통계분석B 2강

2022.02.17 (목) 14:30 ~ 16:30

chwhint@gmail.com

회귀분석

선형회귀

- x 기준으로 y가 어느정도 변하는지를 선형적으로 추정할 때
- Equation

  $y = a_o + a_1 x$

  a_0: 절편

  a_1: 기울기

  a_i : 계수

  x : 독립변수, 예측변수(input) → list일수도 있다.

  y : 응답변수, 종속변수(output) → list 일수도 있다.

- 선형회귀에 대한 예측값(prediction, outcome, output)

  $\hat{y} = \hat{a_0} + \hat{a_1} x$

  추정을 통해 얻어진 값:

  보통 (^, hat)를 붙여서 표현

- 잔차(Residual)

  예측값에 대한 실제 값과의 오차 e_i

  $y_i = a_0 + b_i x_i + e_i$

  잔차는 실제 값에서 예측값을 빼서 계산

  $\hat{e_i} = y_i - \hat{y_i}$

- Fitting 방법

  RSS를 최소화 하는 계수 a_i를 찾는다.

  잔차제곱오차(RSS, Residual Sum of Squares)

  $$RSS = \sum (y_i - \hat{y_i})^2$$

  최소제곱회귀, OLS(Ordinary Least Squres) 라는 말로도 사용

관계 확인과 예측

- 역사적으로는 선형관계로 추정되는 항목에 대해 밝히는 것이 주 용도였음
- 빅데이터의 출현 → 예측 모델로서 주로 의미를 지니게 됨
- 회귀 방정식이 인과관계를 명확히 증명하는 것은 아니다.

다중 선형 회귀

- 예측 변수가 어려개인 경우를 선형으로 조합
- 실제값

  $y_= a_0 + x_1 + a_2 x_2 + ... + a_p x_p + e_i$
- 예측값

  $\hat{y} = \hat{a_0} + \hat{a_1} x_1 + \hat{a_2} x_2 + ... + \hat{a_p} x_p$

모형 평가

- 선형회귀 뿐만 아니라, 다양한 모델에서 같은 평가방법을 사용가능
- 제곱근 평균오차(RMSE, Residual Mean Square Error)

$$RMSE = \sqrt{\frac{(\sum(y_i - \hat{y}_i)^2}{n)}}$$

작을 수록 좋은 결과

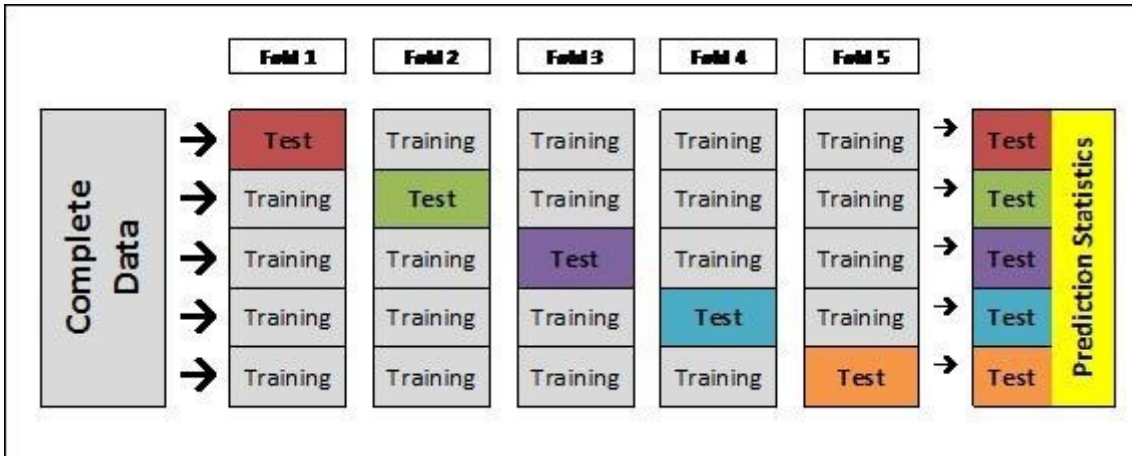
- 결정 계수

  $$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

  1에 가까울수록 좋은 결과

k-fold cross validation

Reference



https://nonmeyet.tistory.com/entry/KFold-Cross-Validation교차검증-정의-및-설명

- 샘플 데이터를 비슷한 개수의 k개 묶음으로 나눔
- 그중 첫번째 묶음을 보관해 두고, 나머지 k-1개의 묶음으로 모델을 훈련시킴
- 첫번째 묶음을 복원하고, 다음 번 묶음을 보관해두고 나머지 묶음으로 모델을 훈련시킴
- 각각의 훈련결과를 기록해두고, k 번을 반복
- 모델 평가 지표들을 대표값으로 합쳐 모델의 최종 성능을 평가한다ㅌ

모델의 선택

- 다중 선형회귀와 다른 기타 모델들은 여러가지 모델을 예측변수로 사용가능

- 사고절약의 원리(오컴의 면도날) 원리
- 모든 조건이 동일(혹은 비슷) 하다면, 복잡한 모델 보다는 단순한 모델을 선택하는 것이 좋다.