 Score는 1에 가까울 수록 좋음
    - SST : Total Sum of Square (편차의 총합)
    - RES : Total Sum of Residual (잔차의 총합)