바이어스와 배리언스를 분리 할 것

와이는 에프 더하기 엡실론으로 변경 가능함.

와이 마이너스 에프 는 엡실롭의 제곱

엠에스이!를 가지고 일반화 에러를 체크 한다.

(에이 + 비 + 씨)2 =>

보라 저 수업 중간에 들어와서 집중하려고 했는데 이게 대체 무슨일이죠..

바이어스의 정의는 실제 값과 우리 예측 값의 평균의 차이 (수학적으로 정의한다면)

그래서 바이어스의 평균은 바이어스 이기 때문에 상수가 되서 ㄹ

분산 베리언스 (

선형회귀 모형 1) 엡실론은 각각의 데이터에 대해 독립적이어야 함(데이터와 엡실론은 독립적이어야 한다)

데이터가 독립일 때 평균은 E[X]E[Y] 로 나옴.

엡실론은 평균이 영임!

이의 엡실론은 평균이 영임..

마지막 남은 항

결국에 남는 것은 엠에스이를 분해를 해봤더니 바이어스와 에러의 평균 임.( 피할 수 없는 에러+분산) => 이게 엠에스 이임.

엠에스티가 정해져 있음.

따라서 바이어스와 베리언스

선형회귀에서는 정규분포를 가정함.

이건 제가

모델에 따라 엠에스이는 달라질 것..

바이어스와 배리언스가 만나는 지점을 찾으면 됨 -> 바이어스 베리언스 트레이드 오프

엠에스 이에서 각각의 바이어스와 베리언스가 차지하는 비율은 정해져있다.

머신러닝의 아트란 바이어스와 베리언스가 만나는 지점을 찾는 것

엠에스 이에는 바이어스와 베리언스로 구성되어 있고 바이어스와 베리언스는 항상 트레이드 오프하고.

좋은 바이어스와 베리언스를 찾는 게 중요함.

일반화의 오류

베리언스를 줄이는 방법은 없을까요?-> 모델을 규제하면 됨!-> 과대 적합을 갑소하기 위해 모델을 규제하면 됨.-> 24개의 피쳐를 가진 모델에서 3개의 피쳐를 가진 모델로 바꾼다(24새의 피쳐의 영향력을 줄인다)

바이어스를 줄이는 방법은 모델을 복잡하게 하는 것

선형회귀는 모델의 가중치를 제한함.

-릿지 리그레션 : 간단한 선형 회귀/목적함수에 빨간 거를 더함 -> 이빨건거의 목적은 전체를 미니마이징 하는 것.-> 그래서 세터의 값은 작어지는 것

세터 자체의 값들은

규제 해야하는 수량이 많아지면 많아 질수록 각각의 값들은 작아짐.

알파가 무한히 커지면 커질수록 모델의 차수(세터)가 0에 가까워 지는 지(수평선에 가까워지는 지)-> 그러나 알파를 아무리 키워도 세터가 0이 되지는 않음 (알파가 0이 되면 기존의 선형회귀가 될 것)

-랏쏘 리그레션

규제항이 바뀜(절대값으로-> 절대값의 특징은 어떤 특성의 가중치를 완전히 제거함)

릿지보다 우선적으로 사용하지는 않음

학습시키다보면 수학적인 특성 때문에 학습 및 업데이트 하다보면 0이 될 수도 있음-> 0이라는 것은 특성을 지워버리는 것 -> 그래서 전체 피쳐가 얼마 되지 않을 때 사용한다.

패널티 항의 목표는 세터의 값을 작게 만드는 것

그러나 랏쏘는 세터의 값을 작게 만들 뿐만 아니라 0으로 만들 수 있음. 그래서 랏쏘보다는 릿지를 더 많이 사용 함.

-엘라스틱 넷

랏쏘와 릿지의 절충점.

라이브러리에서 불러와서 쓰면 됨.

언제 어떤 것을 써야 할까요? ->

-라쏘

-엘라스틱 넷

(사용 할때는 릿지부터 쓰고 피쳐들간에 연관성이 높다고 하면 엘라스틱을 쓰면 됨)

규제 하는 이유는 오버 피팅이 높아서 쓰는 것임(모델 각각의 특성을 외워 버려서

선형회귀 모형에서 영향력은 줄인다는 것은 세타를 줄인다는 것임.

세터를 줄여주는 것 중에 랏쏘, 엣지, 엘라스틱이 있고

각각의 규제항을 둬서 세터를 작게 하는 것임.

사람들은 GML

* 머신 러닝과 통계모형의 차이

통계 모형은 최적화

머신러닝에서 다중공성선 문제 신경 쓸피요 없음

개별 변수에 대해서 통계적인 진단이 필요가 없음 (왜? 최적화로 풀었기떔ㄴ에)

오차는 모집단/ 잔차는 표본에서

모집단의 회귀선을 ㅅ

잔차는 오차의 샘플 플레이스 모델이다.

종속 변수는 독립 변수의 선형 조합 이어야 거다.

선형회귀의 가정은 1)선형 조합이기만 하면 된다.

2)오차항의 자기 상관관계가 없어야 한다.(자기상관관계)

:시계열 데이터 (오늘 내가 받은 메시지의 개수가 30개 화요일에 45개 목요일에 60개….. 일요일에 2개 이런식으로 있다고 할 때, 각각의 데이터는 시계열 데이터 이고 이 데이터를 기반으로 선형회귀를 하여 예측 값을 만든다고 할 때.)

실제 값과 예측 값의 차이인 잔차는 시계열에 영향을 받을 것.

시계열이라고 해도 시계열에 상관관계가 없을 수도 있음.

시계열 데이터는 시간에 따라 데이터가 영향을 받는 것을 자기 상관 관계임.

3) 정말 중요한 가정. 오차는 평균이 0이면서 정규분포를 따라야 한다.(굉장히 중요한 가정임)

각각 다른 분포가 같은지에 대한 가설 검정

다중공선성이 존재하지 않거나 거의 없어야 한다.

다중 공선성이란 독립 변수들이 서로 상관관계가 있는 것.

변수 1이 증가 하는데 변수 2가 유의적으로 증가하는 것 -> 다중 공선성

다중 공선성을 체크하기 위해서 베리언스 일플레이션 팩터를 사용한다.

내 모델이 데이터를 얼마나 잘 설명해주고 있나가 알 스퀘어 값임.

VIF가 크면 클수록 다중공선성이 크고 반대의 경우는 반대에 해당.

결정계수를 만들어서 VIF를 구함!

다중공선선을 확인 하는 방법은 각 피쳐 들의 선형회귀를 만들고 통제 피쳐에도 그 선형회귀가 적용되는 지 확인 하는 것.

등 분산성 때문에 로지스틱 리그레션이 생성 된 것.

등분산성의 굉장히 어긋나는 데이터가 카데고리컬 데이터(0또는 1을 가짐)임.

* 그 이유는 등분산이라고 해서 분산이 무조건 같다는 게 아니라 피
* 등분산성을 가져야만 선형회귀를 할 수 있어
* 그래서 등분산성이 없는 경우 로지스틱 리그레션을 사용함.

어느 정도 선형성을 띄어야 함!

각각 의 변수에 대해서 같은 분산을 가져야 함

* 선형 회귀의 가정 5가지은 상식으로도 알고 있어야 함!

선형 회귀 모델링 단계에서 4번은 반복적으로 실행함.

선형회귀모형을 만들었다고 해서 머신 러닝에서는 옵티마이제이션 하니까 필요없는 피쳐들은 삭제함.

그러나 선형회귀 모형은 정규분포를 적용하므로 각각의 유의성을 증명해야함.

통계학에서 사용하는 선형회귀 모형에서는 모수를 정규방정식의

엑스의 베리언스를 분모로 엑스와 와이의 베리언스 (정규방정식의 부분)으로 구함.

각각의 데이터의 유의성을 검증 하지 않고 주어진 식에 데이터를 넣어 버림.

그래서 가정이 필요함 (1. 상관관계가 없다\_기울기가 0이 아니다)

모델의 유의성을 검토성을 확인하기 위해서는 엑스와 와이의 상관관계를 검토해야함.

베터 1이 0으로부터 충분히 멀리 떨어져 있는지 가 중요함-> 파이선에서 사용하는 통계 패키지를 사용함.

가설 검정을 구할 때 중요한 것은 베터 값이 0으로부터 얼마나 떨어져 있을까가 중요함

검정은 가설을 이용해서 검정을 하는 것 뿐이지 라이브러리를 통해서 검정 여부 확인 하면 됨.

* 알 결정 계수

데이터가 가지고 있는 정보량은 데이터의 분산으로 정의되는 경우가 꽤 많음.

내 회귀모형이 데이터를 얼마나 잘 설명하니 (데이터의 분산을 얼마나 잘 정의 하니) 임.

SST

SSE

SSR( 와이의 표준 평균)-> 아무것도 모를 때, 와이의 평균값으로 정하면 됨.

\*내가 내 모델을 안만들고 예측값이 표본 평균으로 주는 거랑

내가 만든 내 모델의 차이를 보기도 함! -> 내 모델의 설명력 이고 이것을 결정 계수라고 한다.

전체 데이터의 변동에서 내 모델

표준 평균에서 떨어져 있을수록/결정계수가 크면 클수록 내 모델의 설명력이 좋은 것임.

(컨벤션은 0.7 이고 카카오는 0.9 로 잡음)

SST,SSE,SSR 의 식과 그림을 반드시 알아야 함!

알 스퀘어 값을 그대로 쓰지는 않고 수정된 알 스퀘어 값을 사용함-> 왜? 모델에 새로운 변수가 추가될 수도 있기 때문에 -> 상관관계가 높으면 높을수록 알스퀘어 값이 높아지는 것임 -> 이것이 규제와 똑같다고 볼 수 있음 -> 상관관계가 강하면 알 스퀘어 값에서 결정이 될 것 -> 수정된 알 스퀘어 결정계수는 규제화가 된 결정 계수다!

그래서 에드저스티드 알 스퀘어를 사용하자!

## -AIC (Akaike Information Criterion, 아카이케 정보 기준

케이가 들어있다는 것은 규제화(패널티\_ 피쳐가 많으면 많을수록 줄을려고 할것)가 되어 있다는 것임.

엠엘이의 라이클리 후드와 에이아이씨의 라이클리 후드는 같은 것임(에이아이씨의 라이클리 후드는 작으면 작을수록 좋음)

케이가 크면 클수록 에이아이씨의 값은 커질 것

새로운 변수를 넣어서 에이아이씨가 커지면 해당 변수는 좋지 않은 것

케이는 패널티와 같은 것임.

이디에이

# 전진제거법:내가 보는 미지의 데이터가 있을 때 데이터 마이닝 할 것이 너무 많음. 따라서 마이닐 할 때 2~3개만 먼저 하고 성능 검토 하고 성능 안좋으면 다른 거 빨리 뽑는 것

매트릭스 곱을 위해 1을 추가 함.

1. 엑스와 와이의 상관관계가 0이냐 아니냐 -> 피 밸류

* 피밸류가 크면 클수록 귀무가설을 기각 하지 않음( 엑스와 와이의 관계는 유의하지 않다)
* 피 밸류가 0.5보다 작아야 함.
* 피밸류가 크면 상관관계가 없으니까 삭제해버림.
* 그러나 삭제하기전에 유의 성이 없는 데이터 삭제
* 그리고 브이아이피를 계산할 것

선형 회귀 모형에서는 1) 유의성이 낮은 것부터(피밸류가 높은 것부터) 제거

와이를 제거한 선형회귀를 하고 서머리를 볼 것 2) 브이아이에프 높은 거 제거

순서임!

( 피 밸류 -> 브이아이 에프 -우선순위는 피 밸류, 브이 아이 에프 임)

|  |
| --- |
|  |
| **ep. Variable:** | quality | **R-squared:** | 0.299 |
| **Model:** | OLS | **Adj. R-squared:** | 0.297 |
| **Method:** | Least Squares | **F-statistic:** | 129.4 |
| **Date:** | Sun, 31 Mar 2019 | **Prob (F-statistic):** | 2.45e-203 |
| **Time:** | 01:03:34 | **Log-Likelihood:** | -3080.2 |
| **No. Observations:** | 2742 | **AIC:** | 6180. |
| **Df Residuals:** | 2732 | **BIC:** | 6240. |
| **Df Model:** | 9 |  |  |
| **Covariance Type:** | nonrobust |  |  |

* 이걸로 전체 선형 회귀를 볼 것
* 피 밸류(0.5 보다 큰것) 높은 거 다 제거 하고 나서 VIF 값(2 혹은 4)이 높은 거 제거 할 것(다중공산성)
* 마지막 단계는 테스트 셋을 넣은 것!!!!!!!!