06.02 기저함수 모형과 과최적화

1) 비선형 모형

기본적인 선형회귀모형 : 입력변수의 선형조합 ($w^{T}x$)

선형회귀모형의 한계 : 비선형 데이터의 회귀모형을 만들 수 없음

대안 : 비선형 회귀모형

비선형 회귀모형 : x에 대해선 비선형, w에 대해선 선형 (2page 상단)

*비선형 모델을 구현하면서 선형모델의 방법론을 그대로 사용.
*대신, 어떤 비선형함수를 얼마나 사용할지가 중요

비선형 함수 생성 => **기저함수 활용**

2) 기저함수

기저함수 : 함수의 수열 (규칙이 정해져 있어, 규칙에 따라 여러개의 비선형함수를 만들어낼 수 있음)
ex) 다항 기저함수 (2page 하단)

*비선형모형 : 가중치(모수) 갯수는 독립변수의 갯수가 아닌, 비선형함수의 갯수에 의존
ex) 다항 기저함수 사용 시, 2차까지 하면 가중치 갯수는 3개, 10차까지 하면 가중치 갯수는 11개

기저함수 종류 : 체비셰프 다항식, 방사 기저함수, 삼각 기저함수, 시그모이드 기저함수

3) 과최적화

1. 과최적화의 이유
1) 모형의 모수(parameter)가 과도하게 많거나
2) 다중공산성

2. 과최적화가 만드는 문제
1) non-training data 입력 시, 오차가 커짐 (cross-validation 오차)
2) 샘플이 조금만 달라져도 가중치 계수의 값이 크게 달라짐 (추정의 불안정성)
*12,13page

06.03 교차검증

- in-sample testing VS outofsample testing
- 과최적화 ==>> 교차 검증 결과, 두 경우의 성능 testing 결과가 크게 다름(R^2)

1. sklearn 교차검증

1) 단순데이터 분리 " tarin_test_split( ) "

2) 교차검증

3) 교차검증 반복 " cross_val_score( ) "

    # 교안의 statsmodelsOLS 클래스 생성해, statsmodels 패키지 모형 객체 사용 가능하도록 변환

K-Fold 교차검증

- 데이터 수가 적을 때, 데이터를 나눠 여러번 testing 진행 (6page 하단)

cross_val_score()

- 11page

벤치마크 검증 데이터

- 11page

06.04 다중공선성과 변수 선택

1) overfitting 주요원인 2가지

1) 모수 갯수가 너무 많아서

2) 다중공선성 (1page 하단)
*x1과 x2가 거의 같은 데이터라면, 모형이 어떻게든 이를 구분하려 overfitting하게 됨

3) 다중공선성에 따른 overfitting 방지법 : 독립변수 제거

- VIF 활용해 의존적인 변수 삭제 (VIF, Variance Inflation Factor)
- PCA를 활용한 의존적인 변수 삭제
- 정규화(regularized) 방법 사용

2) VIF

7page, 8page, 13page