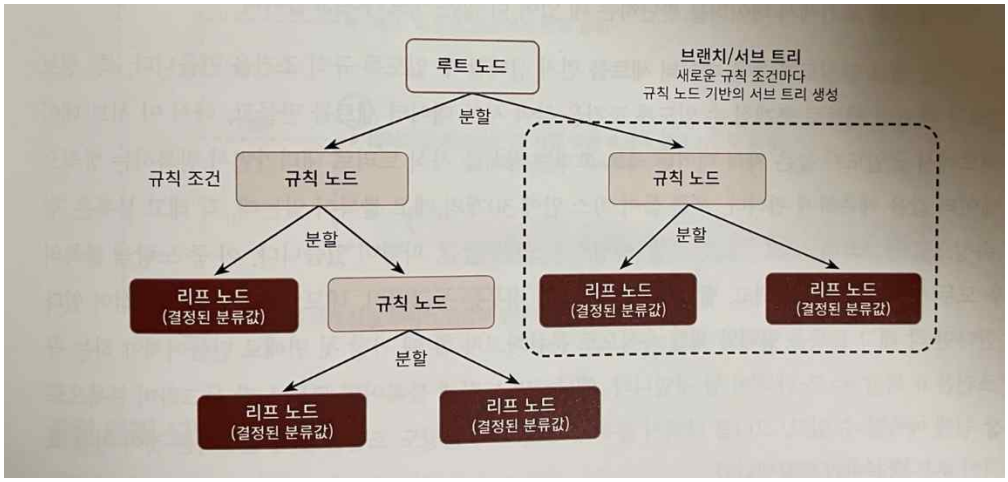
4장

2. 결정 트리

- 결정 트리: 데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리 기반의 분류 규칙을 만드는 것
- 결정 트리의 구조:



- 규칙이 많을수록, 즉 트리의 깊이가 깊어질수록 결정 트리의 예측 성능이 저하될 수 있다. 정보 균일도가 높은 데이터 세트를 먼저 선택할 수 있도록 규칙 조건을 만들어야 한다.
- 정보의 균일도를 측정하는 대표적인 방법은 엔트로피를 이용한 정보 이득 지수와 지니 계수가 있다
  - 엔트로피: 주어진 데이터 집합의 혼잡도
  - 지니 계수: 불평등 지수. 지니 계수가 낮을수록 데이터 균일도가 높다
- 결정 트리의 장점:
  - 쉽다, 직관적이다
  - 피처의 스케일링이나 정규화 등의 사전 가공 영향도가 크지 않다
- 결정 트리의 단점
  - 과적합으로 성능이 떨어질 수 있다
- 결정 트리 과적합
  - 이상치까지 분류하는 복잡한 모델은 학습 데이터 세트의 특성과 약간의 다른 형태의 데이터 세트를 예측하면 예측 정확도가 떨어진다.

3. 앙상블 학습

- 앙상블 학습을 통한 분류는 여러 개의 분류기를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법
- 다양한 분류기준의 예측 결과를 결합해서 단일 분류기보다 신뢰성이 높은 예측값을 얻을 수 있다.
- 앙상블 학습의 유형: 보팅(Voting), 배깅(Bagging), 부스팅(Boosting), 스태깅 등
    - 보팅 & 배깅: 여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정

      보팅은 서로 다른 알고리즘을 가진 분류기를 결합하고,

      배깅은 각각의 분류기가 모두 같은 유형의 알고리즘 기반 (예: 랜덤 포레스트)
    - 부스팅: 여러 개의 분류기가 순차적으로 학습을 수행하되, 앞에서 학습한 분류기가 예측이 틀린 데이터에 대해서는 올바르게 예측할 수 있도록 다름 분류기에게는 가중치를 부여하면서 학습과 예측을 진행
    - 스태깅: 여러 가지 다른 모델의 예측 결과값을 다시 학습 데이터로 만들어서 다른 모델로 재학습히켜 결과를 예측
    - 보팅 유형
- 보팅 유형
    - 하드 보팅: 예측한 결과값들중 다수의 분류기가 결정한 예측값을 최종 보팅 결과값으로 선정
    - 소트프 보팅: 분류기들의 레이블 값 결정 확률을 모두 더하고 이를 평균해서 확률이 가장 높은 레이블 값을 최종 보팅 결과값으로 선정 (일반적)

4. 랜덤 포레스트

- 배깅은 같은 알고리즘으로 여러 개의 분류기를 만들어서 보팅으로 최종 결정하는 알고리즘. 랜덤 포레스트가 배깅의 대표적인 알고리즘.
- 랜덤 포레스트는 여러 개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측 결정을 하게 된다.
    - 부트스트래핑: 여러 개의 데이터 세트를 중첩되게 분리
    - 서브세트 데이터: 부트스트래핑으로 임의로 만들어진 데이터