

교 육 세 미 나

ToBig's 8기 최서현

Ensemble

앙상블

Contents

Unit 01 | 왜 앙상블을 ?

Unit 02 | Bagging

Unit 03 | Random Forest

Unit 04 | Boosting

Unit 05 | ADA Boost

Unit 06 | XG Boost

Unit 01 | 왜 앙상블을 ?

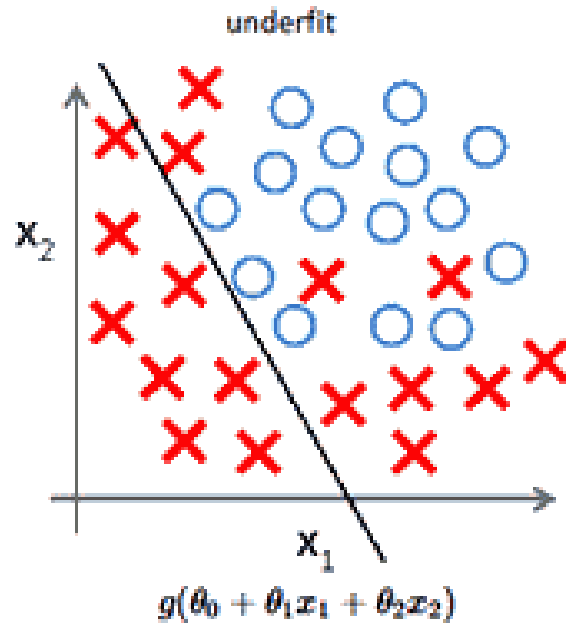
Ensemble : '전체적인 조화', '합주단'을 의미



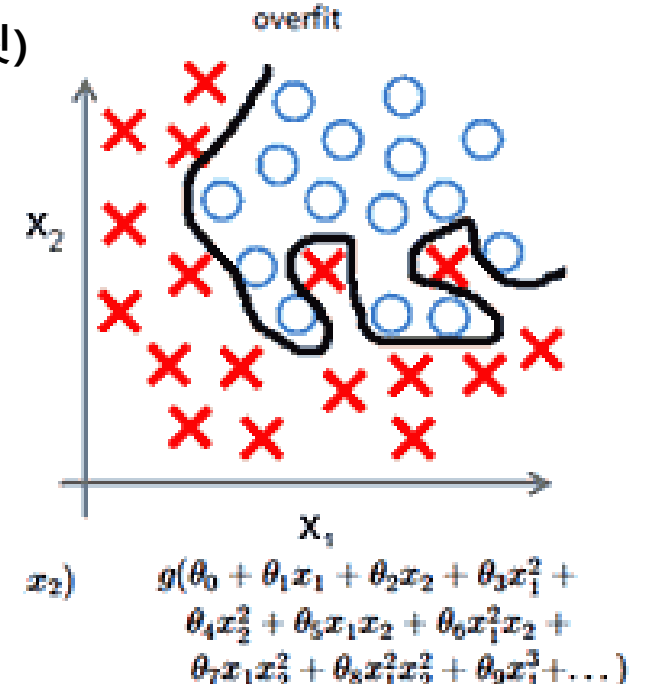
Unit 01 | 왜 앙상블을 ?

Bias – Variance Trade off Problem

Variance ↑
모델 복잡도 ↑
테스트셋에서는 정확도가 낮음 (오버핏)



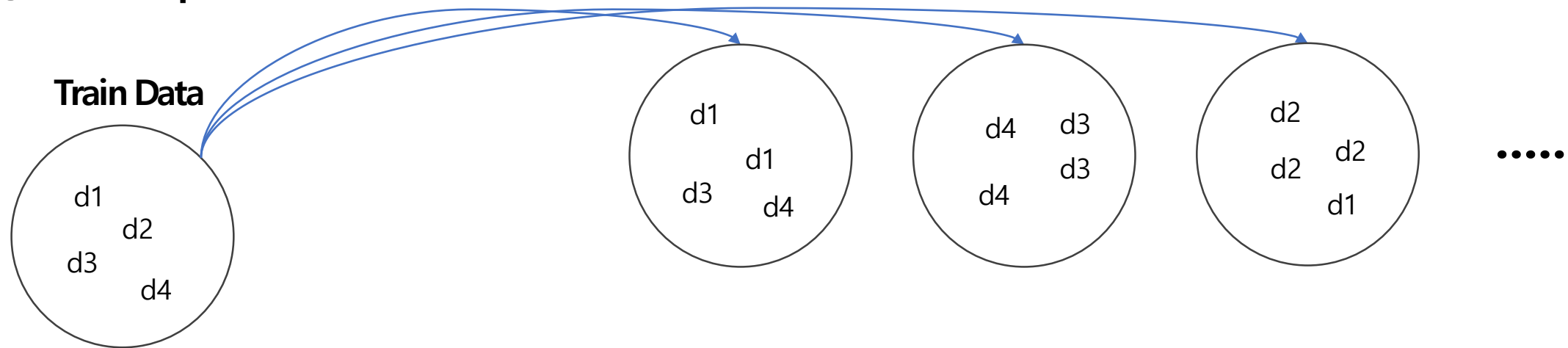
Bias ↑
모델 성능 자체가 그냥 낮음(언더핏)



Unit 02 | Bagging

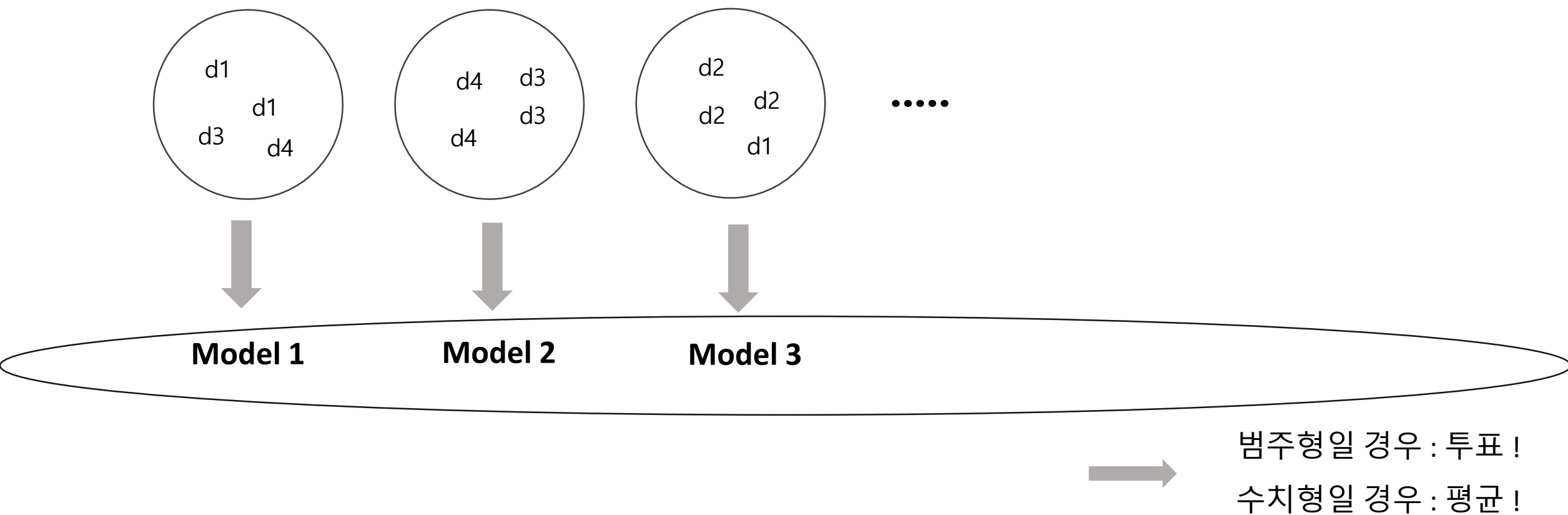
Bagging

① Bootstrap



Unit 02 | Bagging

② 모델을 만들고, Voting / Averaging



Unit 02 | Bagging

Bagging

Variance를 낮추기 좋다.

서로 다른 여러 개의 모델들의 결과들을 모두 고려하여 최종 결과를 내기 때문에
하나의 트레인셋에 너무 치중된 채 트레인이 되는 경우 (오버핏 혹은 과적합)를 피할 수 있는 것이다.
동시에 모델의 높은 정확도도 유지할 수 있다.



과적합 우려가 큰(= variance가 큰) 머신러닝 기법들에 적용하기 좋은 앙상블 방법이다.

ex) 깊이가 깊은 Decision Tree

Unit 03 | Random Forest

Random Forest

= 변수 랜덤 Bagging + Decision Tree

	온도	습도	풍속	비 여부
d1	15	60	15	1
d2	21	10	1	1
d3	3	70	5	0
d4	7	2	30	0

	온도	습도	비 여부
d1	15	60	1
d1	15	60	1
d3	3	70	0
d1	15	60	1

Tree 1

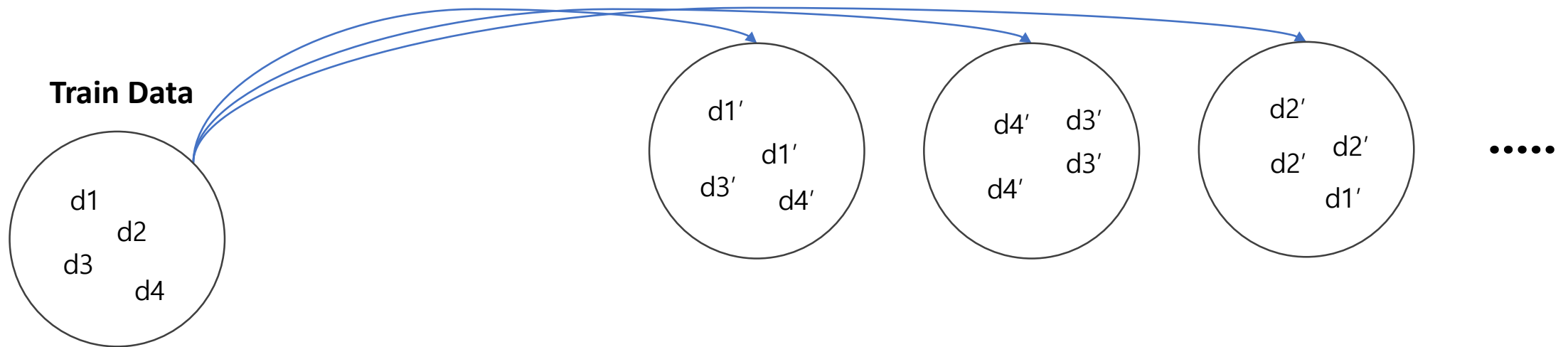
	습도	풍속	비 여부
d4	2	30	0
d2	10	1	1
d3	70	5	0
d3	70	5	0

Tree 2

Unit 03 | Random Forest

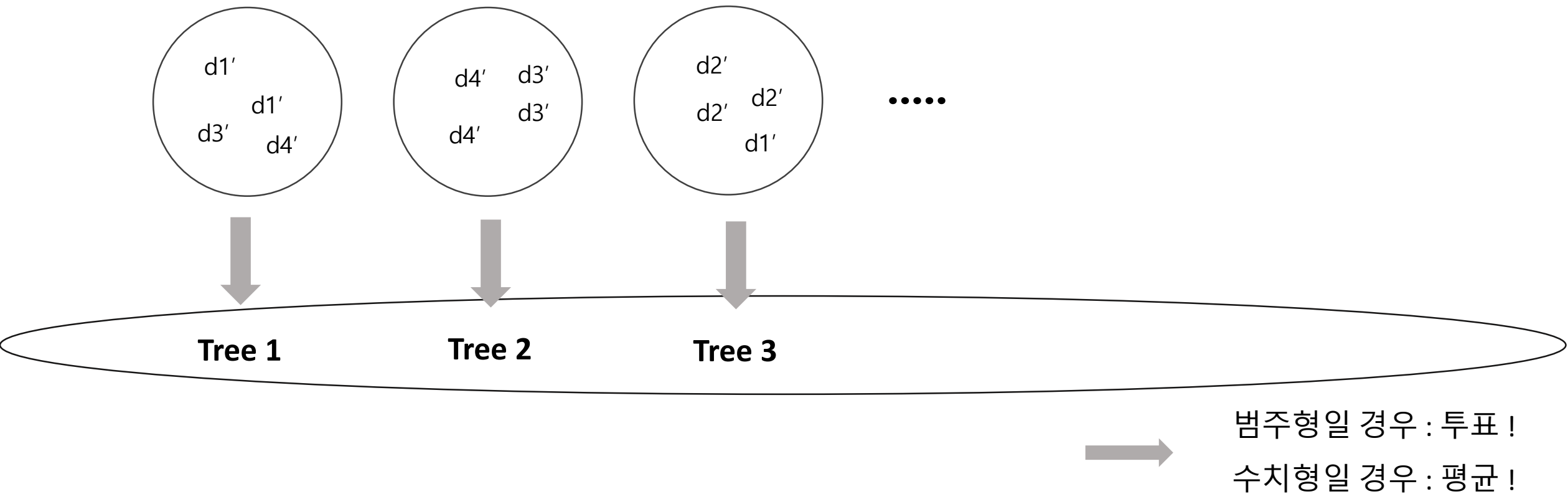
① 변수 랜덤 Bootstrap

※ d1' : d1에서 변수가 랜덤으로 일부만 추출된 상태



Unit 03 | Random Forest

② Voting / Averaging



Unit 04 | Boosting

Boosting

- ① 모델이 오분류한 관측치들에 가중치를 줌
- ② 위 과정을 거친 관측치들로 새로운 학습데이터를 만들어 이에 기반한 새로운 모델을 만듦
- ③ 새로운 모델이 오분류한 관측치들에 또 가중치를 줌
- ④ 위 과정을 반복하여 만들어진 여러개의 모델들로 voting or averaging ! (순차적 학습)

Unit 04 | Boosting

Boosting

Bias를 낮추기 좋다

모델의 정확도를 높이기 용이하므로 언더핏 문제 (Bias가 높은 상태, 혹은 모델의 예측 성능 자체가 낮은 상태) 를 해결할 수 있는 것이다.

동시에 서로 다른 여러 모델들을 모두 고려하므로 Variance도 낮게 유지할 수 있다.

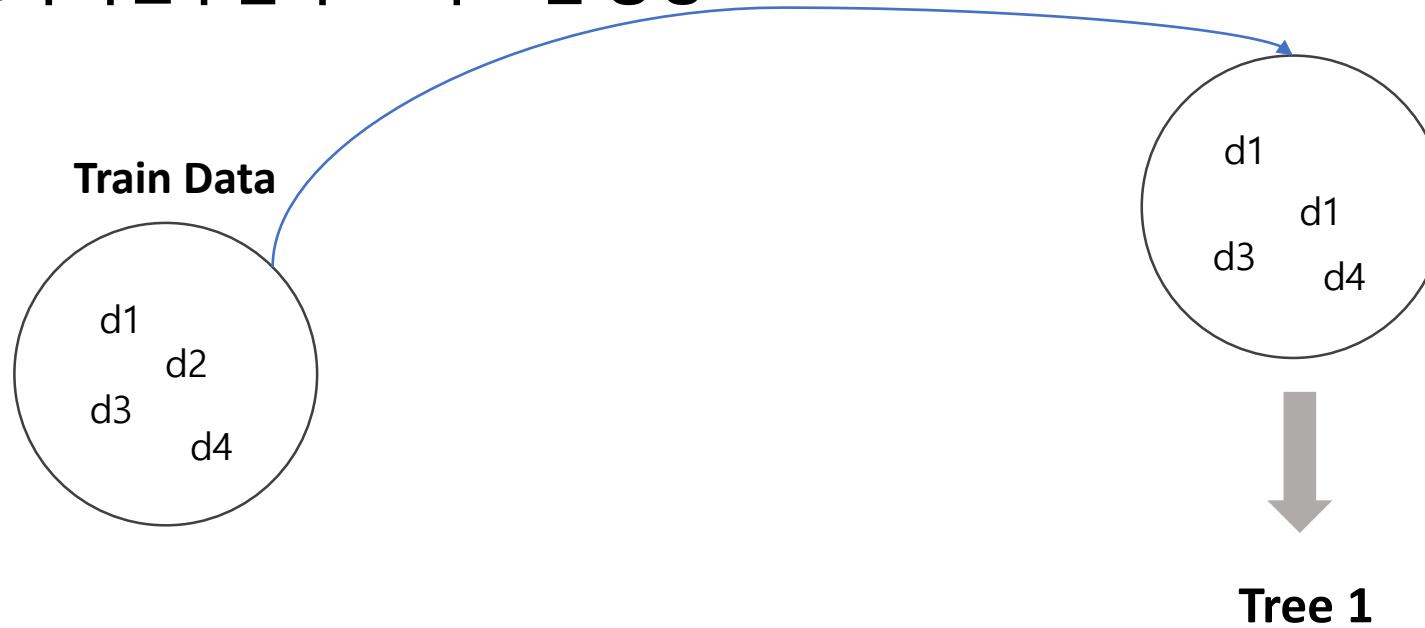


언더핏 위험이 큰(Bias가 높은) 모델들에 사용하기 적합하다
ex) 깊이가 얇은 Decision Tree, 선형분류기

Unit 05 | ADA Boost

ADA Boost

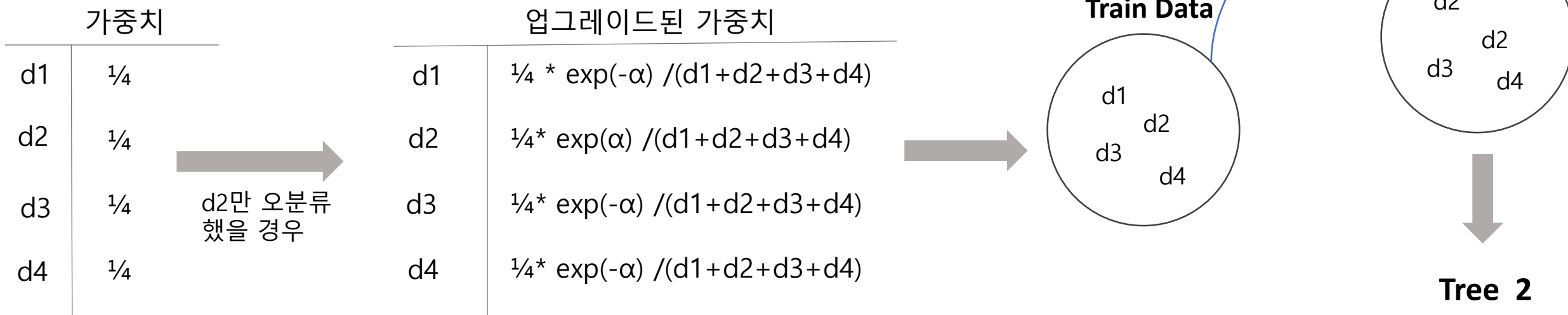
① 1회 복원추출 후 트리 모델 생성



Unit 05 | ADA Boost

ADA Boost

② Train Data로 테스트를 한 후에
오분류된 데이터들이 추출될 확률 높인 후 ① 번과정 다시 시행



※ α (신뢰도) : $\frac{1}{2} * \ln(1-e/e)$

※ e (에러율) : 오류데이터 가중치 합 / 전체 데이터 가중치 합

Unit 05 | ADA Boost

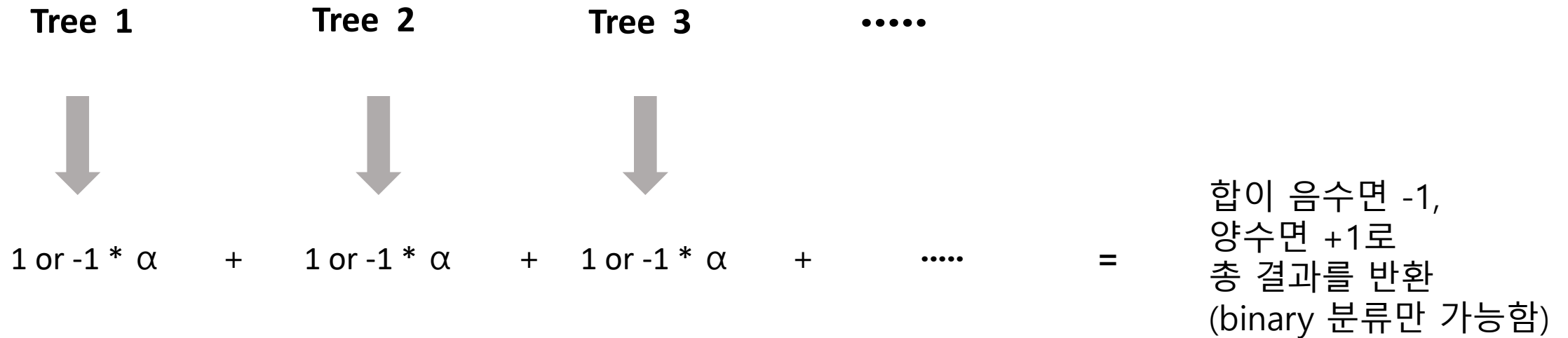
ADA Boost

- ③ 에러율이 0이 될 때까지 혹은 트리 모델 수가 일정한 수에 도달할 때까지 위 과정들을 계속 반복

Unit 05 | ADA Boost

ADA Boost

④ 신뢰도(α)를 곱하여 voting !



Unit 05 | ADA Boost

ADA Boost

현재 R과 Python에서 제공되는 ADA Boost 패키지는 다항 분류, 수치형 데이터 회귀까지 가능한 확장버전임 ! (ADA - SAMME, M1, R2 Algorithm . . .)

Unit 06 | XG Boost

XG Boost (Extreme Gradient Boosting)

- 목적 함수를 정의하고 이를 최소화 하는 값을 찾아 가중치를 업그레이드함
(≡ Gradient Descent)

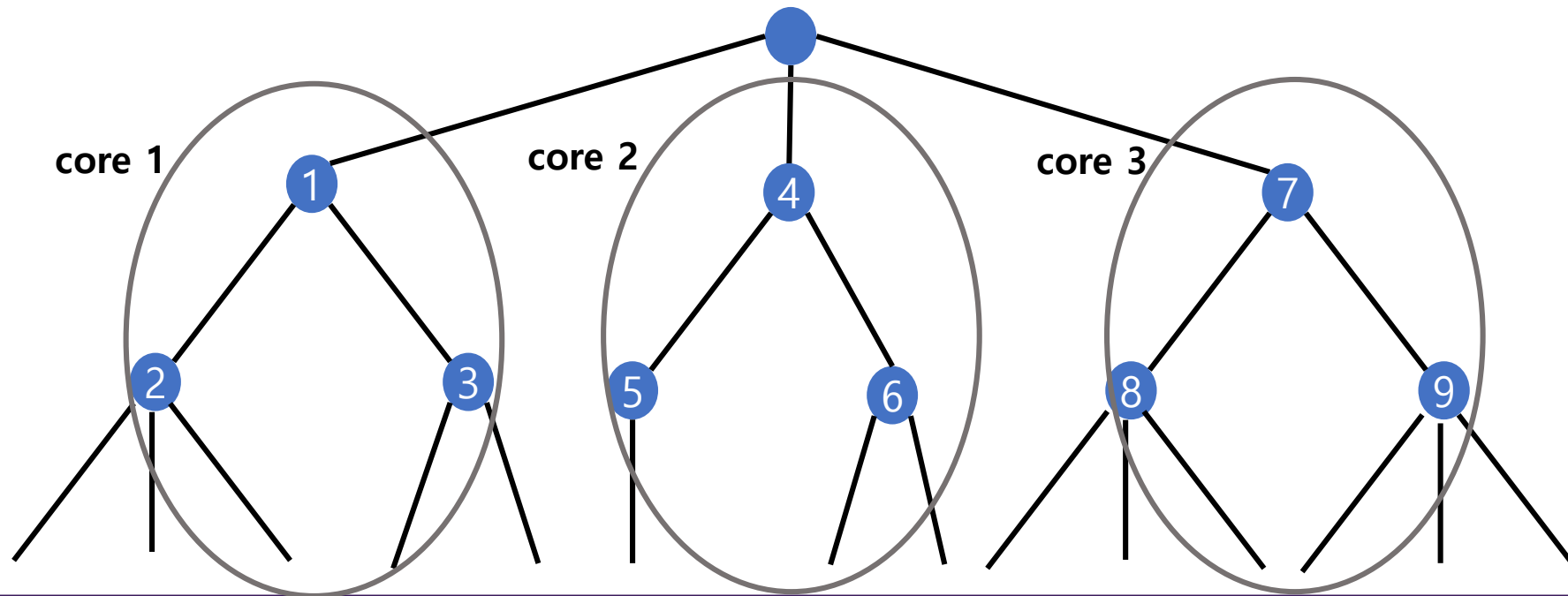
$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Training loss Complexity of the Trees

Unit 06 | XG Boost

XG Boost (Extreme Gradient Boosting)

- CPU 병렬처리. 코어들이 각자 할당 받은 변수들로 제각기 가지를 쳐 나감

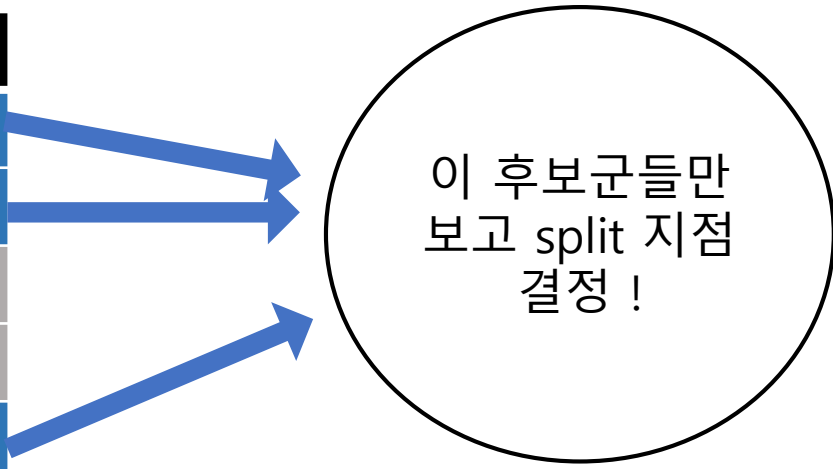


Unit 06 | XG Boost

XG Boost (Extreme Gradient Boosting)

- 연속형 변수에서 split 지점을 고려할 때 모든 값들을 살펴보고 결정하기보단 일부분만을 보고 결정을 함 !

ID	수학	영어
1	100	90
2	95	89
3	83	87
4	40	30
5	20	15



이 후보군들만
보고 split 지점
결정 !

Unit 06 | XG Boost

XG Boost (Extreme Gradient Boosting)

- Sparse Awareness가 가능. Zero데이터를 건너뛰면서 학습이 가능함 ! 그래서 인풋을 더미변수화 하면 속도가 상승!

ID	거주지역
1	서울
2	대전
3	대구
4	부산
5	제주도

<원데이터>



ID	서울	대전	대구	부산	제주도
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1

<더미 매트릭스>

Unit 06 | XG Boost

XG Boost (Extreme Gradient Boosting)

<https://www.r-bloggers.com/parallel-computation-with-r-and-xgboost/>

<http://matthewemery.ca/Why-Kagglers-Love-XGBoost/>

<https://www.youtube.com/watch?v=ufHo8vbk6g4&index=2&list=PLUHTiqMkyW3sLXEA08IXNcUuJe-bw1Qry>

<https://www.dataiku.com/learn/guide/code/python/advanced-xgboost-tuning.html>

<https://stats.stackexchange.com/questions/228260/does-it-make-a-difference-to-run-xgboost-on-hot-encoded-variables-or-single-fact>

<https://arxiv.org/pdf/1603.02754.pdf>

Q & A

들어주셔서 감사합니다.