DA Week5 preview 박보영

회귀
평균과 같은 일정한 값으로 돌아가려는 경향을 이용
선형결합에서 회귀계수와 독립변수. 최적의 회귀계수를 찾아내는 것!
선형/비선형. 단일/다중.
분류=이산값, 회귀=연속값

단순선형회귀(독립1, 종속1)
실제-회귀=오류=잔차.
mean absoulte error, rss=비용함수, 손실함수->반환값을 감소시키자!

비용최소화-경사하강법
고차원 방정식에서 사용. 점직적으로 반복적인 계산을 통해 W파라미터 값을 업데이트하면서 오류값이 최소가 되는 W구함.
미분하여, 미분값이(기울기가) 감소하는 방향으로 w를 업데이트함.
기울기=0. 극값에서 비용함수가 최소.
수행시간 오래걸림->확률적 경사하강법은 일부데이터만 이용.
피처 여러개인 경우, Xmat nxm matrix로 생각.

평가 지표	설명	수식
MAE	Mean Absolute Error(MAE)이며 실제 값과 예측값의 차이를 절댓값으로 변환해 평균한 것입니다.	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Yi - \hat{Y}i\rvert$
MSE	Mean Squared Error(MSE)이며 실제 값과 예측값의 차이를 제곱해 평균한 것입니다.	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2$
RMSE	MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 것이 RMSE(Root Mean Squared Error)입니다.	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}$
R^2	분산 기반으로 예측 성능을 평가합니다. 실제 값의 분산 대비 예측값의 분산 비율을 지표로 하며, 1에 가까울수록 예측 정확도가 높습니다.	$R^2 = \frac{\text{예측값 } Variance}{\text{실제값 } Variance}$

이 밖에 MSE나 RMSE에 로그를 적용한 MSLE(Mean Squared Log Error)와 RMSLE(Root Mean Squared Log Error)도 사용합니다.

다항회귀
2,3,차같은 다항식에서.ploynomialFeatures
차수를 높일수록 과적합의 문제.
편향-분산 트레이드오프
-낮은차수=고편향, 높은차수-고분산
편향과 분산은 반비례 관계. 편향을 낮추면 분산이 높아지고 전체 오류도 낮아짐.

규제선형모델
RSS최소화방법+과적합방지회귀계수크기제어=
비용함수목표=min(RSS)+a크기제어W
a크면 w작게하여 과적합개선, a작으면 w커도돼서 학습개선
a를 통한 규제. L1규제 릿지. L2규제 라쏘.

릿지회귀
norm1
a커질수록 회귀계수 작아짐

라쏘회귀
norm2
불필요한 회귀계수를 급격하게 감소시켜 0으로 만들고 제거함.
적절한 피처만 회귀에 포함시키는 피처 선택의 특성을 가짐.

엘라스틱넷회귀
L1,2 결합. rss(w)+a2*w2+a1*w1
라쏘 회귀의 회귀계수 급변 완화하기 위해 등장
시간 오래걸림.

선형회귀모델은 정규분포 선호. 왜곡된 분포도는 부정적인 예측성능->standardscler, log(!)

로지스틱 회귀, 분류
시그모이드 함수. 반환값으로 분류
가볍고 빠르고 이진분류예측성능 뛰어남.

회귀트리
리프노드에 속한 데이터 값의 평균값을 구해 회귀 예측값을 계산
트리 모델은 회귀에도 가능. CART 알고리즘을 기반으로 하고 있기 때문.
분할되느 데이터 지점에 따라 브랜치를 만들면서 계단 형태의 회귀선. max_depth크면과적합