Machine Learning

Chapter 6 지도 학습(Supervised Learning)

START

스마트인재개발원
Smart Human Resources Development

학습목표

- 앙상블 개념을 이해 할 수 있다.
- Tree계열 앙상블 모델을 사용 할 수 있다.
- Grid search를 이해하고 모델의 하이퍼파라미터를 튜닝 할 수 있다.

Decision Tree Ensemble

앙상블이란?

Decision Tree Ensemble

Ensemble(앙상블)

- 앙상블(ensemble)은 여러 머신러닝 모델을 연결하여 더 강력한 모델을 만드는 기법

Decision Tree Ensemble

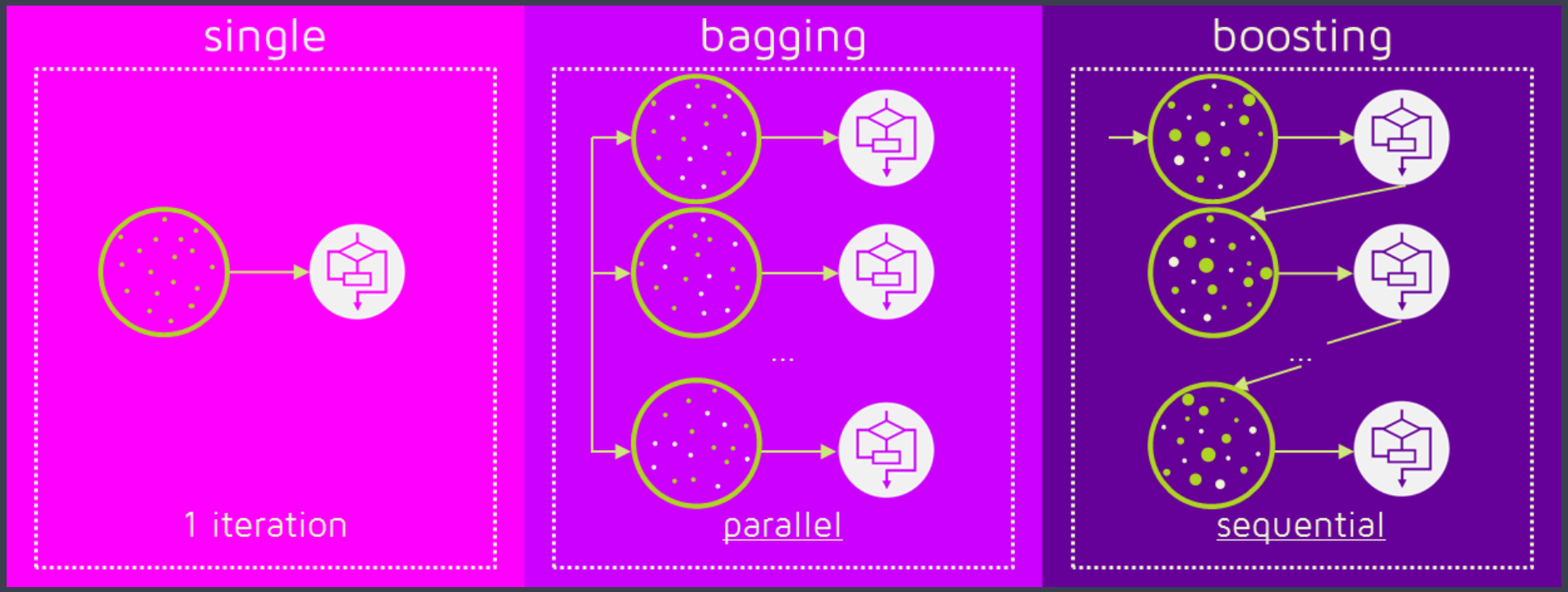
Decision Tree Ensemble(결정트리 앙상블)

- 개별 결정트리의 과대적합되는 단점을 보완하는 모델
- 다수결 법칙 또는 평균등으로 통합하여 예측 정확성을 향상
- 결정트리 모델들이 서로 독립적
- 결정트리 모델들이 무작위 예측을 수행하는 모델보다 성능이 좋을경우

→ 서로 독립적인 다양한 모델을 만들자

Decision Tree Ensemble

배깅(Bagging) VS 부스팅(Boosting)



Decision Tree Ensemble - RandomForest

RandomForest

- 서로 다른 방향으로 과대적합된 트리를 많이 만들고 평균을 내어 일반화 시키는 모델

- 다양한 트리를 만드는 방법 두 가지
  - 트리를 만들 때 사용하는 데이터 포인트 샘플을 무작위로 선택한다.
  - 노드 구성시 기준이 되는 특성을 무작위로 선택하게 한다.

Decision Tree Ensemble - RandomForest

장단점 및 주요 매개변수(Hyperparameter)

- 생성 할 트리의 개수 : n_estimators
- n개의 데이터 부트스트랩 샘플 구성
  (n개의 데이터 포인트 중 무작위로 n 횟수만큼 반복 추출, 중복 된 데이터가 들어 있을 수 있다.)
- 무작위로 선택될 후보 특성의 개수 : max_features
  (각 노드 별로 max_features 개수 만큼 무작위로 특성을 고른 뒤 최선의 특성을 찾는다.)
- max_features를 높이면 트리들이 비슷해진다.

Decision Tree Ensemble - RandomForest

장단점 및 주요 매개변수(Hyperparameter)

- 결정트리의 단점을 보완하고 장점은 그대로 가지고 있는 모델이어서 별다른 조정 없이도 괜찮을 결과를 만들어낸다.
- 트리가 여러 개 만들어지기 때문에 비전문가에게 예측과정을 보여주기는 어렵다.
- 랜덤하게 만들어지기 때문에 random_state를 고정해야 같은 결과를 볼 수 있다.

Decision Tree Ensemble - RandomForest

장단점 및 주요 매개변수(Hyperparameter)

- 텍스트 데이터와 같은 희소한 데이터에는 잘 동작하지 않는다.
- 큰 데이터 세트에도 잘 동작하지만 훈련과 예측이 상대적으로 느리다.
- 트리 개수가 많아질 수록 시간이 더 오래 걸린다.

Decision Tree Ensemble - GradientBoosting

GradientBoosting

- 정확도가 낮더라도 얕은 깊이의 모델을 만든 뒤, 나타난 예측 오류를 두 번째 모델이 보완한다.
- 이전 트리의 예측 오류를 보완하여 다음 트리를 만드는 작업을 반복한다.
- 마지막까지 성능을 쥐어짜고 싶은 경우 사용한다, 주로 경진 대회에서 많이 활용.
  (GradientBoosting을 더 발전시킨 XGBoost도 있음)

Decision Tree Ensemble - GradientBoosting

장단점 및 주요 매개변수(Hyperparameter)

- 보통 트리의 깊이를 깊게하지 않기 때문에 예측 속도는 비교적 빠르다. 하지만 이전 트리의 오차를 반영해서 새로운 트리를 만들기 때문에 학습속도가 느리다.
- 특성의 스케일을 조정하지 않아도 된다.
- 희소한 고차원 데이터에는 잘 동작하지 않는다.

Decision Tree Ensemble - GradientBoosting

장단점 및 주요 매개변수(Hyperparameter)

- 생성 할 트리의 개수 : n_estimators
  (트리가 많아질 수록 과대적합이 될 수 있다.)
- 오차를 보정하는 정도 : learning_rate
  (값이 높을 수록 오차를 많이 보정하려고 한다.)
- 트리의 깊이 : max_depth
  (일반적으로 트리의 깊이를 깊게 설정하지 않는다.)

Decision Tree Ensemble

Titanic 데이터 활용
Decision Tree Ensemble 분류 실습

Grid Search

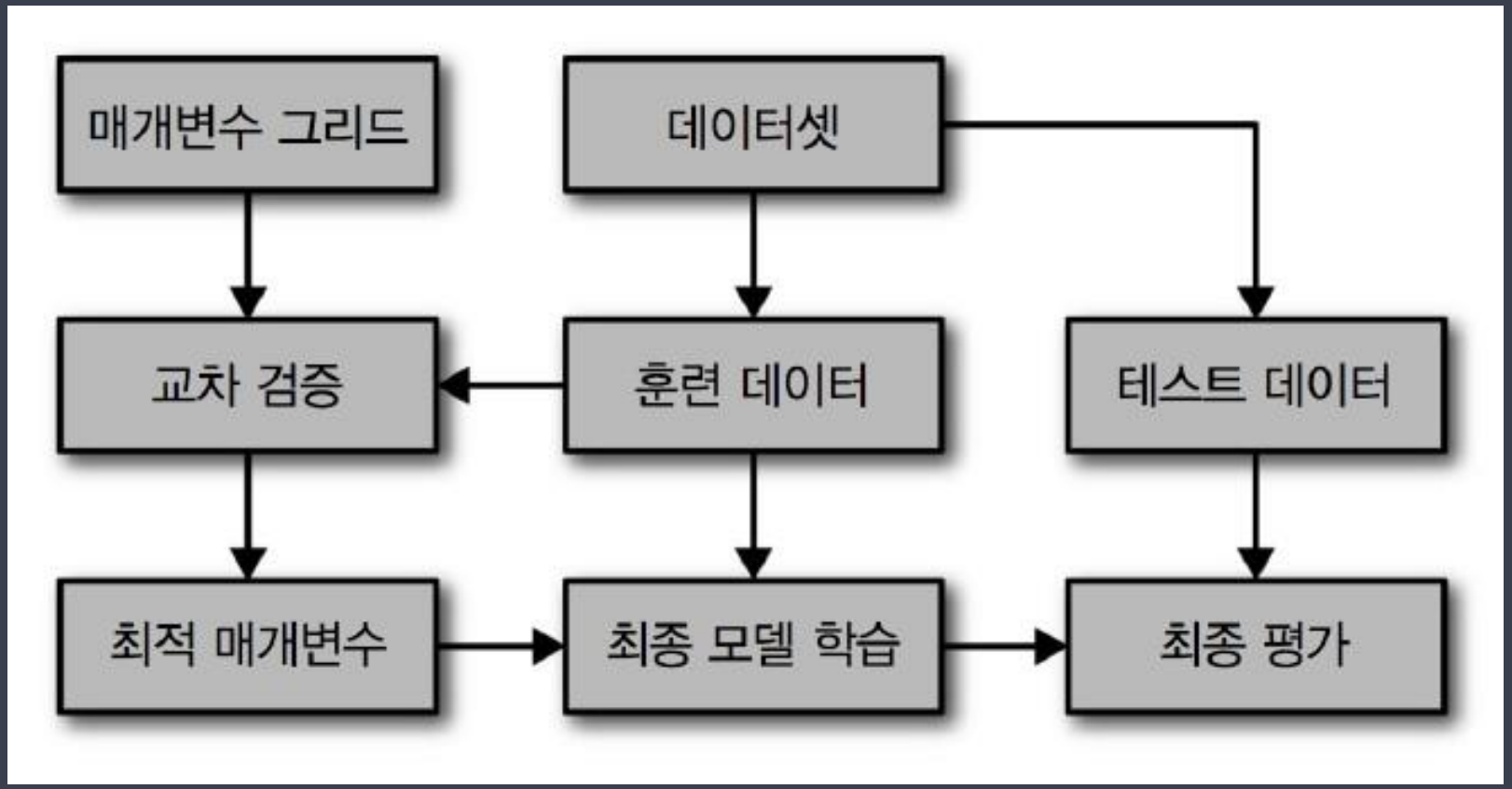
Grid Search

Grid Search

하이퍼파라미터를 여러 개 조정하여 모델을 만들 경우
사용하는 방법 **(하이퍼파라미터 튜닝)**

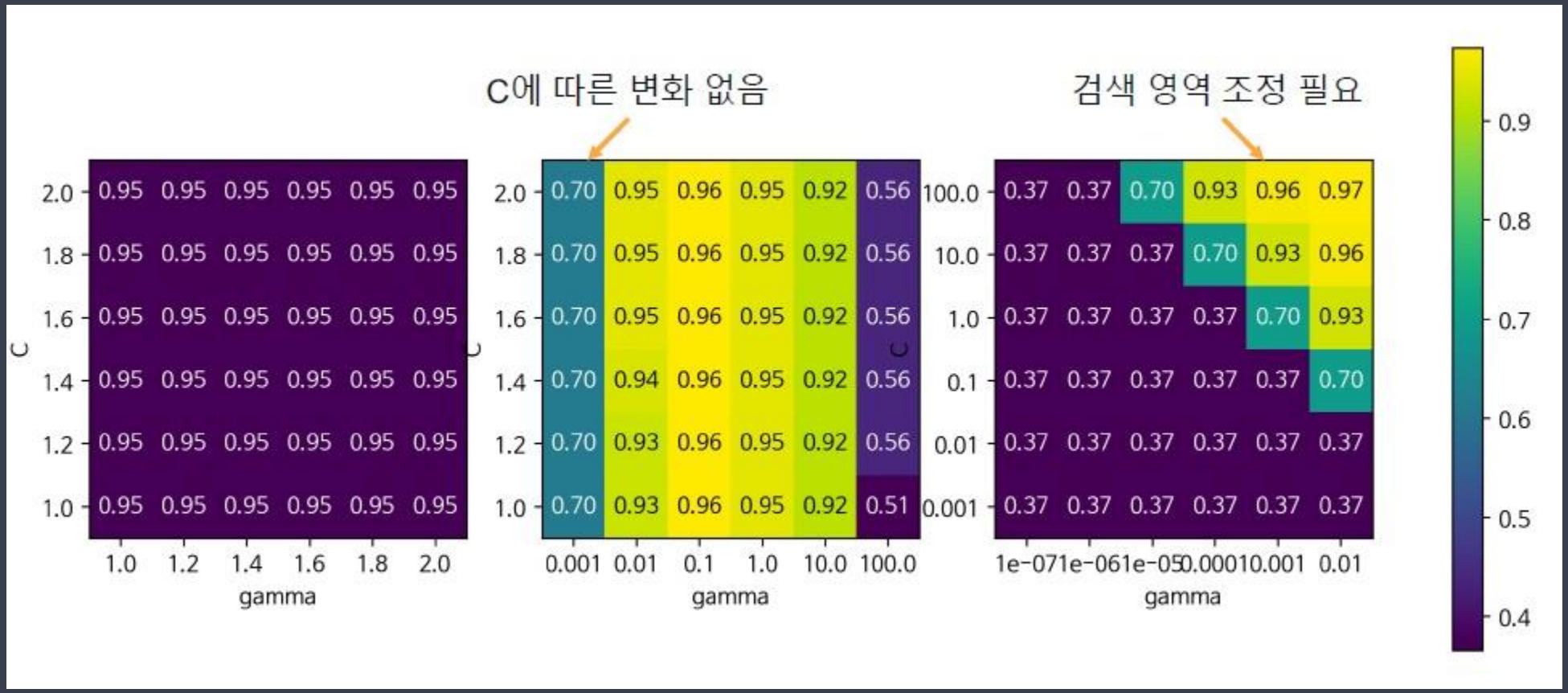
Grid Search

하이퍼파라미터 튜닝 전체 과정



Grid Search

하이퍼파라미터 튜닝 결과 분석



Grid Search

Grid search vs Random search