모델 성능 평가

모형 평가

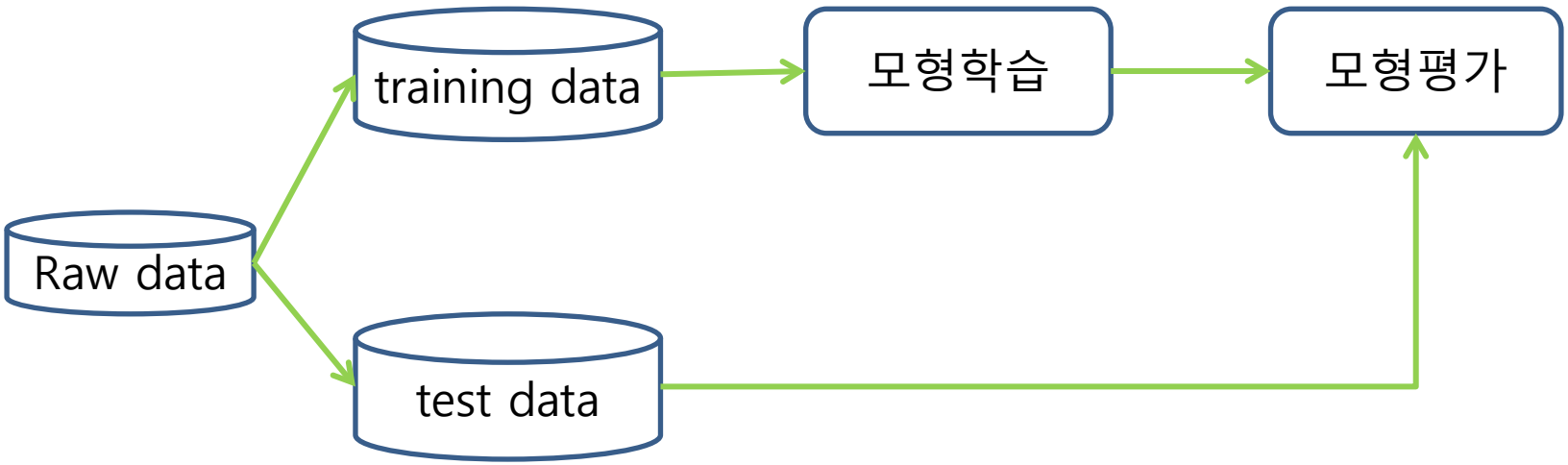
- 분류 분석 모형의 평가는 예측 및 분류를 위해 구축된 모형이 임의의 모형보다 더 우수한 분류성과를 보이는지와 고려된 서로 다른 모형들 중 어느 것이 가장 우수한 예측 및 분류 성과를 보유하고 있는지 등을 비교 분석하는 과정

- 모형 평가의 기준
  - 일반화의 가능성
    - 같은 모집단 내의 다른 데이터에 적용하는 경우에도 안정적인 결과를 제공하는 것을 의미
    - 데이터를 확장하여 적용할 수 있는지에 대한 평가 기준
  - 효율성
    - 분류 분석 모형이 얼마나 효과적으로 구축되었는지 평가
    - 적은 입력변수를 필요로 할수록 효율성이 높다
  - 예측과 분류의 정확성
    - 구축된 모형의 정확성 측면에서 평가하는 것으로 안정적이고 효율적인 모형을 구축하였다 하더라도 실제 문제에 적용했을 때 정확하지 못한 결과만을 양산한다면 그 모형은 의미를 가질 수 없다.

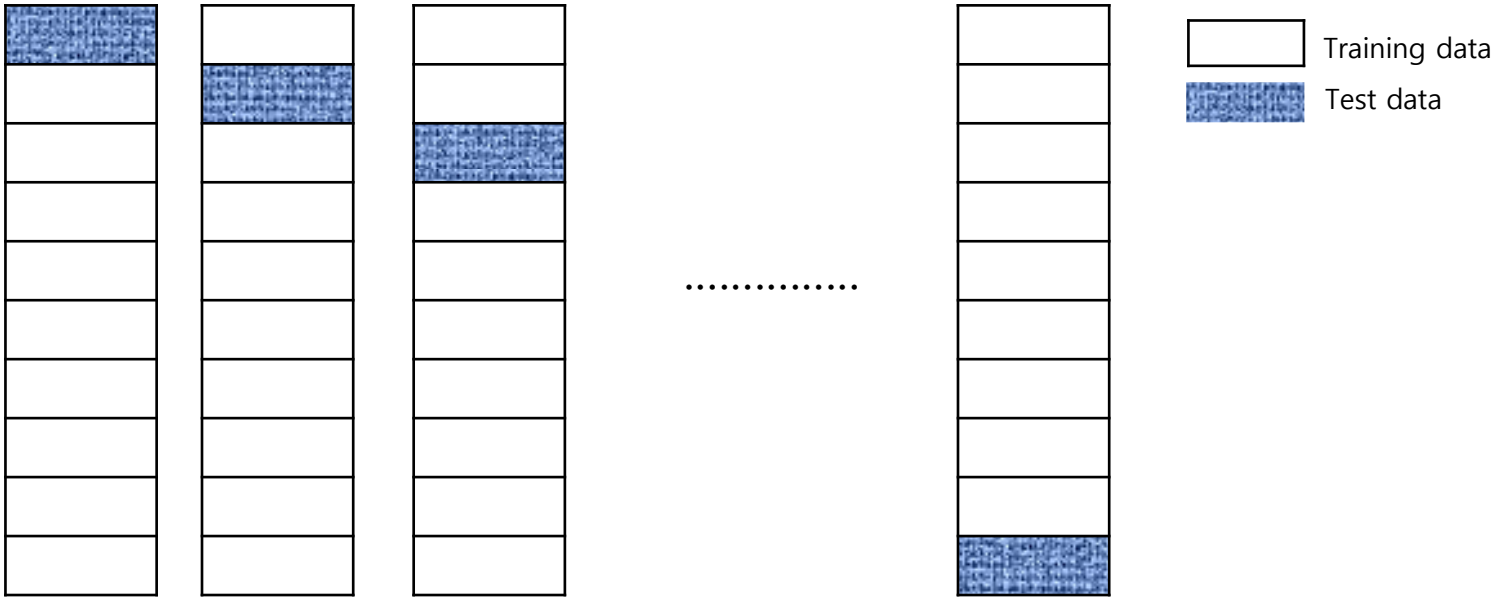
모형 평가

- 훈련용 자료(training data) : 모형 구축
- 검증용 자료(test data) : 모형의 성과 검증
- 과적합화(overfitting) 문제 해결하기 위한 방법
  - 홀드아웃
  - 교차검증
  - 붓스트랩

- 홀드아웃 방법(holdout method)
  - 훈련 데이터셋과 테스트 데이터셋으로 분할하는 절차
  - 데이터의 2/3는 훈련용, 데이터의 1/3은 테스트용



- 교차 검증(cross-validation)
  - 주어진 데이터를 가지고 반복적으로 성과를 측정하여 그 결과를 평균한 것
  - k-fold 교차검증
    - 전체 데이터를 사이즈가 동일한 k개의 하부집합(subset)으로 나누고 k번째 하부집합을 검증용 자료로, 나머지 k-1개의 하부집합을 훈련용 자료로 사용
    - 앞의 방법을 k번 반복 측정하고 각각의 반복측정 결과를 평균 낸 값을 최종 평가로 사용. 10-fold가 일반적.



- 부트스트랩 샘플링(bootstrap)
  - 평가를 반복한다는 측면에서 교차검증과 유사하나 훈련용 자료를 반복 재선정한다는 점에서 차이가 있다.
  - 관측치를 한번 이상 훈련용 자료로 사용하는 복원추출(sampling with replacement)에 기반
  - 0.632 붓스트랩
    - 63.2%의 관측치가 훈련용 데이터로 사용, 나머지 36.8%의 관측치는 검증용 데이터로 사용

- 예측모델의 성능평가
  - 평균오차(Average error), MAPE, RMSE
  - 향상차트(lift chart)

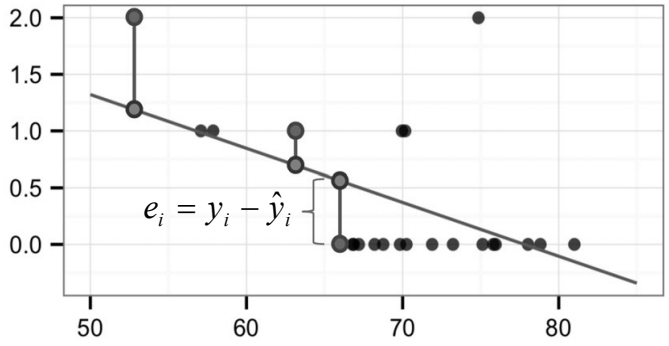
- 분류모델의 성능평가
  - 정오행렬(Confusion matrix), 분류행렬(Classification matrix)
  - 정확도(accuracy)
  - 오분류율(error rate)
  - 민감도(sensitivity)
  - 특이도(specificity)
  - ROC 도표
  - Lift Chart

- 예측모델(prediction model) : 결과변수가 수치형
- 모델 성능
  - 오차 = 실제값 – 예측값
  - 편차 = 데이터값 – 평균
  - 잔차 = 회귀모형의 적합도 = $y_i - \hat{y}_i$
- 오차 (error)
  - 데이터 마이닝 성능평가에서 사용
  - 오차 $e_i = y_i - \hat{y}_i$
- 평균오차(Mean error : ME)
  - 예측이 평균적으로 반응의 예측을 초과하는지 미달하는지 확인

$$ME = \frac{1}{n}\sum_{i=1}^{n} e_i$$ (-) : 실제값 < 예측값

- 절대평균오차(Mean absolute error : MSE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$



- 평균백분율오차(Mean percentage error : MPE)
  - 오차의 방향을 고려하여 예측이 실제값에서 얼마나 벗어나는지에 대한 백분율

$$MPE = 100 \times \frac{1}{n} \sum_{i=1}^{n} \frac{e_i}{y_i}$$

- 절대평균백분율오차(Mean absolute percentage error : MAPE)

$$MPE = 100 \times \frac{1}{n} \sum_{i=1}^{n} \left| \frac{e_i}{y_i} \right|$$

- 평균제곱오차의 제곱근(Root mean squared error : RMSE)
  - Training dataset(AE) → test dataset(RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^{\,2}}$$

분류 성능 측정

- 분류기의 예측 이해
  - 분류 모델을 평가하는 목적(지도학습의 경우)
    - 미래의 케이스에 모델의 성능이 어떻게 추론할지 더 잘 이해하기 위해서
    - 과적합(overfitting) 방지 필요

- 혼동행렬

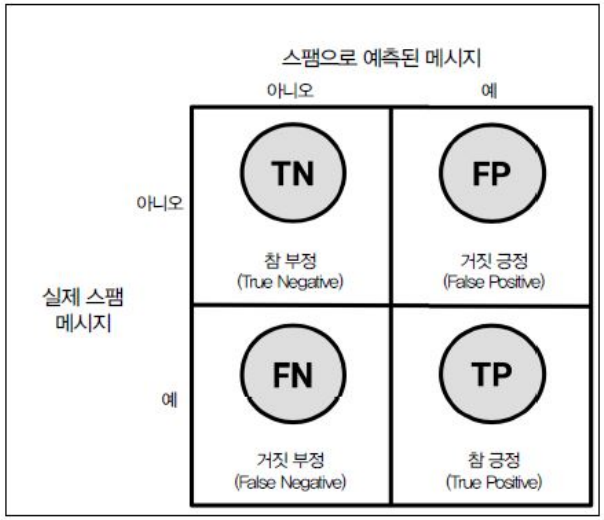


그림 10.3: 긍정과 부정 클래스 사이를 구분하면 혼동 행렬에 상세 사항이 추가된다.

분류 성능 측정

- 통계치
  - 정확도

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

  - 오류율

$$\text{error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = 1 - \text{accuracy}$$

  - 민감도(참 긍정률)

$$\text{sensitivity, recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  - 특이도(참 부정률)

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

  - F-measure(F-척도)
    - 정밀도와 재현율을 하나의 값으로 결합한 성능 척도

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

분류 성능 측정

- 카파 통계량
  - 우연히 정확한 예측을 할 가능성을 설명함으로써 정확도를 조정한다.
  - Pr(a) : 실제 일치
  - Pr(e) : 예상 일치

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

  - 0 ~ 1까지의 범위를 가진다.
    - 매우 좋은 일치 = 0.8 ~ 1(모델의 예측과 실제 값이 완벽하게 일치하는 경우)
    - 좋은 일치 = 0.6 ~ 0.8
    - 보통 일치 = 0.4 ~ 0.6
    - 어느 정도 일치 = 0.2 ~ 0.4
    - 거의 일치하지 않음 = 0.2보다 작음

- 정밀도 : 모델이 긍정 클래스를 예측할 때 예측이 얼마나 정확한지 여부

$$\text{정밀도} = \frac{TP}{TP + FP}$$

- 재현율 : 결과가 얼마나 완벽한지에 대한 척도. 전체 긍정 개수에 대해 참 긍정 개수로 정의

$$\text{재현율} = \frac{TP}{TP + FN}$$

ROC(Receiver Operating Characteristic curve, 수신자 조작 특성 곡선)

: 레이더 이미지 분석의 성과를 측정하기 위해 개발된 그래프

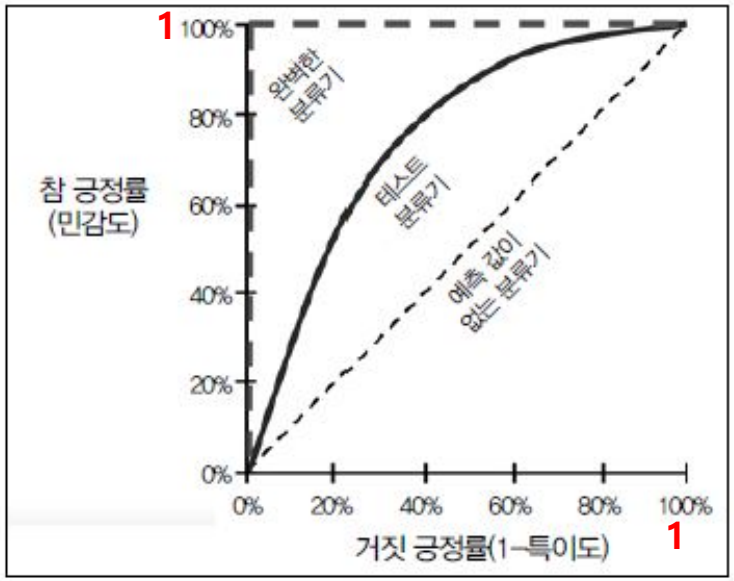


그림 10.4: ROC 곡선은 완벽한 분류기와 쓸모없는 분류기에 상대적으로 분류기 형태를 묘사한다.

- 두 분류 분석 모형을 비교하여 분석 결과를 가시화할 수 있다.
- x축, y축 모두 1인 경우 모두 True로 분류한 경우이며, x축, y축 모두 0인 경우 모두 False로 분류한 경우이다.

AUC(Area Under the Curve) : 곡선 아래 영역

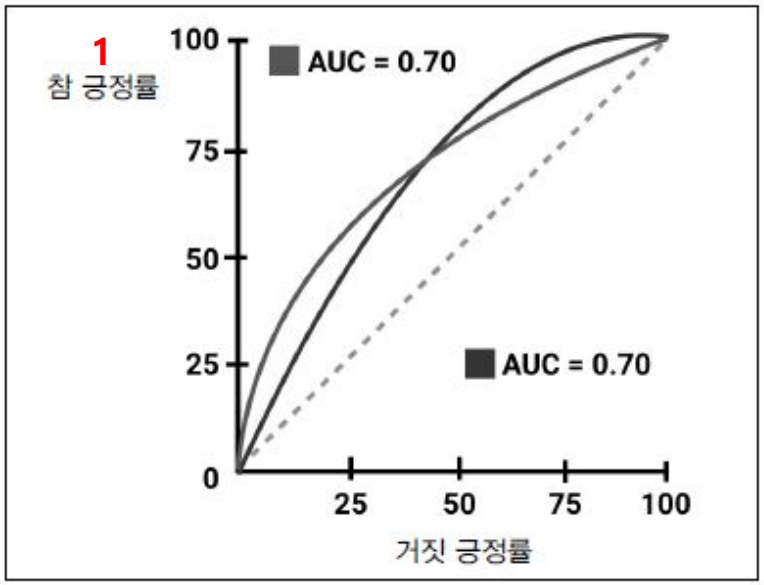


그림 10.5: ROC 곡선에서는 동일한 AUC를 갖지만 다른 성능을 가질 수 있다.

1

• 모형의 성과를 평가하는 기준은 ROC 그래프의 밑부분 면적(AUC)이 넓을수록 좋은 모형으로 평가한다.

•모형의 AUC가 1에 가까울수록 좋은 모형으로 평가된다.