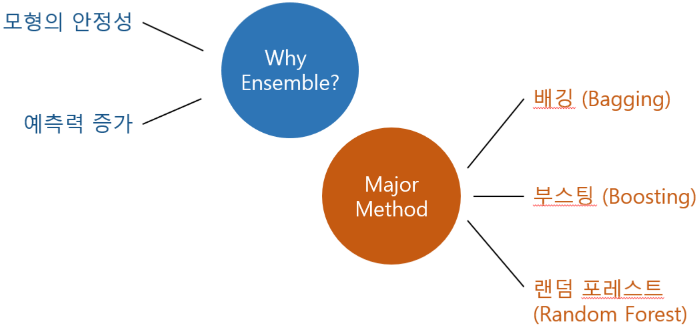
Ensemble

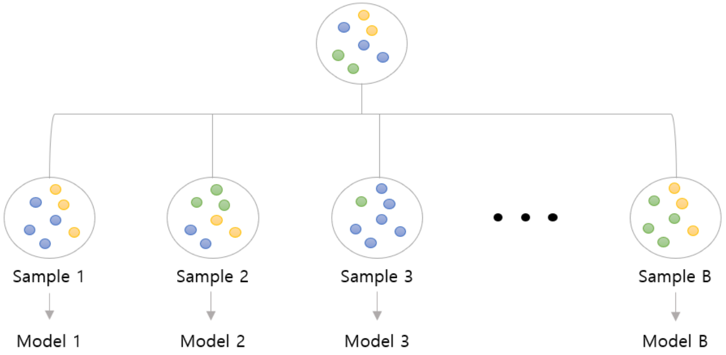
- 여러 개의 예측모형들을 결합하여 하나의 예측 모형을 만드는 방법



Bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.
- The "bootstrap data sets" is created by sampling with replacement, and is the same size as our original data set.

- Bagging (Bootstrap aggregation)
  - Bootstrap을 이용하여 모형을 생성



Bagging - 알고리즘

- $L = \{(x_i, y_i)\}_{i=1}^{n}$: training data

1. For $b = 1, \dots, B$
   1) 부스트랩 $L^{*(b)}$ 샘플 생성
   2) 각 샘플에 대하여 예측 모형 $f^{(b)}(x)$ 구축
2. B개의 예측 모형 결합 : 최종모형 $\hat{f}(x)$ 구축
   - 회귀모형 : $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{(b)}(x)$
   - 분류모형 : B개중 가장 많이 분류된 범주로 분류

Bagging

- 평균을 사용하여 불안정한 예측모형의 분산 감소
- 일부러 과적합된 예측모형에 bagging을 적용하였을 때 효과 증가
- Bagging은 여러 예측모형에 대해 유용한 방법이지만, 특히 의사결정나무에 유용하게 사용될 수 있음
- 개별 트리는 분산은 크고 편차는 작지만, B개의 트리의 결과를 평균을 하게 되면 분산을 줄이면서 예측력을 향상 시킬 수 있음
- Bagging은 최적의 트리를 구축하는 것보다 더 빠를 수도 있음 (가지치기 등의 모형 선택 단계 생략 가능)

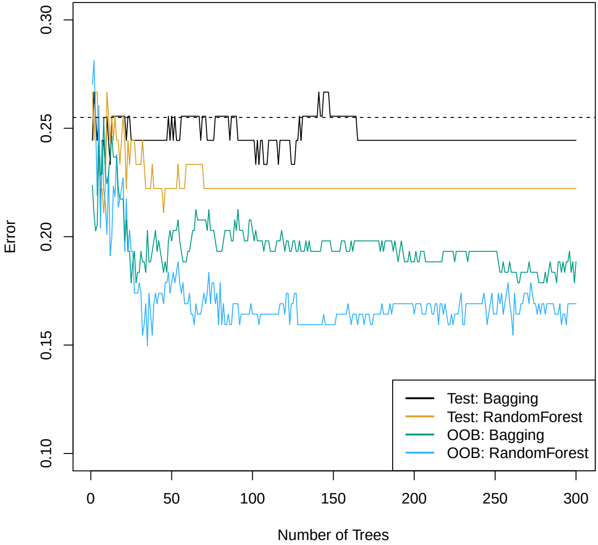
Out-of-Bag Error Estimation

- 배깅모형에서 test error를 계산하는 방법
- bootstrap sample을 추출하면 평균적으로 원래 데이터의 2/3 정도만 선택됨

$$\lim_{n\to\infty}\left(\frac{n-1}{n}\right)^n = \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1} = 0.368$$

- 선택되지 않는 나머지 데이터를 out-of-bag(OOB) 관측값이라고 함
- 이 OOB 관측값을 이용하여 test error를 계산할 수 있음

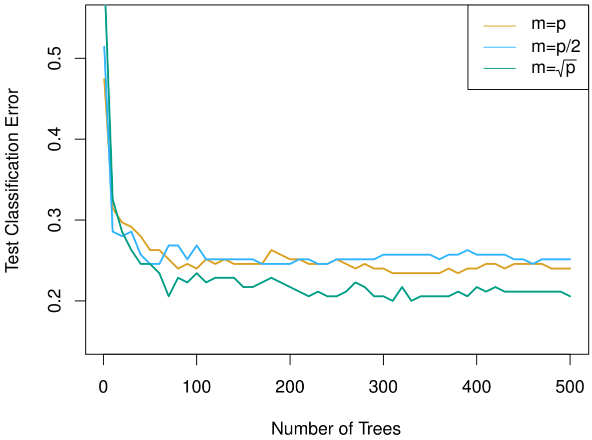
Bagging the Heart Data



Random Forest

- bootstrap sample + random sample of m predictors
- 전체적인 방법은 bagging과 동일
- 하나의 트리를 형성하는 과정에서, 각 노드에서 전체 p개의 설명변수 중 m개만을 임의로 추출하여 분리 규칙 생성
- 일반적으로 $m \approx \sqrt{p}$
- 분리에 사용하는 설명변수를 무작위로 선택함으로써 트리 사이의 상관성을 줄여 앙상블의 효과를 향상 시킬 수 있음 (분산 감소)

Random Forest



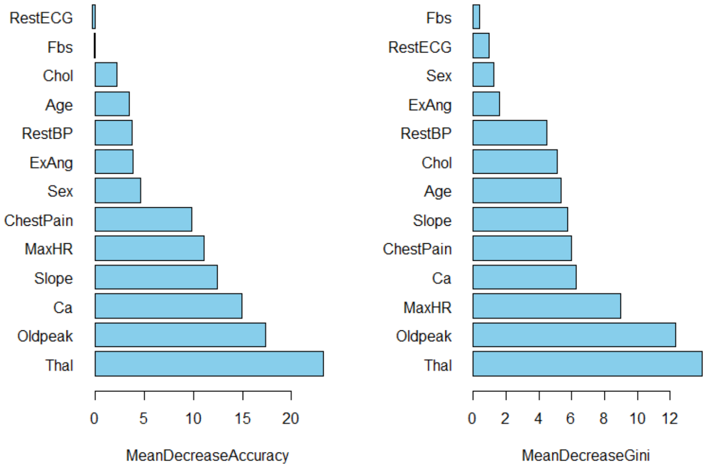
Random Forest

- 예측력은 Boosting과 비슷
- 분리규칙을 생성할 때 사용하는 변수의 숫자가 훨씬 작아지기 때문에 효율성이 높고, 배깅/부스팅보다 속도가 빠름
- 트리의 장점 : 설명력, 단점 : 정확도가 떨어짐
- 랜덤포레스트는 많은 수의 트리를 생성하기 때문에 예측력을 향상시킬 수 있지만, 트리의 강점이 사라짐
- 대신, 변수의 중요도를 측정할 수 있음

변수중요도

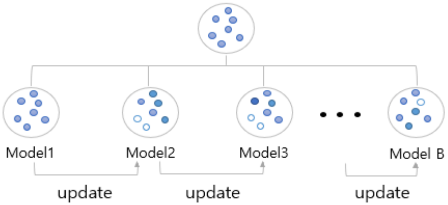
- 정확도 감소
  - b번째 트리 형성 $\rightarrow$ OOB error 측정(e_b) $\rightarrow$ OOB에서변수 X_i 를 임의로 섞음, OOB error 측정(r_i), $i = 1, ..., p$
  - $d_i = r_i - e_b, \;\; b = 1, ..., B \;\; \Rightarrow$ 변수중요도 $= \bar{d}_i / s_{d_i}$
- 불순도 감소
  - 트리 형성 과정에서 분리 규칙으로 사용된 변수들에 대해 RSS 나 불순도가 얼마나 감소했는지를 모든 B개의 트리에서 측정하여 평균한 값을 변수 중요도로 사용

변수중요도


RestECG
Fbs
Chol
Age
RestBP
ExAng
Sex
ChestPain
MaxHR
Slope
Ca
Oldpeak
Thal
0 5 10 15 20
MeanDecreaseAccuracy
Fbs
RestECG
Sex
ExAng
RestBP
Chol
Age
Slope
ChestPain
Ca
MaxHR
Oldpeak
Thal
0 2 4 6 8 10 12
MeanDecreaseGini

Boosting

- 약한(weak) 예측 모형을 결합하여 매우 정확한 예측 모형을 만드는 방법
- 예측모형을 순차적으로(sequentially) 학습하여 먼저 학습된 모형의 결과가 다음 모형의 학습에 정보를 제공
- 이전 모형의 결점을 보완하는 방향으로 학습이 이루어짐



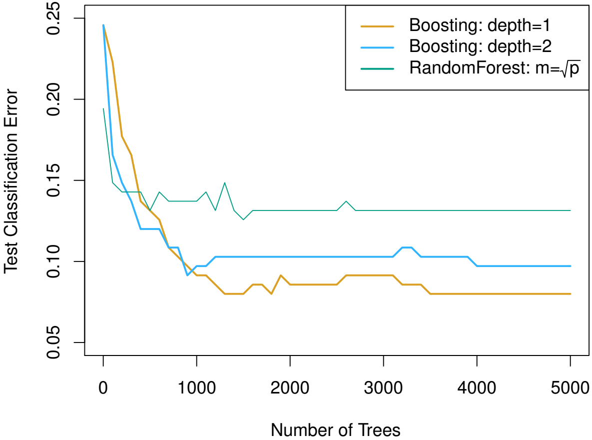
AdaBoost : Adaptive Boosting(1997)

1. 초기가중치설정 : $w_i = 1/n, i = 1, \ldots, n$
2. 다음을 $b = 1, \ldots, B$ 만큼 반복
   - 2.1 가중치 w_i를 이용하여 bootstrap sample 추출하여 분류 모형 $f_b(x) \in \{-1, 1\}$ 생성 ($y_i \in \{-1, 1\}$)
   - 2.2 오차률 계산 : $err_b = \dfrac{\sum_{i=1}^{n} w_i I(y_i = f_b(x_i))}{\sum_{i=1}^{n} w_i}$
   - 2.3 $\alpha_b = \log \dfrac{1 - err_b}{err_b}$
   - 2.4 가중치 갱신 : $w_i \leftarrow w_i \times \exp\left(\alpha_b I(y_i \neq f_b(x_i))\right)$
3. 최종모형 결정 : $f(x) = sign\left(\sum_{b=1}^{B} \alpha_b f_b(x)\right)$

Boosting

- 예측력이 매우 좋음
- Bagging과는 달리 과적합할 가능성이 있음 (B 선택 중요)
- 약한 예측모형을 기본 모형으로 사용하는 게 적합하기 때문에 decision stump (하나의 분할을 갖는 트리, 즉 terminal node의 갯수 $=$ 2)를 많이 사용함
- 물론 기본 예측 모형으로 분할이 몇 번 일어난 모형을 사용할지 선택 가능

Example



Boosting: depth=1
Boosting: depth=2
RandomForest: m=√p
Test Classification Error
0.05
0.10
0.15
0.20
0.25
0
1000
2000
3000
4000
5000
Number of Trees

Bagging vs. Boosting

비교	Bagging	Boosting
특징	병렬 앙상블 모델	연속 앙상블
목적	variance 감소	bias 감소
적합한 모형	High variance Low bias	Low variance High bias
대표 알고리즘	Random Forest	Adaboost Gradient Boosting
Sampling	Random sampling	Random sampling with weight on error