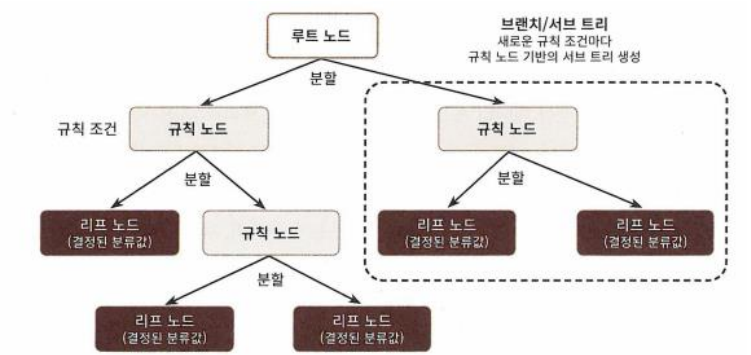
결정 트리

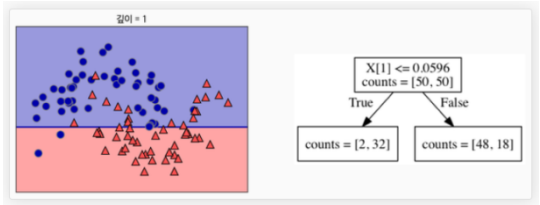
학부연구생 이현재

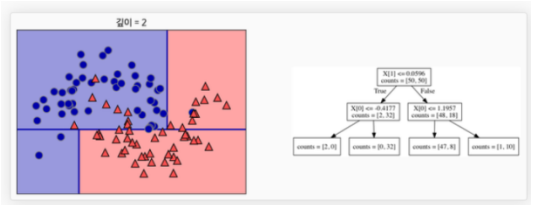
결정 트리 (Decision Tree)

- 정의 : 데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리 기반의 **분류** 규칙을 만드는 것
  - If, else 기반으로 규칙 표현, 스무고개
  - 규칙 노드(Decision Node) : 규칙 조건
  - 리프 노드(Leaf Node) : 결정된 클래스 값
  - 새로운 규칙 조건마다 서브 트리(Sub Tree) 생성
  - 많은 규칙이 있다는 것은 분류를 결정하는 방식이 더욱 복잡→**과적합** → 트리의 깊이가 깊어질수록 결정 트리의 예측 성능 저하



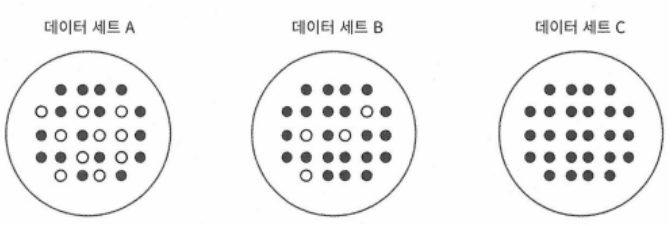
- 가능한 적은 규칙 노드로 높은 예측 정확도를 가지려면 데이터를 분류할 때 최대한 많은 데이터 세트가 해당 분류에 속할 수 있도록 규칙 노드의 규칙이 정해져야 한다.
  - 최대한 **균일**한 데이터 세트를 구성할 수 있도록 분할하는 것이 중요





정보의 균일도

- 균일도

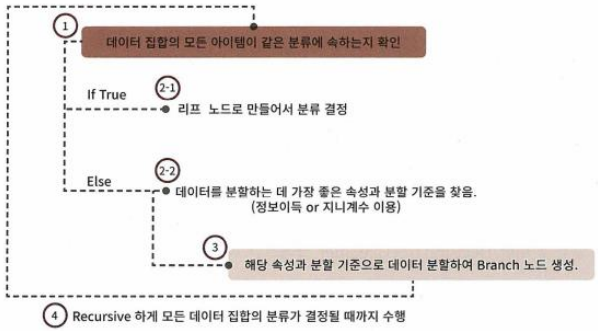


  - 균일도가 가장 높은 데이터 세트는 C, 다음이 B, 마지막으로 A
  - 데이터 세트의 균일도는 데이터를 구분하는 데 필요한 정보의 양에 영향을 미침.
  - 규칙 노드는 정보 균일도가 높은 데이터 세트를 먼저 선택할 수 있도록 규칙을 만듬

    (**균일도가 높은 것을 최우선으로 분류**, 예는 하단의 설명 참고)
  - 균일도를 측정하는 대표적인 방법은 **정보 이득 지수**와 **지니 계수**

- 정보 이득 지수
  - 1 – 엔트로피 (엔트로피가 높은 데이터 세트는 A, B, C 순)
  - 엔트로피 : 주어진 데이터 집합의 혼잡도, 불순도가 높으면 엔트로피가 높고 불순도가 낮으면 엔트로피가 낮음
  - **정보 이득이 높은 속성을 기준**으로 분할**(정보 이득이 높다 = 엔트로피가 낮다 = 불순도가 낮다 = 균일도가 높다)**

- 지니 계수
  - 머신러닝에서의 지니계수는 지니계수가 낮을수록 데이터 균일도가 높은 것으로 해석
  - **지니 계수가 낮은 속성을 기준**으로 분할**(지니계수가 낮다 = 균일도가 높다)**
  - 사이킷런의 결정트리 알고리즘은 기본으로 지니계수를 이용
    (지니 계수가 낮은 조건을 찾아서 분할)



Prof. Hyerim Bae (hrbae@pusan.ac.kr)

결정 트리 모델의 특징

- 장점
  - 정보의 균일도라는 룰을 기반으로 하고 있어서 알고리즘이 쉽고 직관적
  - 정보의 균일도만 신경쓰면 되므로 특별한 경우를 제외하고는 각 피처의 스케일링/정규화 같은 전처리 작업이 필요 없음.
  - 선형 회귀 모델과 달리 특성들간의 상관 관계가 많아도 트리 모델은 영향을 받지 않음.
  - 비선형 모델에서 깊이(depth)를 조절(하이퍼 파라미터를 조절)함으로써 모델을 학습 가능
  - 수치형, 범주형 데이터 모두 가능
  - 회귀와 분류 모두 가능
- 단점
  - 새로운 sample이 들어오면 손수무책(다시 학습)
  - 학습 정확도를 높이기 위해서는 모델을 더 복잡하게 만들어야 하는데 그러다 보면 과적합에 걸리기 매우 쉬움
- 결정트리의 파라미터

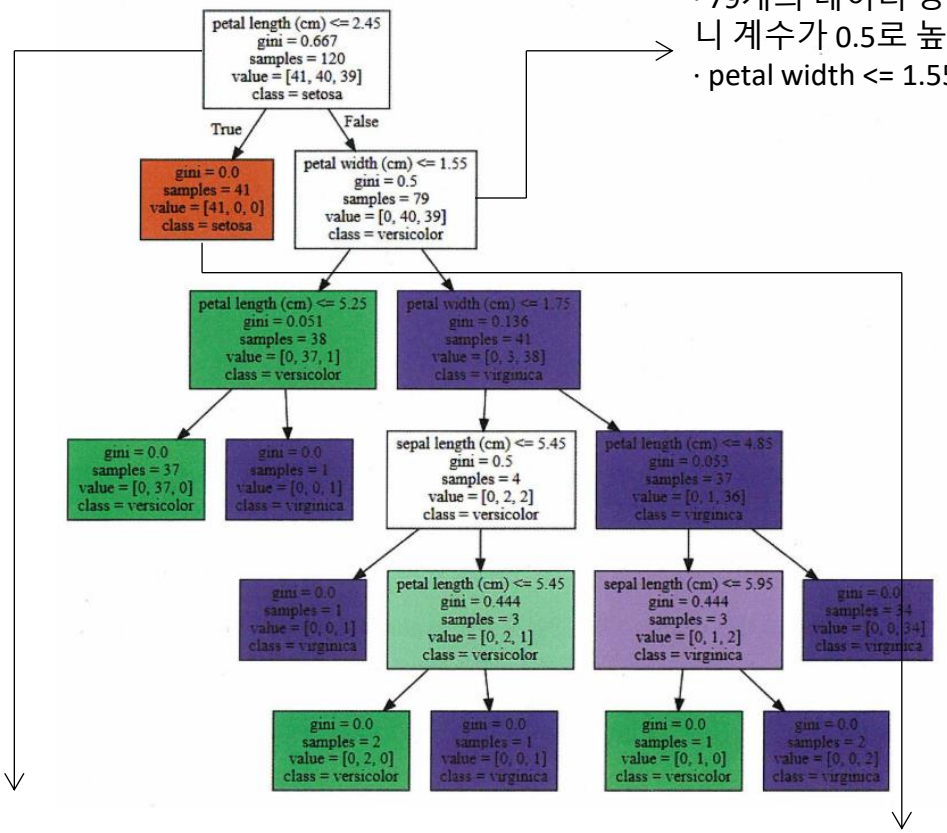
파라미터명	설명
min_samples_split	샘플이 최소한 몇개 이상이어야 split(하위(잎)노드로 분리)할것인가/클수록 과적합방지, 작을수록 정확하게 분리되어 과적합
min_samples_leaf	(잎) 노드가 되려면 가지고 있어야 할 최소 샘플의 수/ 클수록 과대적합방지, 작을수록 과적합/ min_samples_split을 만족해도 min_sample_leaf을 만족하지 않으면 leaf노드가 되지 못한다.
max_depth	얼마나 깊게 트리를 만들것인가/None이면 최대한 깊게(불순도가 0이 될때까지)/클수록 과적합,작을수록 과적합방지
max_leaf_nodes	최대 몇개 잎 노드가 만들어 질때 까지 split(하위(잎)노드로 분리)할것인가
max_feature	최적의 분할을 위해 고려할 최대 피처개수

정지 규칙
–더 이상 분리가 일어나지 않는 기준

1. 더 이상 분리해도 불순도가 줄어들지 않을 때
2. 자식 마디에 남은 sample 수가 너무 적을 때
3. 분석자가 지정한 규제 매개변수에 도달했을 때

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

결정 트리 시각화



· 79개의 데이터 중 Vesicolor 40개, Virginica 39개로 여전히 지니 계수가 0.5로 높으므로 분기할 규칙 필요
· petal width <= 1.55 규칙으로 자식 노드 생성

• 'petal length(cm) <= 2.45'와 같이 피처의 조건이 있는 것은 자식 노드를 만들기 위한 규칙 조건. 이 조건이 없으면 리프 노드.

• gini는 다음의 value=[]로 주어진 데이터 분포에서의 지니 계수

• samples는 현 규칙에 해당하는 데이터 건수

• value=[]는 클래스 값 기반의 데이터 건수. 붓꽃 데이터 세트는 클래스 값으로 0, 1, 2를 가지고 있으며, 0:Setosa, 1:Versicolor, 2:Virginica 품종을 의미. 만일 Value = [41,40,39]라면 클래스 값의 순서로 Setosa 41개, Vesicolor 40개, Virginica 39로 데이터가 구성

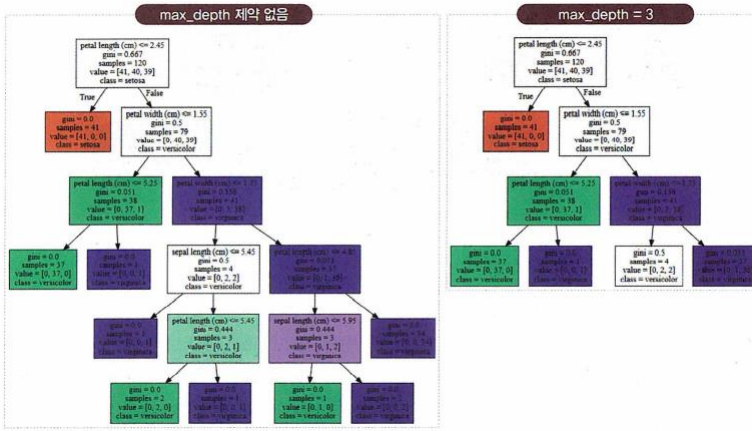
· samples=120은 전체 데이터 개수가 120개
· sample 120개가 41,40,39로 분포되어 있어 지니 계수가 0.667
· petal length <= 2.45 규칙으로 자식 노드 생성
· class = Setosa는 하위 노드를 가질 경우 Setosa의 개수가 41개로 가장 많다는 의미

· 모든 데이터가 Setosa로 결정되므로 클래스가 결정된 리프 노드가 되고 더 이상 규칙을 만들 필요가 없음
· 41개의 데이터가 모두 Setosa이므로 예측 클래스는 Setosa
· 지니 계수는 0

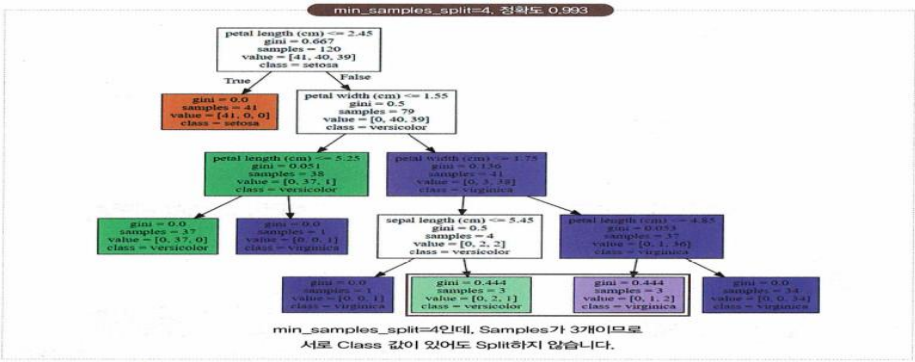
Prof. Hyerim Bae (hrbae@pusan.ac.kr)

결정 트리 시각화

- max_depth
  - 결정 트리는 규칙 생성 로직을 미리 제어하지 않으면 완벽하게 클래스를 구분하기 위해서 다시 자식 노드 생성
  - 이로 인해 매우 복잡한 결정 트리가 만들어져 모델이 쉽게 과적화



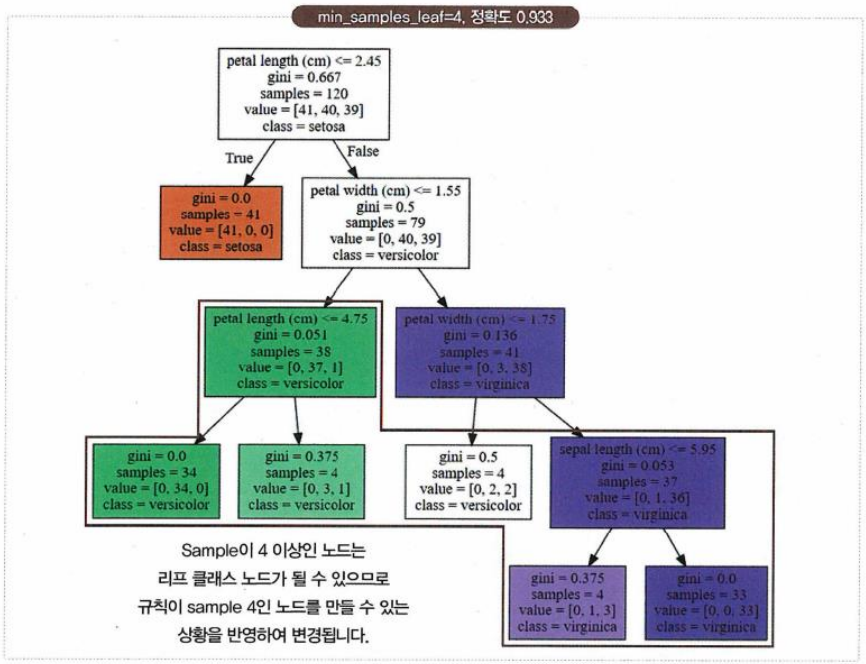
- min_samples_split
  - 자식 규칙 노드를 분할해 만들기 위한 최소한 샘플 데이터 개수
  - Sample이 3개인데, 이 노드 안에 value가 [0,2,1]과 [0,1,2]로 서로 상이한 클래스 값이 있어도 더 이상 분할하지 않고 리프 노드가 된 것을 확인(min_samples_split가 4로 규정)



자연스럽게 트리 깊이도 줄고 간결한 트리 형성

결정 트리 시각화

- min_samples_leaf
  - 리프 노드가 될 수 있는 샘플 데이터 건수의 최솟값(default는 1)
  - Default는 1, 다른 클래스 값이 하나도 없이 단독 클래스로만 돼 있거나 단 한 개의 데이터로 돼 있을 경우에만 리프 노드가 될 수 있음.
  - min_samples_leaf의 값을 키우면 더 이상 분할하지 않고, 리프 노드가 될 수 있는 조건이 완화
  - 즉, min_samples_leaf <= 지정값의 기준만 만족하면 리프노드가 될 수 있음



- min_samples_leaf = 4로 설정하면 샘플이 4 이하이면 리프 노드가 되기 때문에 지니 계수 값이 크더라도 샘플이 4인 조건으로 규칙 변경을 선호하게 되어 자연스럽게 트리 깊이가 낮아지고 간결하게 만들어짐

서포트 벡터 머신 모델 코드 구현

https://github.com/LeeYunseol/Lab_study/blob/main/%EA%B2%B0%EC%A0%95%ED%8A%B8%EB%A6%AC/%EA%B2%B0%EC%A0%95%ED%8A%B8%EB%A6%AC.ipynb

참고 자료

- Pytorch로 시작하는 딥러닝 입문
- 귀퉁이 서재 블로그(https://bkshin.tistory.com/)