1 부스팅

부스팅은 앙상블의 형태로 만드는 일반적인 기법.
연속된 라운드마다 잔차가 큰 레코드들에 가중치를 높여 이련의 모델들을 생성하는 일반 기법.

1. 배깅과 같이 의사결정 트리에 사용됨.
2. 이전 모델의 오차를 줄이는 방향으로 다음 모델을 연속적으로 생성
3. 에이다부스트, 그레디언트 부스팅, 그레디언트 부스팅

1.1 부스팅 알고리즘

그레디언트 부스팅은 에이다 부스팅과 비슷하지만 비용함수를 최적화하는 접근법을 사용했다는 점에서 차이가 있다. 그레디언트 부스팅에서는 가중치를 졸절하는 대신에 모델에 유사잔차를 학습하도록 한다. 확률적 그레이디언트 부스팅에서는 매 단게마다 데이터와 예측변수를 샘플링하는 식으로 그레디언트 부스팅에 랜덤한 요소를 추가한다.

1.2 XG부스트

가장 많이 사용되는 오픈소스 소프트웨어.
하이파라미터중 가장 중요한 파라미터는 subsample과 eta(learning rate)이다.
subsample은 각 반복 구간마다 샘플링할 입력데이터의 비율을 조정
eta는 축소 비율을 결정

1.3 정규화:오버피팅 피하기

오버피팅

1. 학습 데이터에 없는 새로운 데이터에 대한 모델 정확도가 떨어짐
2. 모델의 예측 결과에 변동이 매우 심하고 불안정한 결과를 보임

정규화(regulartizaion) 모델 복잡도에 따라 벌점을 추가하는 형태로 비용함수를 변경하는 방법. 모델을 정규화하기 위한 두 파라미터 α, λ가 있다. 각각 맨해튼 거리와 유클리드 거리를 의미. 리지 회귀와 라소 회귀는 선형회귀에 위에 개념을 도입해서 과적합을 방지하도록 한 모델들이다.

1.4 하이파라미터와 교차타당성검사

교차타당성검사

1. 데이터 k개의 서로 다른 그룹(폴드)으로 랜덤하게 나눔

2. 각 폴드마다 해당폴드에 속한 데이터를 제외한 나머지 데이터를 가지고 모델 학습

3. 폴드에 속한 데이터를 이용해 모델 평가

위 방법으로 최적의 하이퍼파라미터 조합을 찾는다. XG 부스트의 하이퍼파라미터

1. eta/learning rate : [0,1] default = 0.3 python = 0.1

2. n estimators: 부스팅 라운드 횟수. eta 값이 작으면 라운드 횟수 늘려야됨. 오버피팅 방지하는 파라미터 설정이 포함된 경우, 라운드 횟를 늘려도 됨.

3. max depth: 트리의 최대 깊이 기본값 6. 트리가 얕아 노이즈가 많은 데이터에 대해 모델이 복잡한 거짓 상호작용을 회피하는데 도움이 됨. 파이썬에서는 기본값 3

4. subsample 및 colsample bytree : 일부데이터 비복원 샘플링하는 비율 및 예측변수 중 일부변수를 샘플링하는 비율. 기본값 1.0

5. lambda/reg lambda 및 alpha/reg alpha : 오버피팅을 조절하기 위해 사용되는 정규화 파라미터들 파이썬에서는 reg lambda = 1, reg alpha = 0 이 기본값