4주차 예습과제

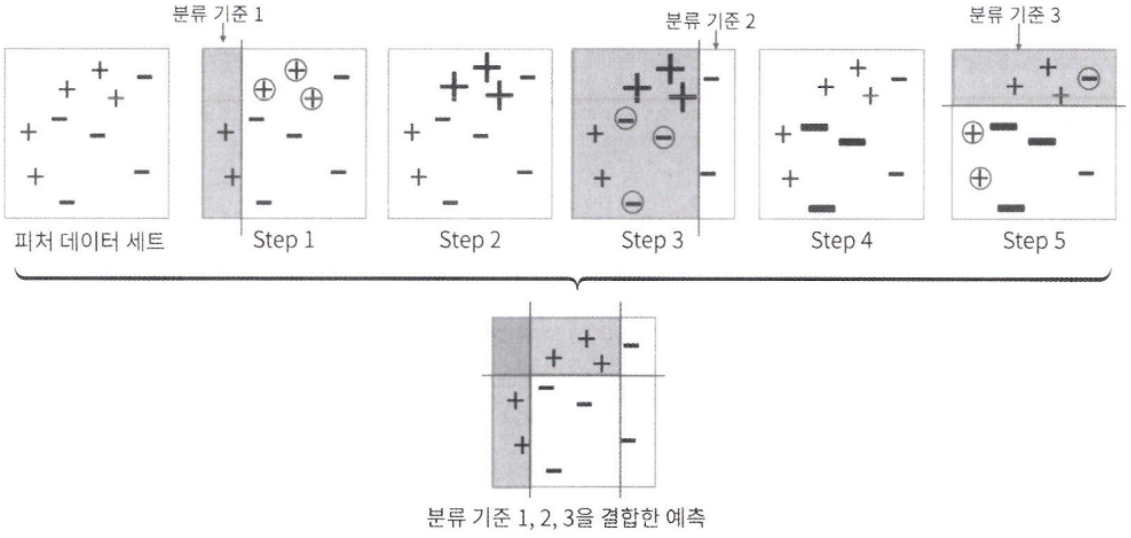
4.5 GBM(Gradient Boosting Machine)

부스팅 알고리즘이란 ?

여러 개의 약한 학습기(weak learner)를 순차적으로 학습-예측하면서 잘못 예측한 데이터에 가중치 부여를 통해 오류를 개선해 나가며 학습하는 방식. 대표적 구현으로 AdaBoost(Adaptive boosting)와 **GBM이 있음.**

AdaBoost란 ?

오류데이터에 가중치를 부여하며 부스팅을 수행하는 대표적인 알고리즘



s1. 첫번째 약한 학습기가 분류 기준1로 + / - 를 분류함 여기서 동그라미 쳐진 +는 잘못 분류된 오류데이터임

s2. 오류데이터에서 가중치 값을 부여함, 가중치가 부여된 오류데이터(+)는.다음 약한 학습기가 더 잘 분류할 수 있도록 크기가 더 커짐

s3. 두번째 약한 학습기가 분류 기준2로 +/- 를 분류함. 동그라미로 표시된(-)은 오류데이터임

s4. s2와 마찬가지로 오류데이터에 더 큰 가중치를 부여하여 크기가 커짐.

s5. 세번째 약한 학습기가 분류기준3으로 +/-를 분류하고 오류데이터를 찾음. 에이다부스트는 이렇게 약한학습기가 순차적으로 오류 값에 대해 가중치를 부여한 예측 결정 기준을 모두 결합해 예측을 수행함.

마지막 아래 그림은 약한 학습기로 세차례 분류한 결과를 모두 결합한 것이다. 개별 약한 학습기보다 훨씬 정확도가 높아졌음을 볼 수 있다.

GBM

에이다부스트와 유사하나, 가중치 부여를 **경사하강법**을 이용하는 것이 큰 차이임.

오류값= 실제 값 - 예측값. 분류의 실제 결과 값을 y, 피처를 $x_1, x_2, \ldots x_n$ 그리고 이 피처에 기반한 예측 함수를 $F(X)$ 함수라고 하면 오류식 $h(x) = y - F(X)$이 됨. 이 오류식 h를 최소하 하는 방향성을 가지고 반복적으로 가중치를 부여하는 것이 **경사 하강법 (Gradient Decent)**임. 이것은 머신러닝의 중요한 기법 중 하나임

GBM 하이퍼파라미터및튜닝

• loss: 경사하강법에서 사용할 비용함수를 지정합니다.
• learning_rate: GBM이학습을진행할때마다적용하는학습률입니다.

•Weaklearner가순차적으로오류값을보정해나 가는데적용하는계수입니다. 0~1 사이의값을지정할수있으며기본값은0.1입니다

•n_estimators: weaklearner의 개수입니다. weaklearner가순차적으로오류를 보정하므로 개수가 많을수록 예측성능이 일정수준까지는 좋아질 수있습니다.
• subsample:weaklearner가학습에사용하는데이터의샘플링비율입니다. 기본값은1이며, 이는전체학습데이터를 기반으로학습한다는의미입니다(0.5이면학습데이터의50 %), 과적합이염려되는경우s ubsample을1보다작은값으 로설정합니다.

4.6 XGBoost(eXtra Gradient Boost)

XGBoost는트리기반의앙상블학습에서가장각광받고있는알고리즘중하나입니다. 유명한캐글 경연대회(Kaggle Contest)에서상위를차지한많은데이터과학자가XGBoost를이용하면서널리 알려졌습니다. 압도적인수치의차이는아니지만. 분류에있어서일반적으로다른머신러닝보다뛰어 난예측성능을나타냅니다. XGBoost는GBM에기반하고있지만, GBM의단점인 느린수행시간및
과적합규제(Regularization) 부재등의문제를해결해서매우각광을받고있습니다. 특히 XGBoost 는병렬CPU환경에서병렬학습이가능해기존GBM보다빠르게학습을완료할수있습니다. 다음은
X GBoost의주요장점입니다.