필수 개념

▼ 1. 지도 학습 vs 비지도 학습

- 지도학습

  데이터와 정답(레이블)을 알고리즘에게 학습시켜 패턴을 훈련시켜 모델을 만드는 것.

- 비지도 학습

  데이터만 제공하여 알고리즘이 알아서 패턴을 찾아내는 모델을 만드는 것.

▼ 2. 분류 vs 회귀

지도학습의 결과 종류

- 분류

  데이터 입력 시, 학습된 레이블 중 가장 높은 확률의 레이블을 결과로 출력하는 것

  e.g. 이것은 스팸메일이냐 —> 네/ 아니오

- 회귀

  입력 변수와 출력 변수 간의 관계를 학습하여, 새로운 입력에 대해 연속적인 값을 예측하는 것.

  e.g. 여기 집값은 얼마쯤 되겠냐? —> 1억 3400

▼ 3. 과대적합 vs 과소적합

> 💡 **데이터에서 특징을 필요 이상으로 추출한 경우 분산이 높아지고, 반대로 필요 이하로 추출하면 편향이 높아진다**

분산 : 통계용어(평균에서 얼마나 떨어져 있는가)가 아님. 모델의 예측 변동성 의미.

편향 : 모델의 패턴을 제대로 학습하지 못 해 생기를 오류로 문제를 단순하게 접근하는 경향을 의미.

- 과소적합

  모델 학습 시 충분한 데이터의 특징을 활용하지 못 한 경우 발생

  학습 데이터, 검증 데이터에서도 낮은 정확도를 보이는 특징이 있다.

  → 해결법

  1. 보다 복잡한 모델 사용하기.
  2. 관련있는 특징을 추가하기.
  3. 많은 데이터 수집하기.

- 과대적합

  학습 데이터에서 필요 이상으로 특징을 학습해서 학습데이터에 대한 정확도는 높지만 검증,테스트 데이터에서는 정확도가 낮은 경우.

  → 해결법

  1. 더 많은 데이터 확보하여 부족한 학습데이터 보충.
  2. 데이터가 충분하지 않는 경우는 학습에 사용된 특징 줄여보기.
  3. 정규화 / 표준화 하기

▼ 4. 회귀모델 성능평가

- **SST**(total sum of squares) : 실제값과 평균값 사이 차이 , 전체 데이터가 얼마나 퍼져있는가를 보여줌
- **SSE** (sum of squared errors) : 실제값과 예측값 사이 차이 == 오차

  이 값이 작을수록 모델이 실제 데이터를 잘 맞추고 있단 의미
- **SSR** (sum of squared regression) : 예측값과 평균값 사이 차이

  이 값이 클수록 모델이 평균보다 잘 예측하고 있단 의미

- R^2 : 모델이 얼마나 데이터를 잘 설명하는가를 나타내는 지표.

1에 가까울수록 잘 설명하고 있고, 0에 가까울수록 모델 설명력이 낮다.

$$SST = SSR + SSE$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- **MSE** (mean squared error) : 예측값과 실제값의 차이를 제곱한 후 평균을 구한 값.

  큰 오류가 있을 때 값이 커짐

- **RMSE** (root mean squared error) : MSE의 제곱근 한 값.

  MSE는 제곱한 값으로 실제값과 단위가 다를 수 있음. 제곱근을 해줌으로써 실제 데이터와 같은 단위로 환산한 것.

- **MAE** (mean absolute error) : 예측값과 실제값의 차이를 절댓값 한 것들의 평균.

  MSE보다 이상치에 덜 민감하다.

- **MAPE** (mean absolute percentage error) : 예측값과 실제값 차이를 절댓값 처리 후 , 실제값으로 나눠 백분율로 변환하고, 그 값들을 평균.

  (예측값 10과 실제값 12의 차이는 2고, 2를 12로 나누면 약 0.167(16.7%))

  ▼ → 예측 모델의 정확도를 평가하는데 쓰임

  1. **직관적인 해석**: MAPE는 예측 오차를 백분율로 나타내기 때문에, 예측이 실제값에 얼마나 근접한지를 직관적으로 이해할 수 있음. 예를 들어, MAPE가 5%라면, 예측값이 실제값에 평균적으로 5% 정도 차이가 난다는 의미.
  2. **비교 가능성**: MAPE는 백분율로 표현되기 때문에, 서로 다른 데이터 세트나 모델 간의 성능을 비교할 때 유용. 예를 들어, 두 모델의 MAPE를 비교하여 어느 모델이 더 정확한지 쉽게 판단할 수 있음.
  3. **스케일 독립성**: MAPE는 데이터의 스케일에 영향을 받지 않는다. 즉, 데이터 값이 크든 작든 상관없이 일관된 평가를 할 수 있다.

▼ 5. 분류모델 성능평가

- **정확도** (accuracy) : 전체 값들 중 정확하게 예측한 값
- **정밀도** (precision) : 예측한 것 중 정확히 예측한 값
- **재현율** (recall) : 실제 값을 정확히 예측한 값 == 민감도

- **특이도** (specificity) : 실제 음성에서 TN인 것.
- **F1 점수** : 정밀도와 재현율의 조화평균 → 정밀도와 재현율이 적절히 요구될 때 사용

  1에 가까울 수록 예측을 정확하게 하고 있다 의미

> 💡 재현율을 높이려면 정밀도를 내려야됨 → 둘이 비례하게 높일 수 없음

→ 재현율을 높이려면 최대한 많은 양성사례를 잡아내고자 양성을 남발함 → 그럼 음성이지만 양성이라고 예측한 값들이 늘어남 → 정밀도가 내려감