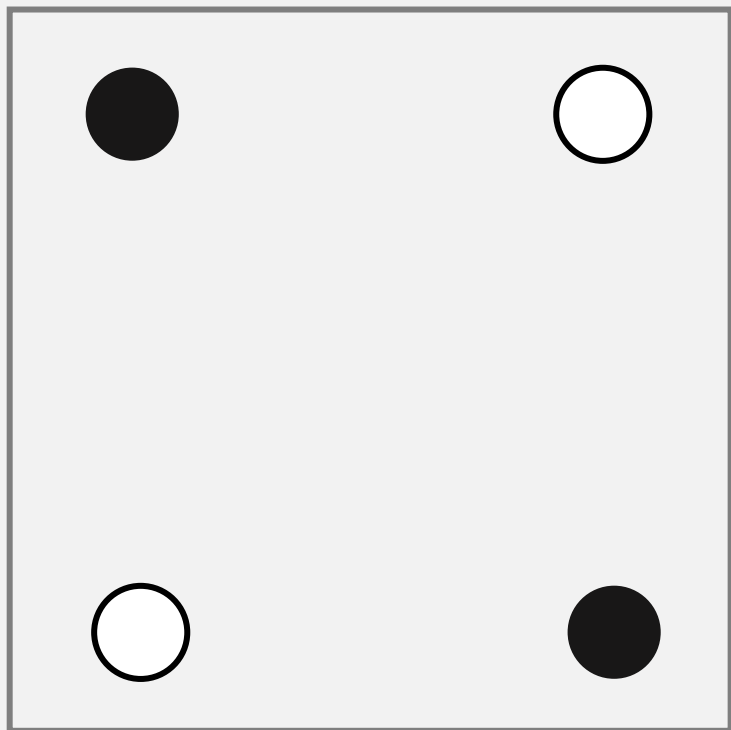


인공지능 기초

머신러닝

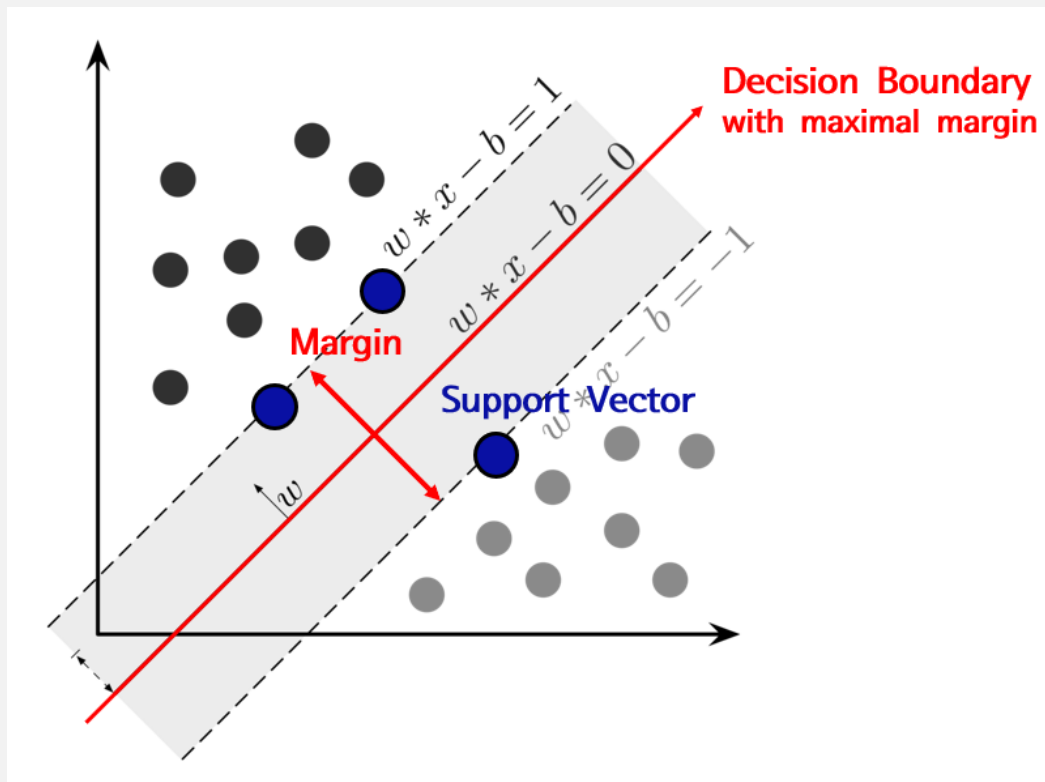
인공지능_ Day06

퍼셉트론의 과제



- 직선을 하나 그어서 직선의 한쪽 편에는 검은 점, 다른 한쪽에는 흰 점만 있도록 해 보시오.
- 이것이 퍼셉트론의 한계를 설명할 때 등장하는 XOR(exclusive OR) 문제

SVM

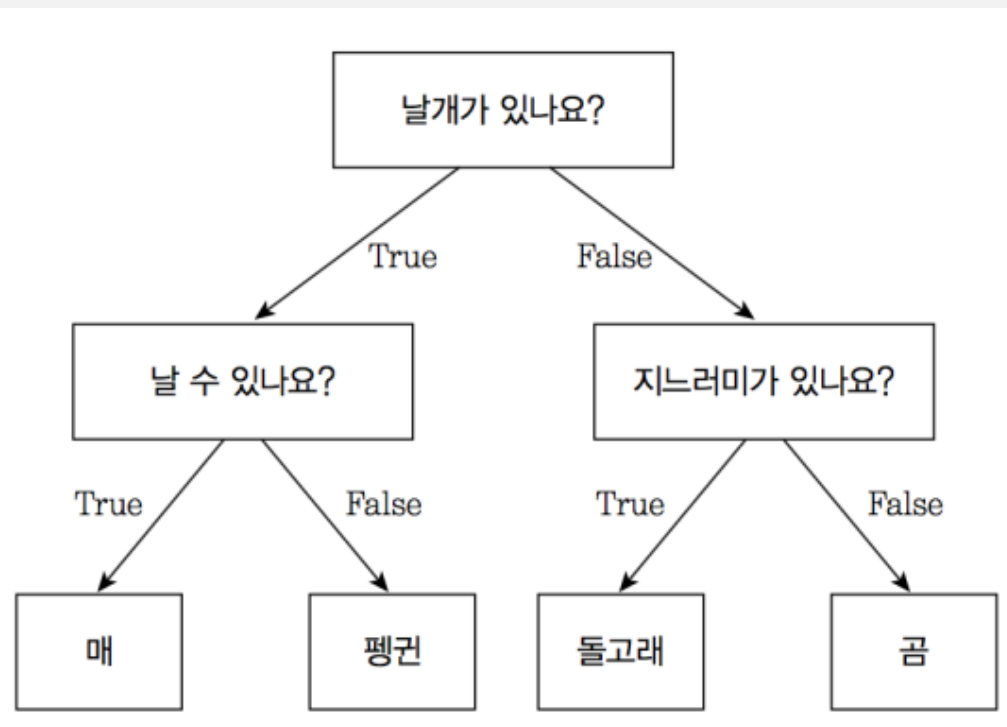


출처 <https://liveyourit.tistory.com/62>

Figure 1. SVM(Support Vector Machine)

- 서포트 벡터 머신은 여백(Margin)을 최대화하는 지도 학습 알고리즘
- 여백(Margin)은 주어진 데이터가 오류를 발생시키지 않고 움직일 수 있는 최대 공간
- 분류를 위한 서포트 벡터 머신 SVC
- 회귀를 위한 서포트 벡터 머신 SVR

Decision Tree

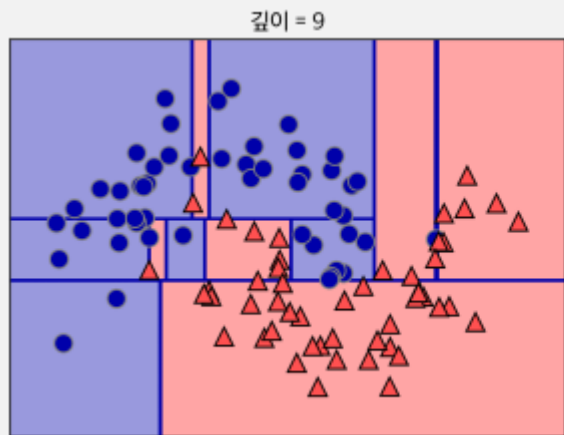
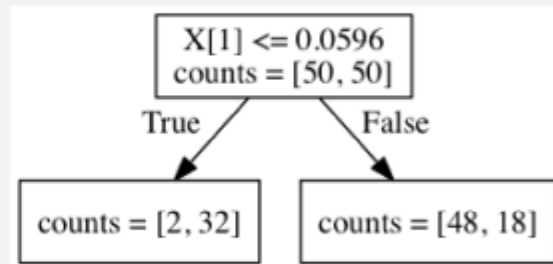
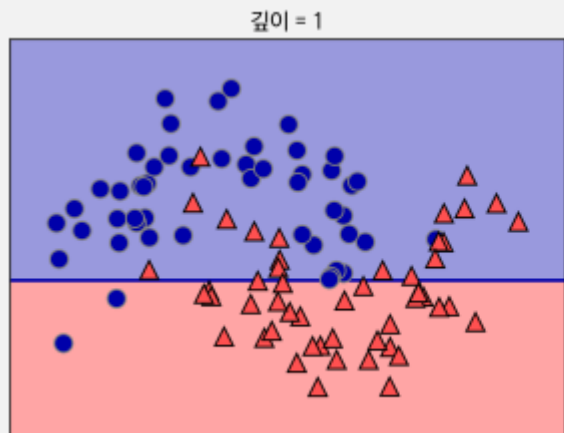


출처 텐서플로우 블로그_결정 트리

- 분류와 회귀 문제에 널리 사용하는 모델
- 기본적으로 결정 트리는 결정에 다다르기 위해 예/아니오 질문을 이어 나가면서 학습

Figure 2. 결정 트리|decision tree

Decision Tree

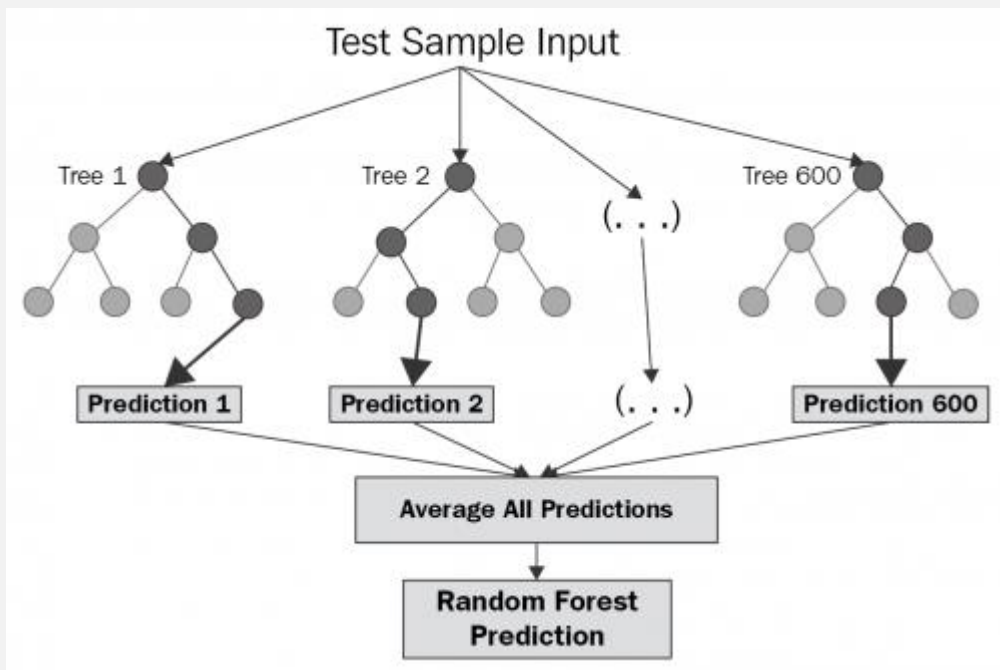


- scikit-learn에서 결정 트리는 DecisionTreeRegressor와 DecisionTreeClassifier에 구현되어 있음

출처 텐서플로우 블로그_결정 트리

Figure 3. 깊이 1과 9인 결정 트리decision tree

Random Forest



출처 IT위키_랜덤 포레스트

Figure 4. Random Forest

- 여러 개의 결정 트리들을 임의적으로 학습하는 방식의 앙상블 방법
- 여러가지 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법
- 특징
 - 임의성: 서로 조금씩 다른 특성의 트리들로 구성
 - 비상관화: 각 트리들의 예측이 서로 연관되지 않음
 - 견고성: 오류가 전파되지 않아 노이즈에 강함
 - 일반화: 임의화를 통한 과적합 문제 극복

Boosting 계열의 모델

Boosting 이란?

- Boosting이란 약한 분류기를 결합하여 강한 분류기를 만드는 과정
- 각 0.3의 정확도를 가진 A, B, C를 결합하여 더 높은 정확도,
예를 들어 0.7 정도의 accuracy를 얻는 게 앙상블 알고리즘의 기본 원리
- Boosting은 이 과정을 순차적으로 실행
A 분류기를 만든 후,
그 정보를 바탕으로 B 분류기를 만들고,
다시 그 정보를 바탕으로 C 분류기를 만듦

Adaptive Boosting (AdaBoost)

- 다수결을 통한 정답 분류 및 오답에 가중치 부여

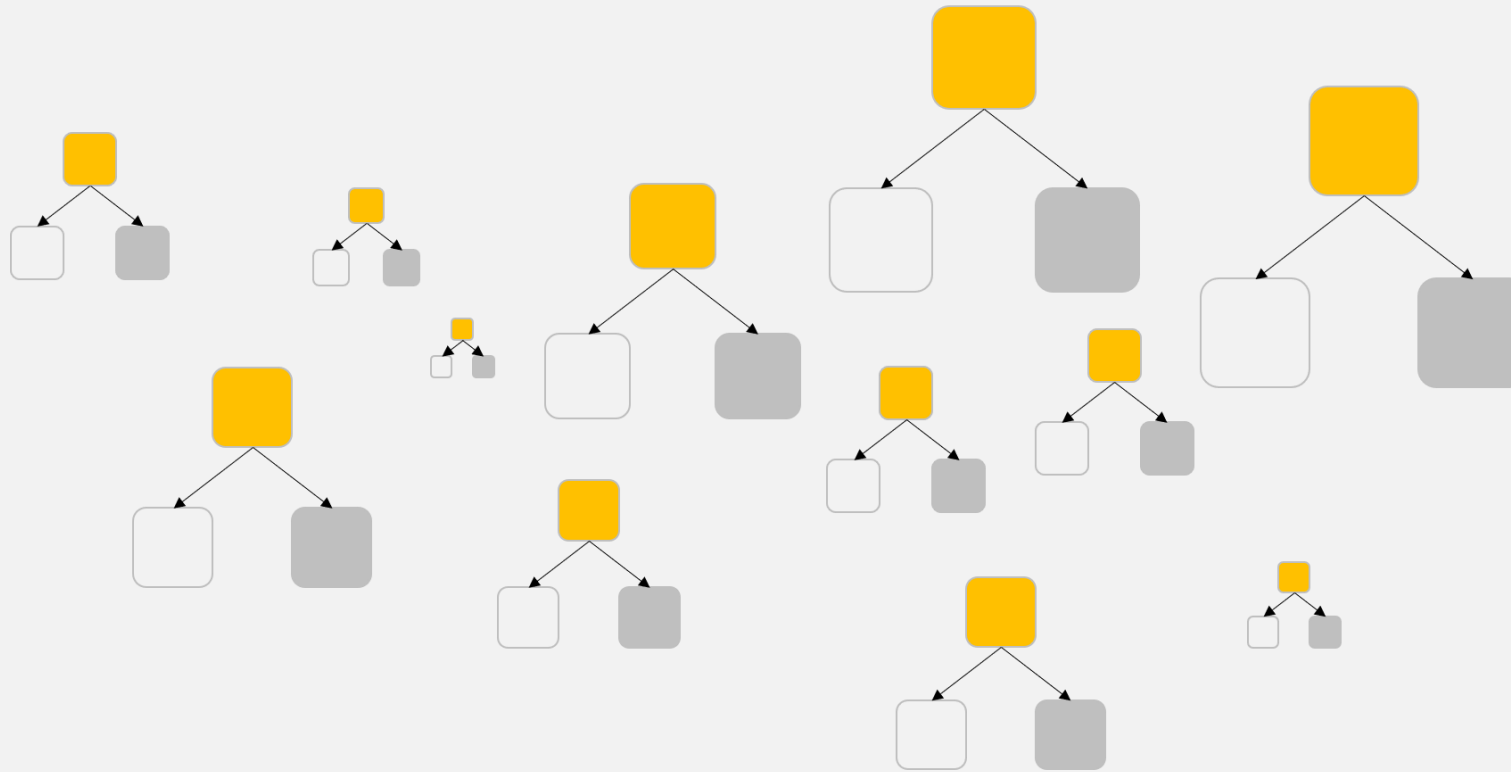
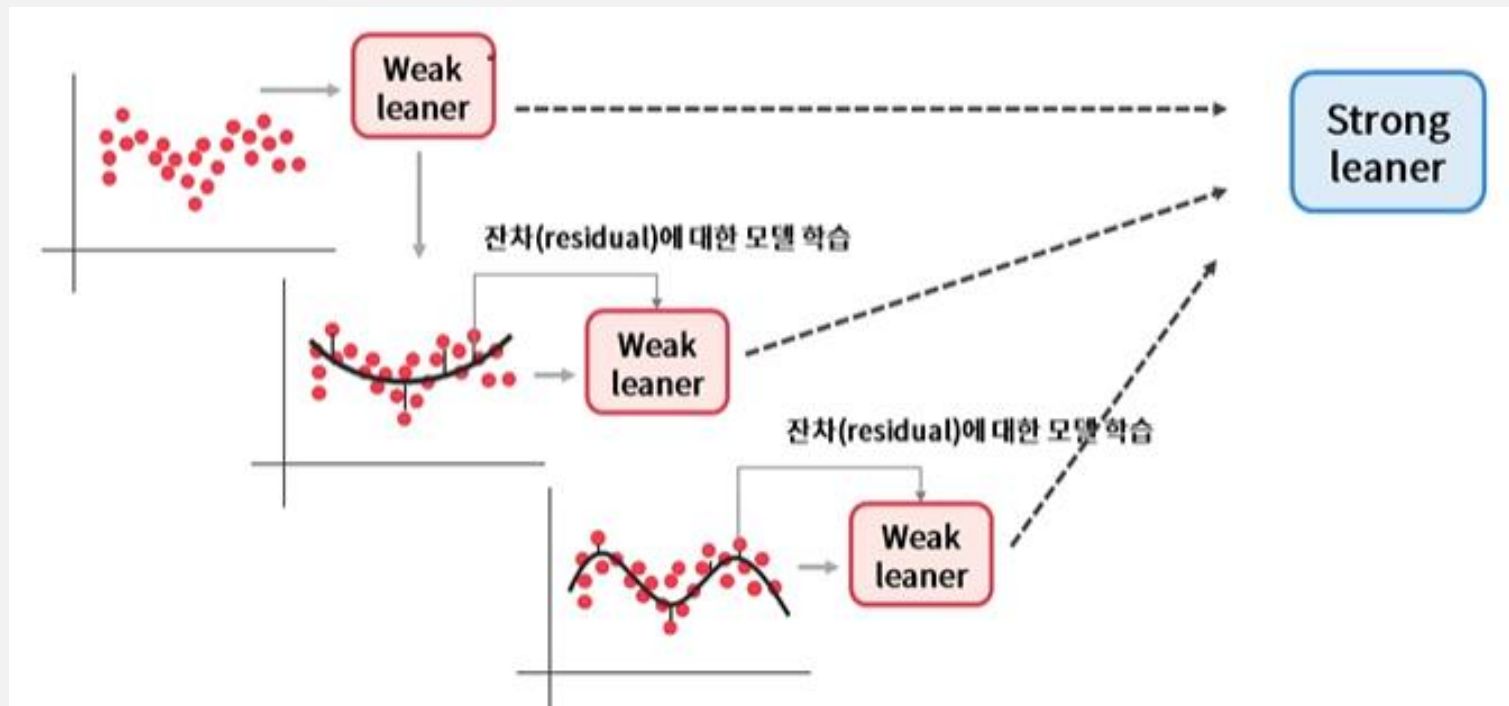


Figure 5. Adaptive Boosting (AdaBoost)

Gradient Boosting Model (GBM)

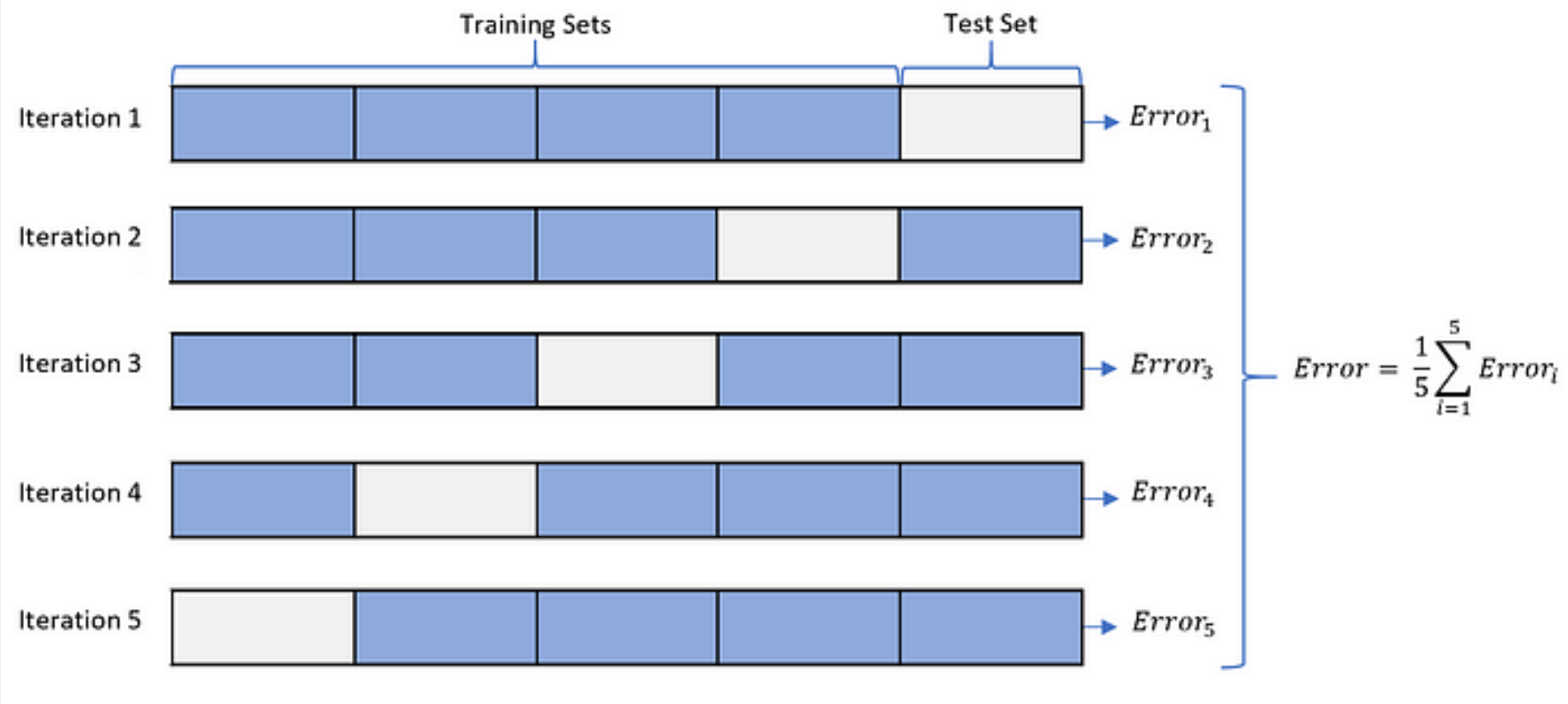
- Loss Function의 gradient를 통해 오답에 가중치 부여
- LightGBM, CatBoost, XGBoost - Gradient Boosting Algorithm을 구현한 패키지



출처 <https://velog.io/@hhhs101/Ensemble>

Figure 6. Gradient Boosting Model (GBM)

K-Fold



- ML 모델에서 가장 보편적으로 사용되는 교차 검증 기법
- K개의 데이터 폴드 세트를 만들어서 K번만큼 각 폴드 세트에 학습과 검증 평가를 수행

1. KFold

- 회귀 문제에서의 교차 검증

2. StratifiedKFold

- 레이블 데이터가 왜곡되었을 경우
- 일반적으로 분류에서의 교차 검증

스케일링 (Scaling)

Normalization (정규화)

- 특성들을 특정 범위(주로 $[0, 1]$) 로 스케일링 하는 것
- 가장 작은 값은 0, 가장 큰 값은 1 로 변환되므로, 모든 특성들은 $[0, 1]$ 범위를 갖게 됨

Standardization (표준화)

- 특성들의 평균을 0, 분산을 1 로 스케일링 하는 것
- 즉, 특성들을 정규분포로 만드는 것

주의사항

- 훈련 데이터에는 `fit_transform()` 메서드를 적용
- 테스트 데이터에는 `transform()` 메서드를 적용

1. MinMaxScaler()

- Min-Max Normalization 이라고도 불리며 특성들을 특정 범위([0,1]) 로 스케일링
- 이상치에 매우 민감하며, 분류보다 회귀에 유용함

2. StandardScaler()

- 특성들의 평균을 0, 분산을 1 로 스케일링 즉, 특성들을 정규분포로 만드는 것
- 이상치에 매우 민감하며, 회귀보다 분류에 유용함

3. MaxAbsScaler()

- 각 특성의 절대값이 0 과 1 사이가 되도록 스케일링
- 즉, 모든 값은 -1 과 1 사이로 표현되며, 데이터가 양수일 경우 MinMaxScaler 와 동일함

4. RobustScaler()

- 평균과 분산 대신에 중간 값과 사분위 값을 사용
- 이상치 영향을 최소화할 수 있음

실습

1. XOR 문제의 해결
2. SVM 모델 (SVC, SVR)
3. Linear 모델 (Perceptron, LogisticRegression은 분류모델, LinearRegression 회귀모델)
4. Tree 모델 (DecisionTreeClassifier, DecisionTreeRegressor)
5. Ensemble 모델 (RandomForestClassifier, RandomForestRegressor)
6. All_Estimator
7. KFold 와 StratifiedKFold
8. Feature Importances

Day06. 인공지능 Study

1. 인공지능 개념 정리 - 머신러닝, 딥러닝
2. 퍼셉트론 (Perceptron)
3. 다층 퍼셉트론 (Multi-Layer Perceptron: MLP)
4. 옵티마이저 (Optimizer)
5. 학습률 (learning rate)
6. 경사하강법 (Gradient Descent)
7. 손실함수 (Loss Function)
8. 활성화 함수 (Activation Function) - Sigmoid, ReLU, Softmax

- 9. 회귀분석
- 10. 결정계수 R2 score
- 11. 분류분석
- 12. 원 핫 인코딩 (One Hot Encoding)
- 13. 난수값 (random_state)
- 14. 정확도 accuracy score
- 15. 과적합 (overfitting)
- 16. 합성곱신경망(CNN)
- 17. 이미지증강(ImageDataGenerator)

18. 자연어처리(Word Embedding)

19. SVM model

20. Decision Tree model

21. K-Fold

수고하셨습니다.