Bagging & Random Forest

김혜빈

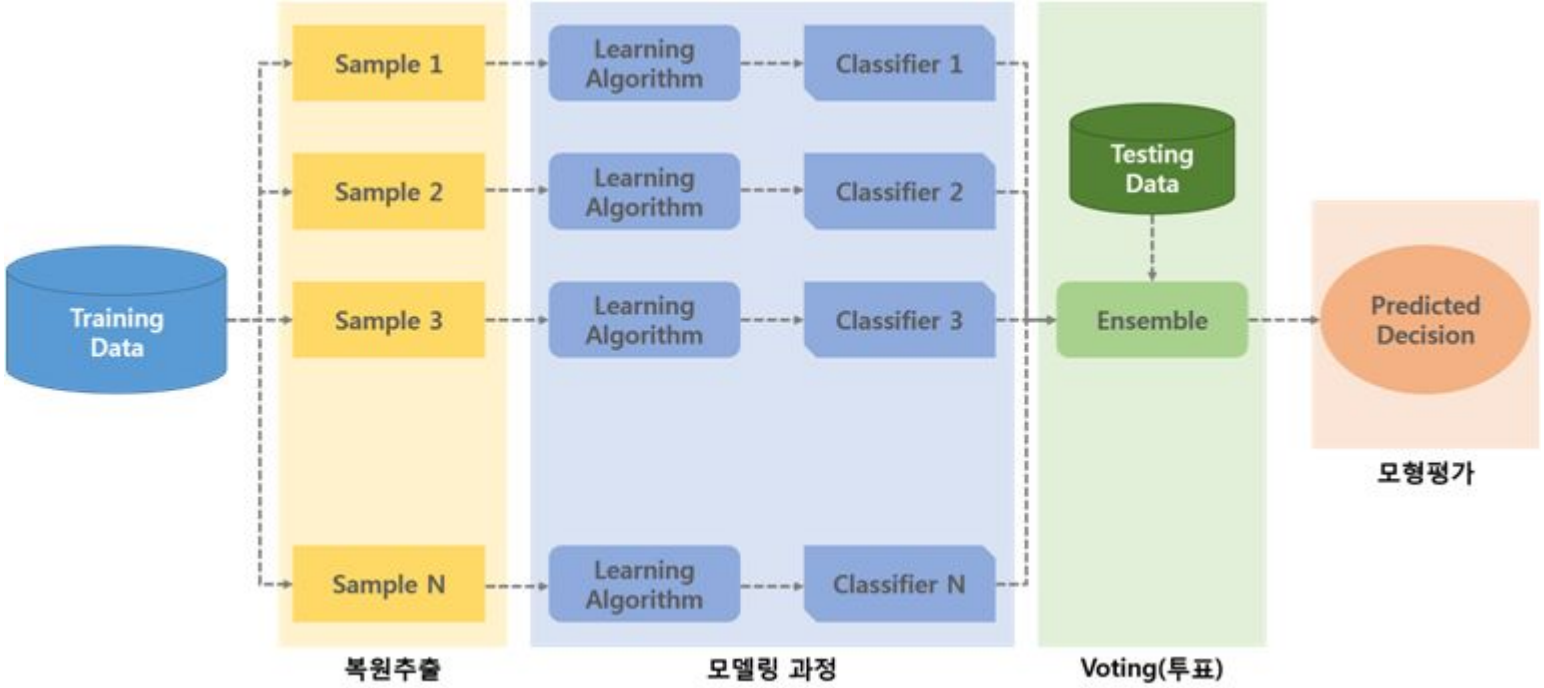
KU-BIG

Bagging

Bagging

- Bootstrap Aggregating의 줄임말
- 훈련용 데이터 집합으로부터 크기가 같은 표본을 여러 번 동일한 크기로 단순확률 반복 추출 하여 각각에 대한 분류기를 생성하고, 생성된 분류기들의 결과를 종합하여 의사결정을 내리는 방법.
- 이 때, 최종 결과값이 연속형일 경우에는 평균값, 범주형일 경우에는 투표를 통해 최종 결과값을 도출함

Bagging Algorithm

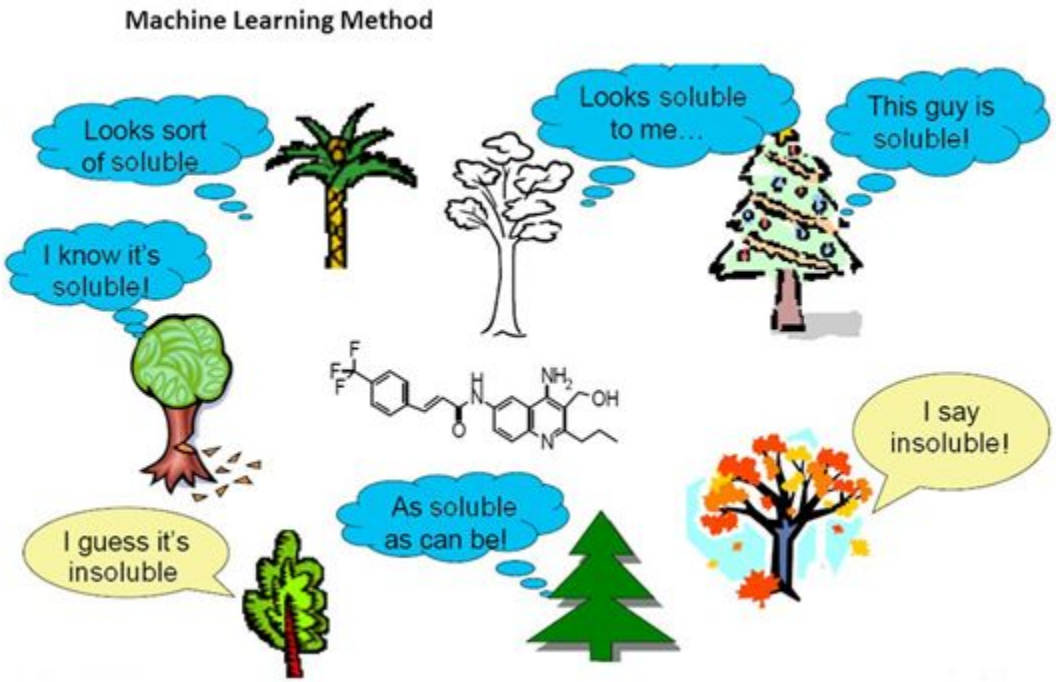


Bagging Algorithm

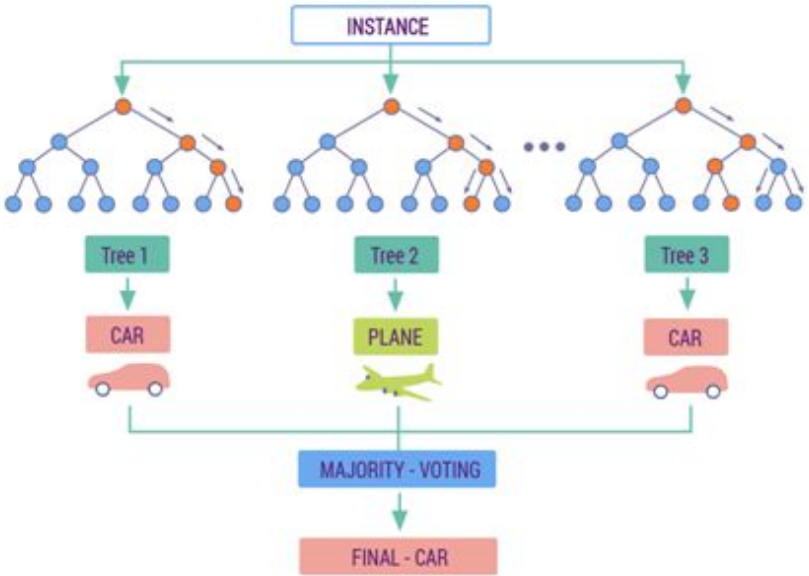
- N개의 training data에서 크기 n의 bootstrap sample을 B개 ($D1, D2, \ldots, DB$) 생성한다.
- 각각의 bootstrap sample들로 B개의 분류기를 학습시킨다.
- 모델들을 결합해 최종모델 산출

Random Forest

Random Forest



Random Forest

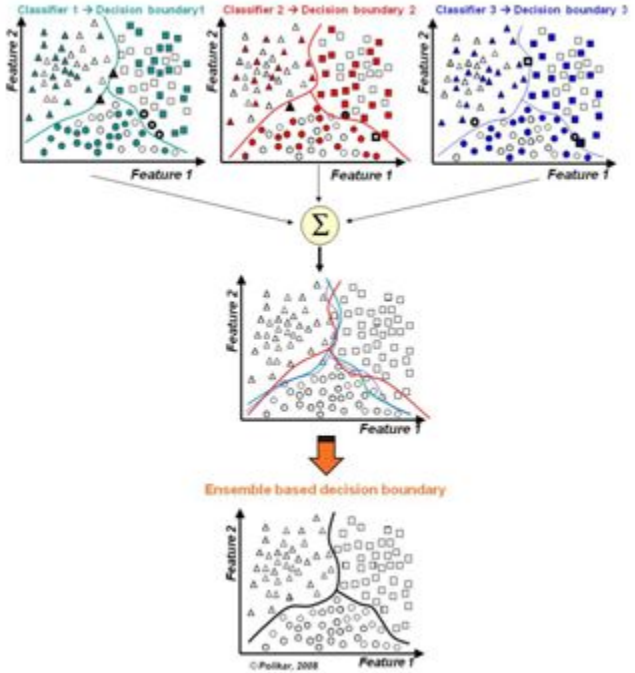


- N개의 training data에서 크기 n의 bootstrap sample을 B개 $(D1, D2, \ldots, DB)$ 생성
- 각 Sample에서 임의의 변수 추출, Bootstrap sample들로 B개의 분류기들 학습시키기

- 모델 결합해 최종모델 산출

- 숲이 완성됐어요! 와!

Random Forest



Random Forest

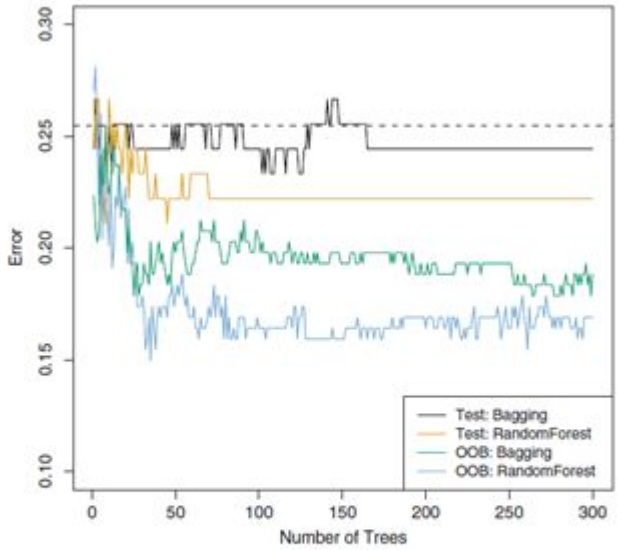


FIGURE 8.8. *Bagging and random forest results for the* Heart *data. The test error (black and orange) is shown as a function of B, the number of bootstrapped training sets used. Random forests were applied with* $m = \sqrt{p}$. *The dashed line indicates the test error resulting from a single classification tree. The green and blue traces show the OOB error, which in this case is considerably lower.*

- Bagging, Random Forest, 그리고 OOB error에 대한 그림
- Bagging 보다 Random Forest가 더 좋은 성능을 보임
- Tree의 개수가 어느정도 커지면 error율이 수렴함

  =나무의 개수는 충분한 정도 이상을 늘릴 필요는 없다


스브스뉴스
웃으면
福
이
옵니
다
감사합니다
KU-BIG 사랑해요