CART(Classification And Regression Trees)

: 트리 생성 알고리즘 중 하나
- 이진 트리로만 이루어져 있다.

터미널 노드에 적합하는 **상수 C**는
어떻게 도출할 수 있을까?

$$f(X) = \sum_{m=1}^{M} c_m I(X \in R_m)$$

회귀: y값들의 평균
분류: 최빈값

$$when \bigcup_{m=1}^{M} R_m = \mathbb{R}^p \; and \; R_m \cap R_p = \emptyset \; (non-overlapping, distinct)$$

$$and \; \mathbb{R}^p \; is \; a \; p-dimensional \; input \; space$$

2) Decision Tree Classifier

비슷한 관측치들끼리 몰려 있는 형태가 이상적

? **불순도가 줄어듦**

회귀 모델

RSS가 줄어드는 방향으로 트리가 나뉘도록 한다

: 관측치들이 어떤 범주에 속할 지에 대한
불확실성이 줄어드는 것

분류 모델

=

불순도가 줄어들도록 트리가 나뉘도록 하자

관측치들이 각자의 범주에 따라 적절히 분류 되었음을 의미

* 예시를 통해 이해해보자!

3) Tree-based Model에서 과적합 피하기

복잡도 (Complexity): 트리모델 > 선형회귀 모델

? Why

분류가 진행되며 관측치들이 나뉘는 과정

=

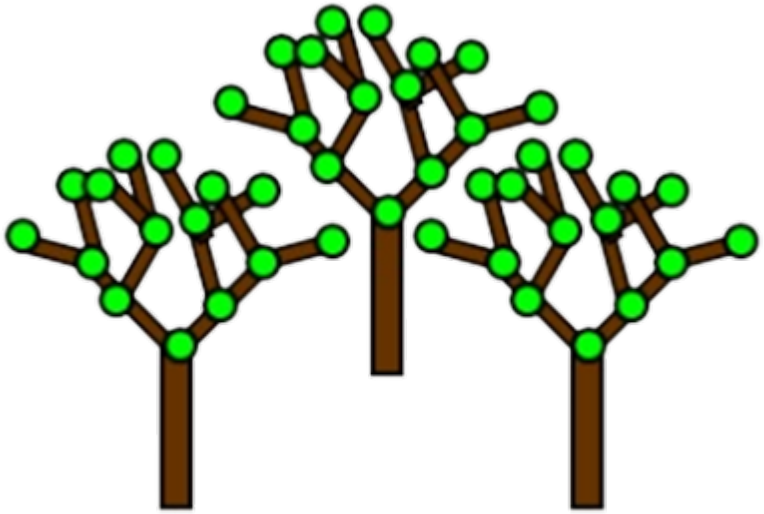
모델의 파라미터가 늘어나는 효과

⇒ 하이퍼 파라미터 값에 따라 모델 모양이 달라짐

3) Tree-based Model에서 과적합 피하기

모델 간 분산이 큰 경우를 해결해주자

> Idea
> **여러 개**의 트리모델을 적합해
> 이들을 **종합적**으로 **평가**해보자!



랜덤 포레스트 모델링 기법

자세한 설명은 뒤에서..

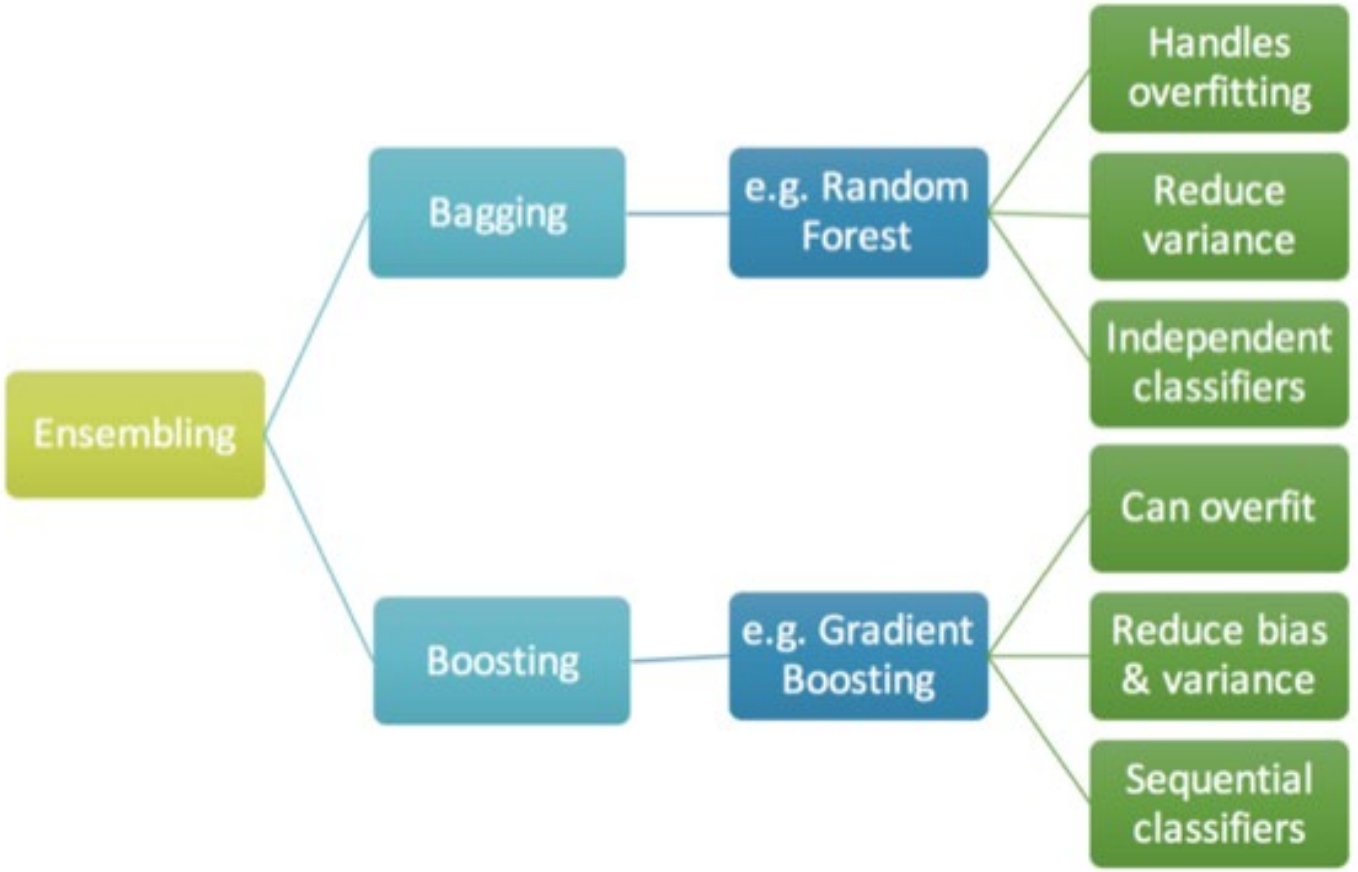
앙상블 모델(Ensemble Model)

지금까진 단일한 트리 기반 모델에 관한 내용이었다!

앙상블 기법은 여러 모델을 독립적으로 학습시킨 후,
각 모델의 결과를 조합하여 최종 결과를 생성

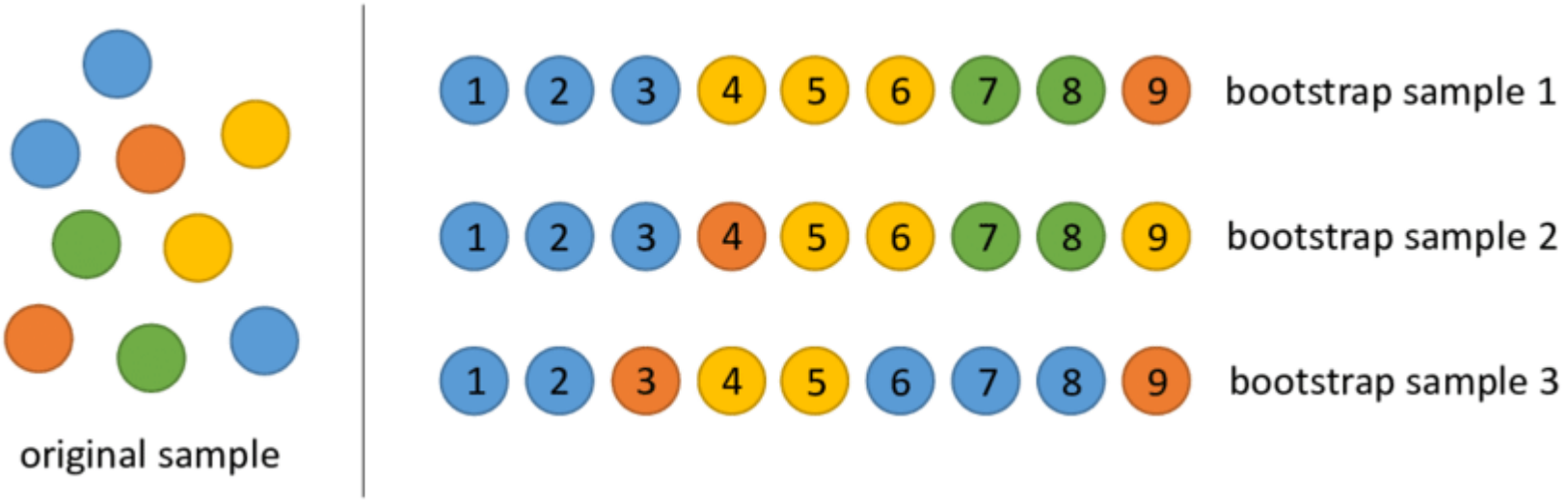
앙상블 모델(Ensemble Model)

앙상블 기법은 아래와 같은 갈래로 분류



배깅(Bagging)

Bootstrap sampling을 하는 이유?



샘플 데이터셋을 각각 다르게 해
모델에 적합 시켰을 때 발생할 수 있는 모델 variance를 최소화

랜덤 포레스트(Random Forest)

랜덤 포레스트는 매 모델링마다 사용할 독립변수를 임의로 선택

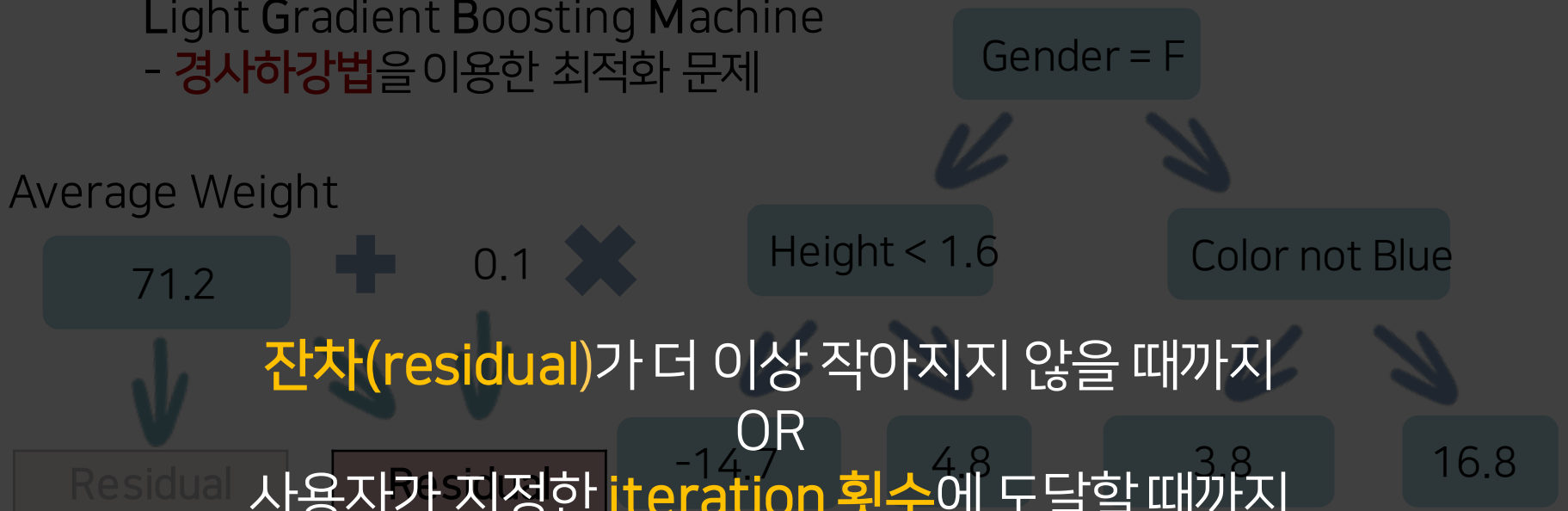
부트스트랩 샘플에 대한 배깅 실시

트리 적합 과정에서 일부 X변수만을 사용

변수의 종류가 다양해지면서 Decorrelation 달성

1) LGBM

Light Gradient Boosting Machine
- 경사하강법을 이용한 최적화 문제



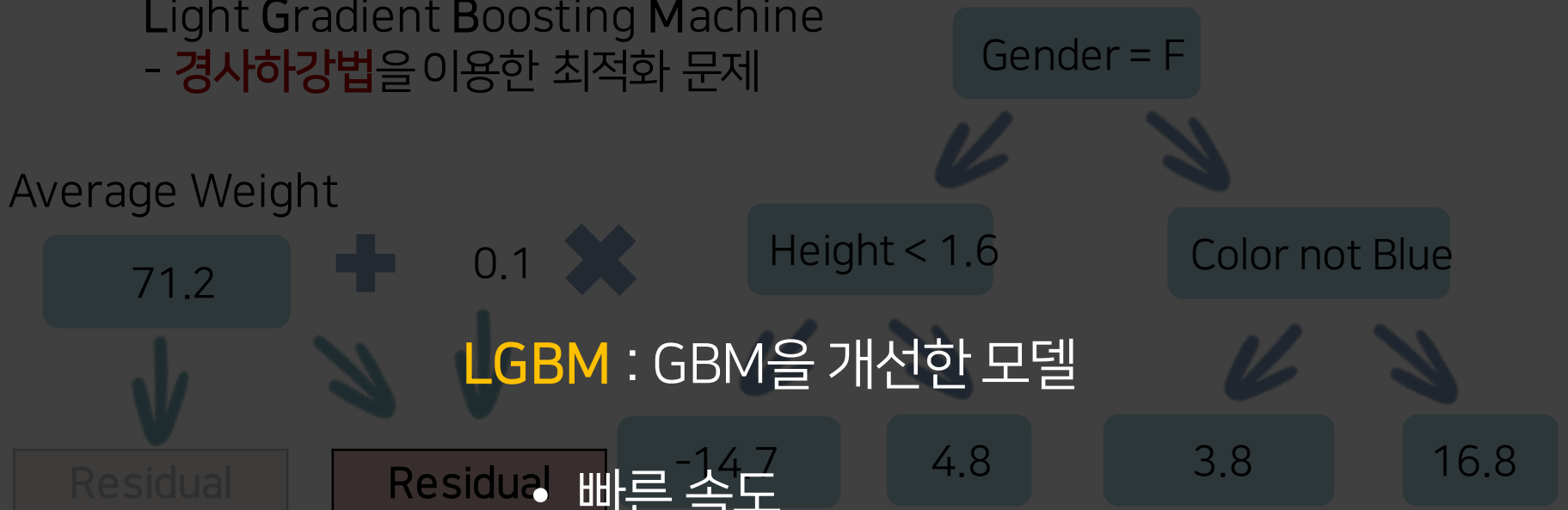
잔차(residual)가 더 이상 작아지지 않을 때까지
OR
사용자가 지정한 iteration 횟수에 도달할 때까지
반복하며 순차적으로 전개됨

Residual
16.8
4.8
-15.2
1.8
5.8
-14.2

Residual
15.1
4.3
-13.7
1.4
5.4
-12.7

1) LGBM

Light Gradient Boosting Machine
- 경사하강법을 이용한 최적화 문제



LGBM : GBM을 개선한 모델

Residual
16.8
4.8
-15.2
1.8
5.8
-14.2

Residual
15.1
4.3
-13.7
1.4
5.4
-12.7

- 빠른 속도
- 적은 메모리 차지
- GOSS 방법을 취함

1) LGBM

Light Gradient Boosting Machine

GOSS (Gradient Based One Side Sampling)
: 큰 error를 보이는 관측치들의 error를 줄이는데 집중

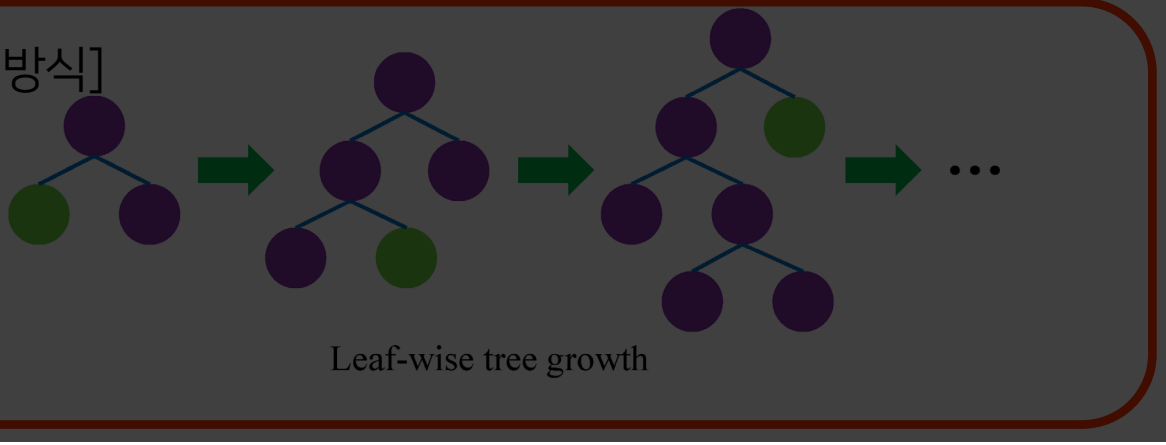
[너비 우선 방식]

LGBM

: 예측 오류 최소화가 목표

Level-wise tree growth

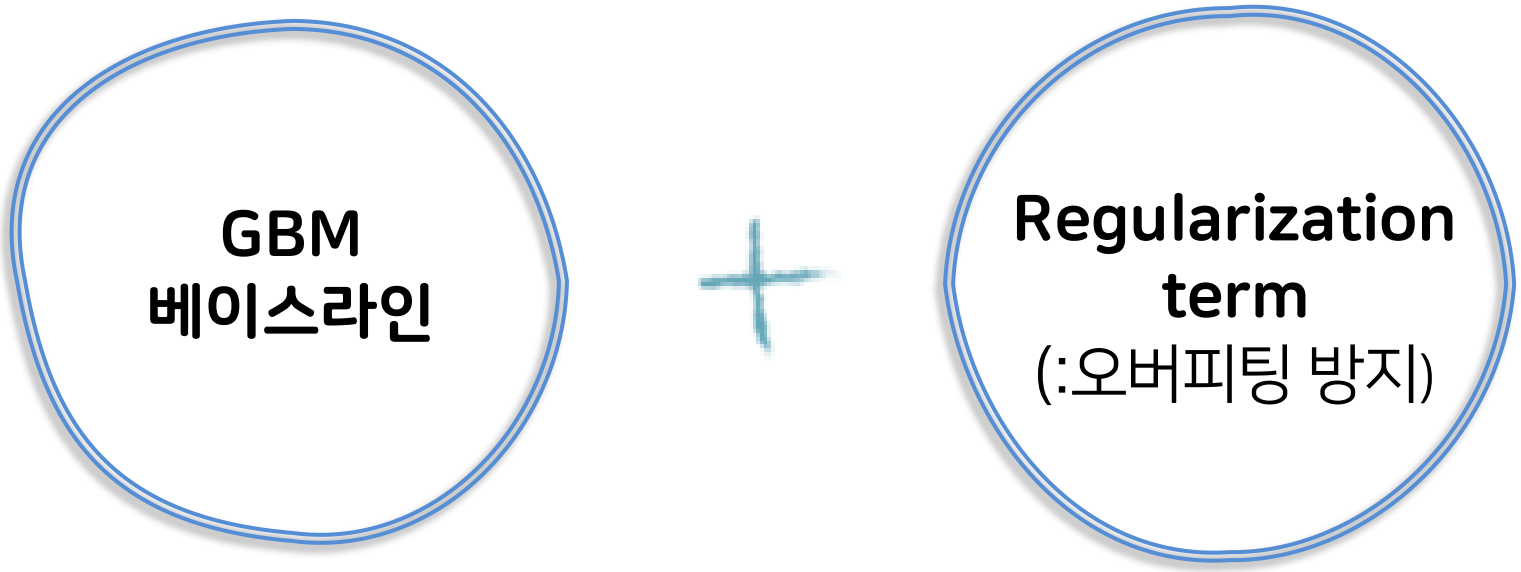
[깊이 우선 방식]



Leaf-wise tree growth

2) XGBoost

a.k.a. GBM Killer



회귀, 분류 문제 모두 수행 가능