7주차

다중회귀분석

* 수치형 설명변수 X와 종속변수 Y의 관계를 오차제곱합을 최소로 하는 선형의 다차식을 구하는 것

\*종속변수 Y가 범주형 데이터 🡪 로지스틱 회귀 이용

로지스틱 회귀

* 종속변수 Y가 범주형 데이터인 경우 적용하는 모델

🡪 종속변수 Y: 범주 1이 될 확률

* 실제 사회 · 자연 현상에서는 특정 변수에 대한 확률값이 선형이 아닌 S자 형태를 따르는 경우가 많음 🡪 선형회귀에서의 종속변수에 수정을 가한 것이 로지스틱 회귀
* 승산(odds): 사건 A가 발생하지 않을 확률 대비 일어날 확률의 비율

Odds = P(A) / P(AC) = P(A) / 1 – P(A) (0 ≤ odds ≤ ∞)

* 매개변수 C값을 조정하는 규제를 통해 정확도(accuracy)를 높임
* 규제 강화(=C값을 낮춤): 복잡도 감소 🡪 데이터를 전체적으로 고려 🡪 과소적합

\*가중치(w) 🡪 0

* 규제 약화(=C값을 높임): 복잡도 증가 🡪 데이터를 개별적으로 고려 🡪 과대적합
* 94~95%의 보통의 정확도를 보이는 kNN에 비해 정확도가 96~98%로 높음
* Ex) 훈련 세트의 정확도가 검증 세트의 정확도와 비슷한 경우 과소적합 상황이라 할 수 있다. 이 경우 과대적합을 유도해야 한다. 이를 위해 C값을 증가시킴으로써 훈련 세트와 검증 세트의 정확도를 모두 높이는 적절한 모델을 설정할 수 있다.

선형 분류 모델(Linear Classifier)

* 로지스틱 회귀
* SVC(Support Vector Classifier)

로지스틱 회귀

* L1 규제 (라쏘 Lasso)
* 규제가 커질수록(=C값이 작아질수록)

1. 훈련 세트와 검증 세트의 손실이 커짐(과소적합)
2. 가중치 값이 0에 가까워짐

* 중요한 특성이 적을 때 이용
* 파라미터 세팅 필요
* L2 규제 (리지 Ridge)
* L1 규제와 비슷한 양상
* 규제가 강해져도 과소적합이 심해지지 않음
* 매개변수 C값을 조정

1. 0에 가깝게(규제 강화) 🡪 데이터 전체의 대략적인 해석(과소적합)

\*훈련 세트의 정확도는 높아도 검증 세트의 정확도는 낮음

1. 크게 🡪 데이터 하나 하나를 개별적으로 해석(과대 적합)

🡪 적합한 매개변수 C값을 조정을 해가며 훈련 세트와 검증 세트 모두 높은 정확도를 보일 수 있는 C값을 찾아야 함

선형 SVC(Linear SVC)

* C값을 낮춤
* 가중치가 0에 가까워짐
* 데이터 다수에 맞추려 함
* 수평결정계(=기울기가 수평)
* 규제 증가
* C값을 높임
* 개개의 데이터 포인트 분석
* 규제감소
* 과대적합

비선형 SVC(Non Linear SVC)

* 매개변수 추가(+kernel, gamma)
* Kernel, C, gamma 3개의 매개변수 이용
* 매개변수 튜닝
* C/gamma 값 증가 🡪 복잡한 모델
* C/gamma 값 감소 🡪 강한 규제

데이터 전처리(최댓값/최솟값 제외, 평균값을 이용 등 데이터셋을 조정하는 것)

* 정확도 높은 SVM(Support Vector Machine)을 위해 데이터 전처리가 필요