5 주차 과제

분류

주하연

1. 앙상블 기법에 대해 조사하시오 (배깅, 부스팅, 보팅, 스태킹)

앙상블 기법 : 여러 개의 분류기를 생성하고 그 예측을 결합함으로써 보다 정확한 예측을 도출하는 기법이다. 이는 강력한 하나의 모델을 사용하는 대신, 보다 약한 모델 여러 개를 조합하여 더 정확한 예측에 도움을 주는 방식이다. 앙상블 기법은 머신 러닝과 통계학에서 주로 사용되고, 앙상블은 “조화”, “집합”을 의미한다.

여러 개의 모델을 생성하여 예측을 결합하여 성능을 향상시켜 예측

➔ 여러 결과들을 취합해서 가장 좋은 결과를 낼 수 있도록 하는 것이 목적

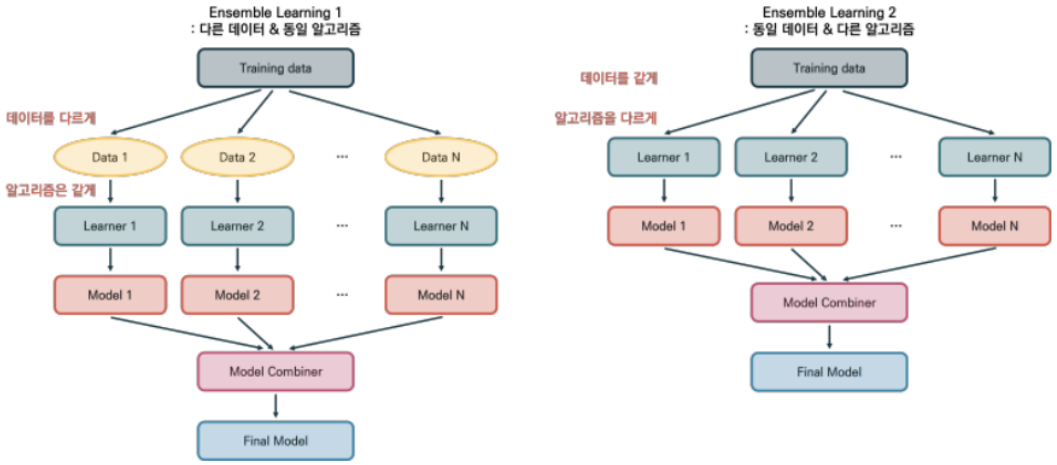
➔ 각각 틀릴 수는 있지만, 최종적으로 종합한 결과는 좋은 결과가 나올 수 있다.

➔ Base-learner : 기본모델

➔ Meta-learner : 기본 모델을 많이 만들고 결합해서 새로 만든 learner

앙상블은 단일 모델보다 더 나은 예측 성능을 제공하는 경우가 많다. 이는 다양한 개별 모델의 예측을 평균화하거나 투표를 통해 결합함으로써 가능하다. 앙상블은 예측 모델 간의 상호 보완성과 다양성을 활용하여 편향을 줄이고 분산을 줄이는 효과를 가져온다. 이로 인해 예측의 안정성과 일반화 능력이 향상될 수 있다.

즉, 더 많은 연산능력을 활용하여 더 좋은 예측력을 가진 것이 앙상블기법이다.



A. 배깅(Bagging)

배깅은 Bootstrap Aggregating의 줄임말로, 원래 데이터셋에서 부트스트래핑(복원

추출)방법을 사용하여 여러 개의 독립적인 예측 모델을 만들고, 각 모델의 예측 결과를 평균화하여 최종 예측을 수행하는 앙상블 기법이다. 각 모델은 독립적으로 학습되기 때문에 병렬처리가 가능하며, 예측 결과의 분산을 줄이는 효과가 있다.

➔ 복원 추출로 다수의 데이터셋을 생성 후, 모델을 각각 훈련시켜 결과들의 평균치를 사용

- 장점

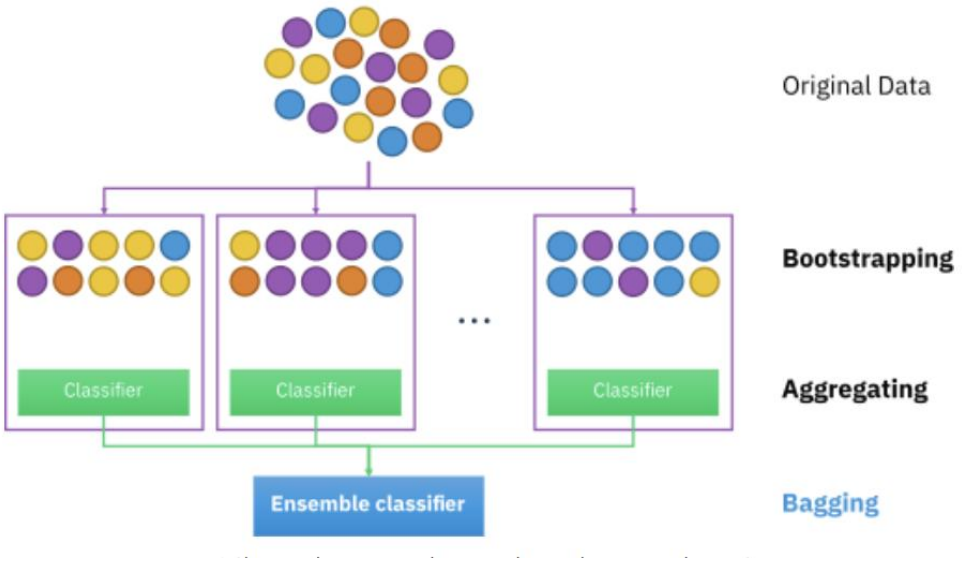
  Stumps(=split)를 딱 한 번만 하기 때문에 매우 안정적(stable)임

  병렬 처리로 동시 학습하여 속도가 빠름

  결과가 잘 나오는 편이기 때문에 tree를 간단히 해도 무방함

- 단점

  안정성에 중점을 두기 때문에, Accuracy가 다소 떨어질 수 있음

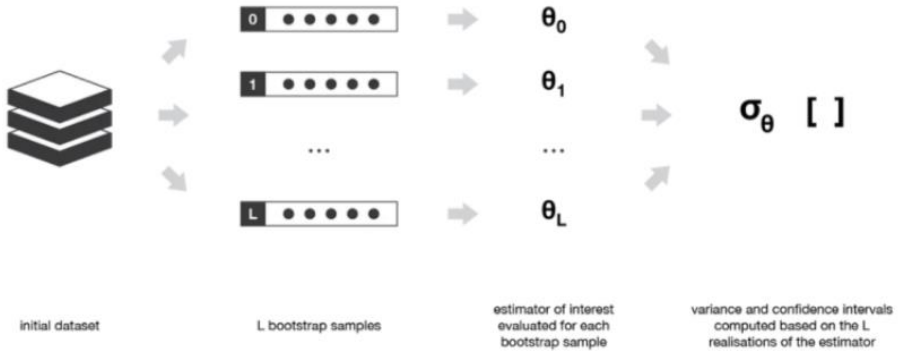


ii. 부트스트랩(bootstrap)

부트스트랩(bootstrap)은 random sampling을 적용하는 방법을 일컫는 말이다. 주어진 데이터셋으로부터 복원 추출을 통해 샘플을 반복적으로 추출하는 과정이다. 이 방법은 통계적 추론이나 예측 모델의 불확실성을 추정하는 데에 활용된다.

부트스트랩은 데이터의 편향을 줄이고, 모델의 예측 성능을 향상시키는 데 효과적이지만, 모델의 복잡도가 증가하고 학습시간이 오래 걸린다는 단점이 있다.

머신러닝에서 random sampling을 통해 training data를 늘릴 수 있다.



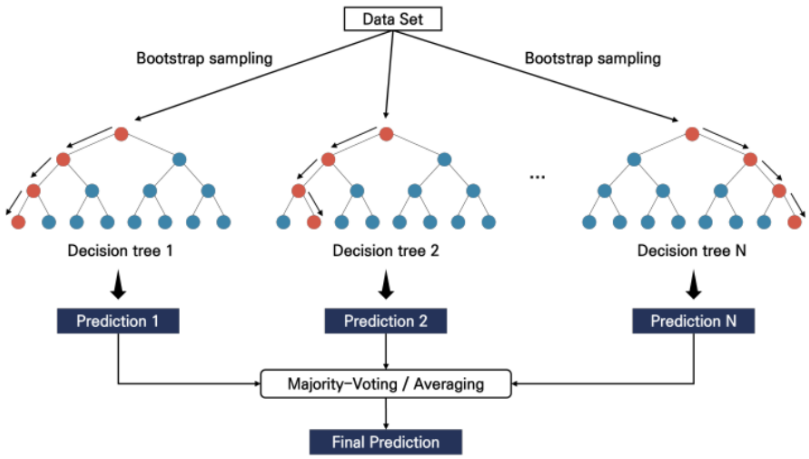
- 부트스트랩 과정
  - ■ 원래 데이터 셋으로부터 크기가 원본 데이터셋과 동일한 샘플을 복원 추출한다. 이렇게 추출한 샘플을 부트스트랩 샘플이라고 한다.
  - ■ 부트스트랩 샘플로부터 통계량(평균, 분산 등)을 계산한다. 이러한 통계량은 원래 데이터셋에서의 특성을 대표하는 추정값이 된다.
  - ■ 위의 과정을 여러 번 반복하여 다수의 부트스트랩 샘플과 그에 해당하는 통계량을 얻는다.

배깅의 대표적인 예시로는 랜덤 포레스트가 있다.

iii. 랜덤포레스트

랜덤 포레스트는 앙상블 기법의 한 종류로, 부트스트랩을 사용하여 여러 개의 결정 트리를 생성하고, 그 예측을 결합하여 최종 예측을 도출하는 방식이다. 이는 데이터의 편향을 줄이고, 모델의 예측 성능을 향상시키는 데 효과적이지만, 모델의 복잡도가 증가하고, 학습 시간이 오래 걸리는 단점이 있다.

즉, 설명변수를 무작위로 선택함으로써, 트리의 다양성을 확보하여 모형간의 상관관계를 줄이고자 하는 것이다.



B. 부스팅

부스팅은 약한 예측 모델들을 순차적으로 학습시켜 강한 예측 모델을 만들어내는 앙상블 기법이다. 각 모델은 이전 모델의 오차에 집중하여 다음 모델을 학습시키는 방식으로 예측 성능을 향상시킨다. 부스팅은 가중치를 사용하여 각 모델의 예측 결과에 가중치를 부여하고, 이를 결합하여 최종 예측을 하게 만든다.

➔ 순차적으로, 복원추출로 가중치를 준다.

➔ 즉, 먼저 생성된 모델을 꾸준히 개선해 나가는 방향으로 학습이 진행

- 장점

  오답에 대해 높은 가중치를 부여하므로 Accuracy가 높게 나타남

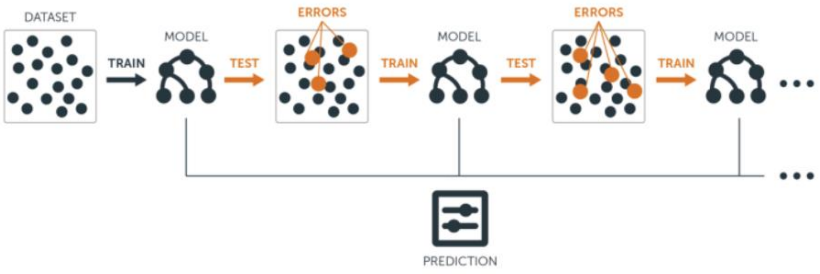
- 단점

  Outlier에 민감

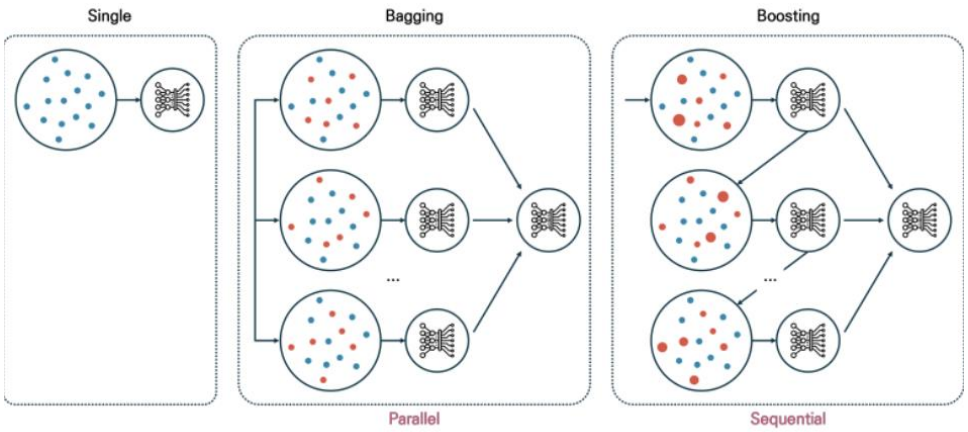
  가중치를 업데이트하며 순차적으로 학습하기 때문에 속도가 느림

  Accuracy 성능이 좋아지는 쪽으로 설계되어 있으므로 과적합 주의

  Accuracy가 낮아질 때도 있기 때문에 Unstable함



ii. 배깅 vs 부스팅

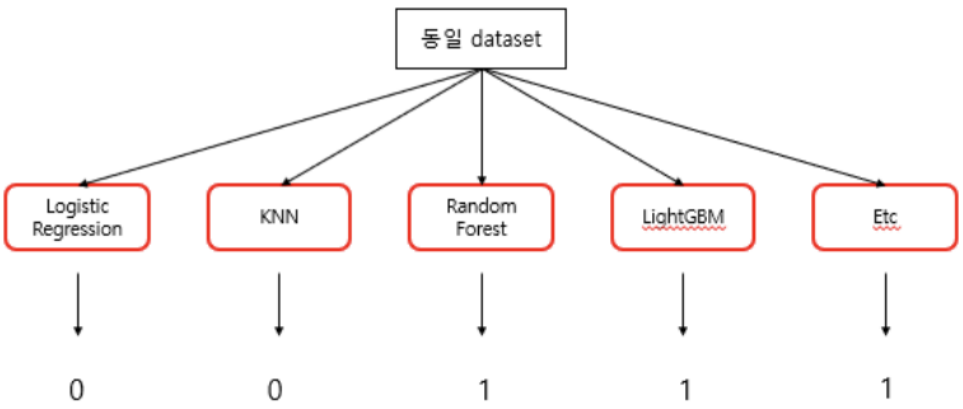


부스팅의 대표적인 예시로는 그래디언트 부스팅과 에이다부스트가 있다.

C. 보팅

보팅은 서로 다른 예측 모델들이 투표를 통해 예측 결과를 결합하는 앙상블 기법이다. 분류 문제에서는 다수결 투표를 사용하여 클래스 레이블을 선택하고, 회귀 문제에서는 예측값들을 평균화하여 최종 예측을 만든다. 보팅은 다양한 예측 모델들의 다른 관점과 특성을 활용하여 더 강력한 예측을 할 수 있게 한다.

➔ 말 그대로 투표를 하는 것을 의미한다. 여러 개의 model(분류기)를 사용하게 된다. 동일 데이터셋에서 여러 개의 Model로 학습을 진행한다.



위 그림처럼 동일 dataset에 대해서 모델별로 다른 결과를 뽑아낼 것이다.

여기서 Voting방식이 Hard Voting과 Soft Voting 으로 나뉘게 된다.

i. Hard Voting

굉장히 단순하게 '다수결 투표' 를 따라간다.

ii. Soft Voting

Hard Voting과는 다르게 각 레이블의 예측 확률의 평균으로 최종 분류를 진행한다. 예측 확률의 평균으로 분류하기 때문에 Hard Voting과 결과가 다르게 나올 수도 있다.

하드 보팅(Hard Voting)	다수결 원칙과 유사 예측한 결과값들 중 다수의 분류기가 결정한 예측값을 최종 보팅 결과값으로 선정	
소프트 보팅(Soft Voting)	분류기들의 레이블 값 결정 확률을 모두 더해 이를 평균내서 확률이 가장 높은 레이블 값을 최종 보팅 결과값을 선정	

D. 스태킹

스태킹은 다양한 예측 모델들을 조합하기 위해 여러 계층의 모델을 구성하는 앙상블 기법이다. 먼저 초기 예측 모델들을 통해 예측을 수행하고, 이 예측 결과들을 새로운 특성으로 사용하여 최종 예측 모델을 학습시킨다.

➔ 다양한 알고리즘으로 학습 및 예측, 예측값들로 다시 학습해서 최종 예측

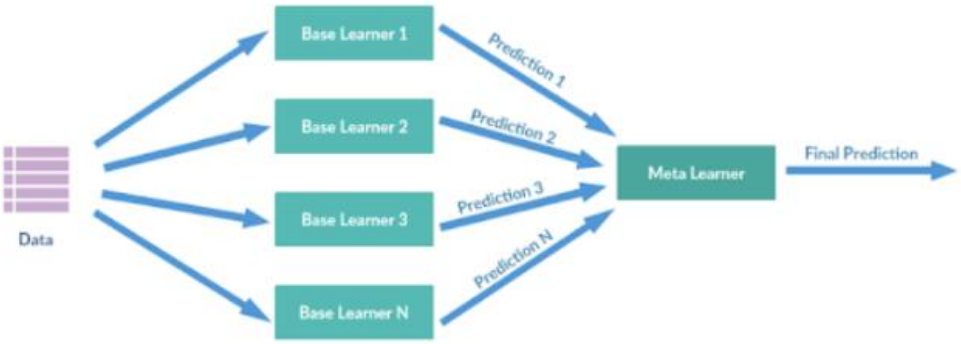
➔ Base model들을 만들어서 그 결과를 다시 input으로 하는 meta learner로 사용

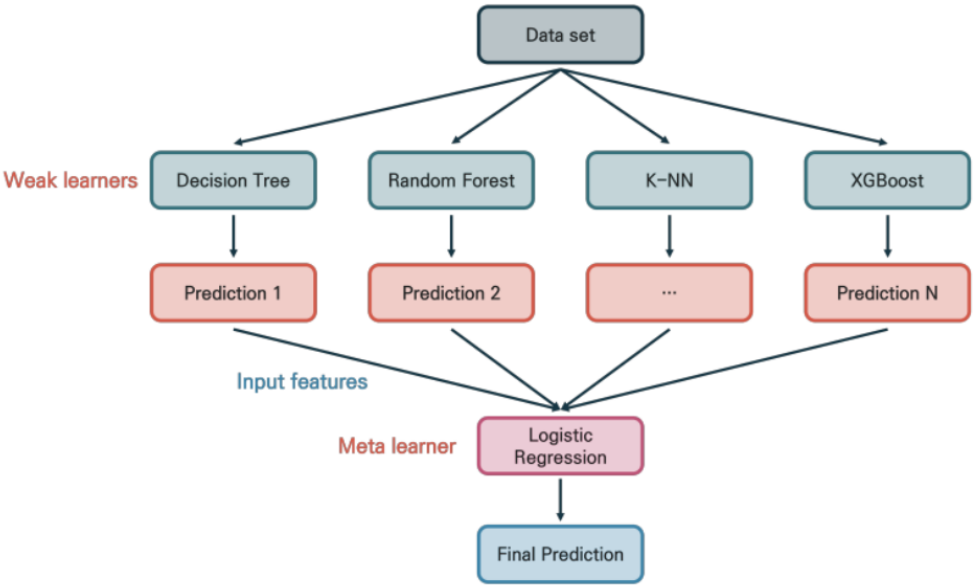
- 장점

  단일 모델에 비해 예측 성능 향상

- 단점

  과적합, 해석문제, 학습시간 오래걸림, 비용 증가, 유지보수 어려움





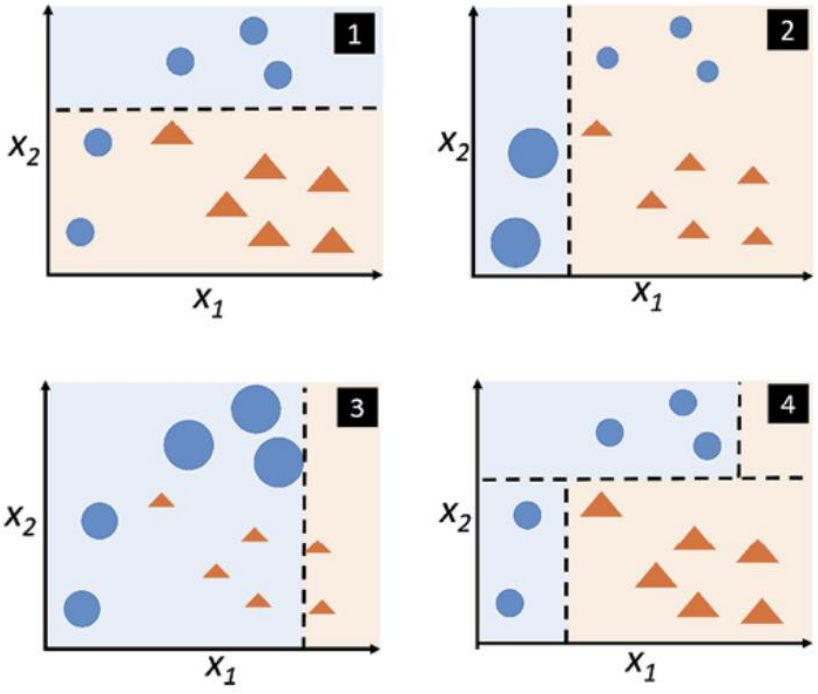
2. 앙상블 기법을 적용한 모델의 종류를 조사하시오.

   A. 배깅 – 랜덤포레스트(1번에서 설명)

   B. 부스팅 - 아다부스트=에이다부스트(AdaBoost)

   오류 데이터에 가중치 부여를 통해 오류를 개선해 나가면서 학습하는 방식

   약한 학습기가 순차적으로 오류 값에 대해 가중치를 부여한 예측 결정 기준을 모두 결합해 예측을 수행한다.



i. 첫 번째 약한 학습기(weak learner)가 분류 기준 1로 분류를 실행하며 오류 데이터가 발생한다

ii. 오류 데이터에 대해서 다음 약한 학습기가 더 잘 분류할 수 있도록 가중치 값을 부여한다.

iii. 두 번째 약한 학습기가 분류 기준 2로 분류를 실행하며, 오류 데이터가 발생한다

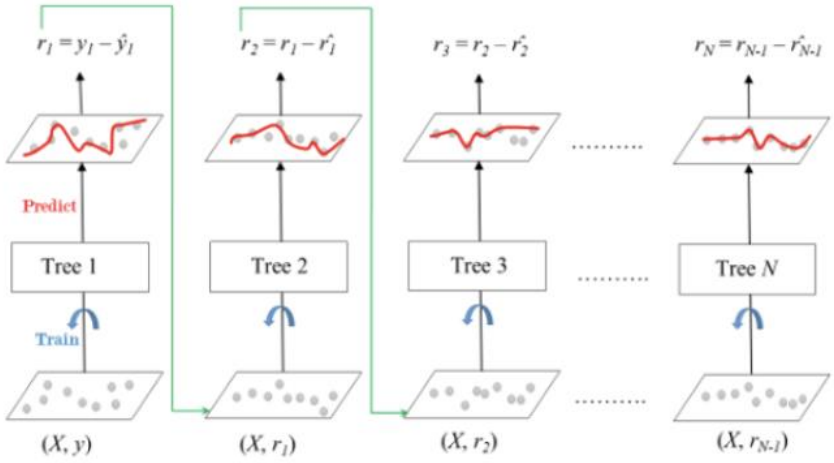
iv. 똑같이 오류 데이터에 대해서 가중치 값을 부여한다

v. 첫 번째, 두 번째 약한 학습기를 모두 결합하여 예측한다. 개별 약한 학습기보다 정확도가 향상한다

C. 부스팅 – 그래디언트 부스팅(Gradient Boosting)

새로운 모델을 만들 때 잔여오차로 학습 : 실제값 – 예측값을 최소화하는 방향성을 가지고 업데이트(경사하강법)

순차적 예측 오류 보정으로 학습을 수행 -> 대용량 데이터의 경우 학습에 많은 시간 필요



D. 보팅 – 하드보팅, 소프트보팅(1번에서 설명)

3. 하이퍼 파라미터의 개념과 하이퍼 파라미터 최적화 (HyperParameter Tuning) 기법에 대해 조사하시오.

A. 하이퍼 파라미터

하이퍼 파라미터는 머신 러닝 모델의 구성을 결정하는 매개변수이다. 이는 모델 학습 과정에서 사용자가 직접 설정해야 하는 값으로, 모델의 성능과 일반화 능력에 영향을 미친다.

예로는 학습률, 에포크 수, 노드 수 등이 있다.

B. 하이퍼 파라미터 최적화

하이퍼 파라미터 최적화는 하이퍼 파라미터의 값을 조정하여 머신 러닝 모델의 성능을 향상시키는 과정이다. 이는 모델의 일반화 성능을 향상시키고, 과적합을 방지하기 위해 중요하다.

i. 그리드 탐색(Grid Search)

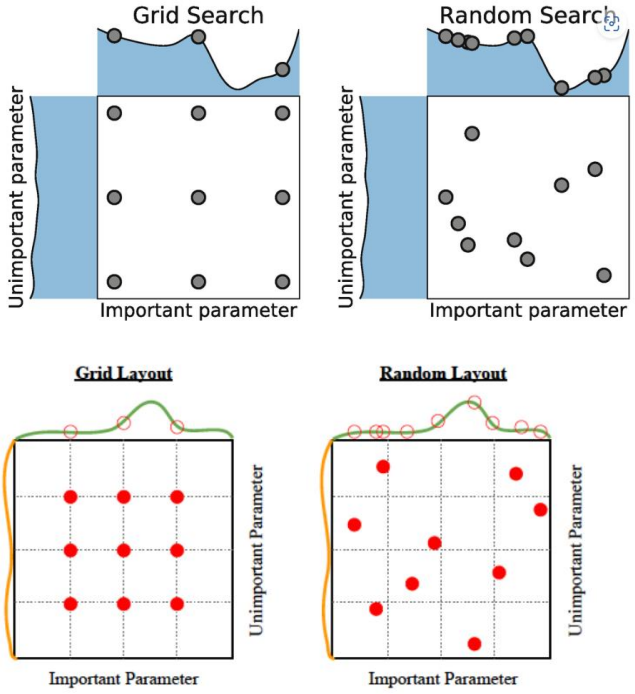
하이퍼 파라미터의 모든 가능한 값을 시도하는 기법이다.

주어진 하이퍼파라미터 공간을 그리드로 구성하고, 각 조합에 대해 교차 검증 등의 방법을 사용하여 성능을 평가하고 최적의 조합을 선택한다. 이는 모든 가능한 조합을 시도하기 때문에 계산 비용이 크지만, 상대적으로 간단하게 구현할 수 있다.

ii. 랜덤 탐색(Random Search)

하이퍼 파라미터의 가능한 값 중에서 무작위로 선택하는 기법이다.

그리드 탐색에 비해 모든 조합을 시도하지 않기 때문에 계산 비용이 낮아지지만, 최적의 조합을 찾는 데에는 더 많은 시도가 필요할 수 있다. 이는 조합 탐색에 효율적이다.



Grid Search vs Random Search (datarian.io)

iii. 베이지안 최적화(Bayesian Optimization)

하이퍼 파라미터의 값을 선택할 때 이전의 실험 결과를 고려하는 기법이다.

이전 시도에서 얻은 정보를 바탕으로 후보 하이퍼파라미터 값들의 가능성을 모델링하고, 이를 활용하여 더 좋은 후보를 선택하는 방식으로 탐색 공간을 효율적으로 탐색한다.

iv. 자동화된 하이퍼 파라미터 최적화 도구

Scikit-learn의 BridSearchCV, RandomizedSearchCV, Optuna, Hyperopt 등과 같은 자동화된 하이퍼 파라미터 최적화 도구를 사용할 수도 있다. 이러한 도구는 하이퍼 파라미터 탐색 과정을 자동화하고, 최적의 조합을 찾는 데 도움을 준다.

4. bostonhousing 데이터를 분할하여 주택 가격을 예측하는 회귀 앙상블 모델을 구축하고, 회귀성능평가지표를 통해 예측성능을 평가하시오.

https://colab.research.google.com/drive/19s4IK7gUsdG6FW7J9yX9Wzmygqgkbo5F?usp=sharing

5. Framingham 데이터를 분할하여 10년 후 관상동맥 질환 발병 여부를 분류하는 분류 앙상블 모델들을 구축하고, 분류성능평가지표를 통해 예측성능을 평가하시오.

https://colab.research.google.com/drive/1e3kA7-u8cpNT6eGahtEz0mEnzbZL-ZjO?usp=sharing