ML SESSION

#5
배깅, 랜덤포레스트
베이지안 서치

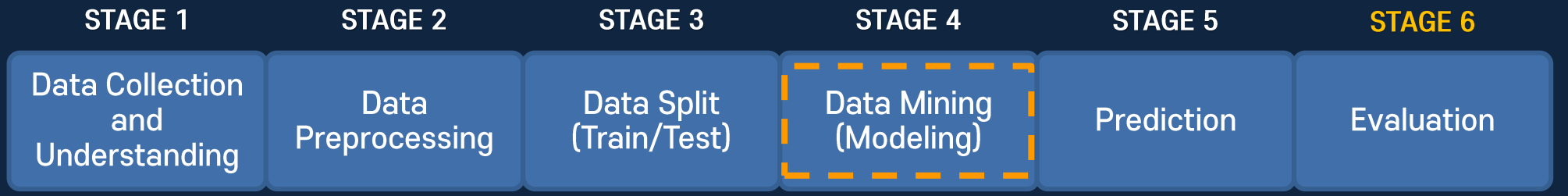
DATA&ANALYTICS

0 우수과제자

멘토	우수과제자
조영진	고민성
김예원	주민지
윤성식	이수빈
장성민	이현지
마민정	이수민
한보혜	이상우

INDEX

0 Modeling Step



1

앙상블

‘앙상블이 좋다’

1

앙상블

앙상블 : 서로 조화롭게 잘 어우러져 화합을 이룬다는 뜻

-> 각각의 악기가 내는 소리가 서로 잘 조화롭게 합쳐져 '더 좋은 소리'를 만듦

1

앙상블

ML에서 앙상블은 ?

앙상블(Ensemble) : 서로 조화롭게 잘 어우러져 화합을 이룬다는 뜻

-> 각각의 악기가 내는 소리가 서로 잘 조화롭게 합쳐져 '더 좋은 소리'를 만듦

각각의 모델 — 예측 — '더 좋은 예측 '

1

앙상블

앙상블(Ensemble) : 서로 조화롭게 잘 어우러져 화합을 이룬다는 뜻

ML에서는 ?

-> 각각의 모델이 내는 예측이 서로 잘 조화롭게 합쳐져 '더 좋은 예측'을 만듦

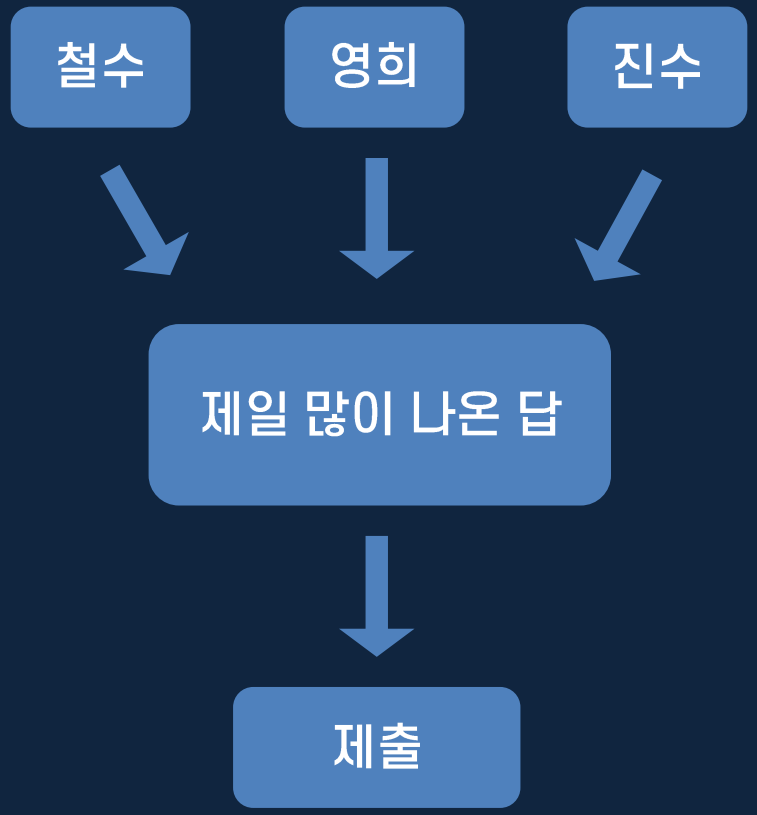
1

앙상블

문제	1	2	3	4	5
정답	A	A	A	A	A

1

앙상블



1

앙상블

정답률

철수

1	2	3	4	5
A	B	A	B	A

60%

영희

1	2	3	4	5
A	B	A	B	A

60%

진수

1	2	3	4	5
A	B	A	B	A

60%

제출

문제	1	2	3	4	5
개수	3A	3B	3A	3B	3A
제출	A	B	A	B	A

60%

1

앙상블

정답률

철수

1	2	3	4	5
A	B	A	B	A

60%

영희

1	2	3	4	5
A	A	B	B	A

60%

진수

1	2	3	4	5
B	A	A	A	B

60%

제출

문제	1	2	3	4	5
개수	2A 1B	2A 1B	2A 1B	1A 2B	2A 1B
제출	A	A	A	B	A

80%

1

앙상블

정답률

철수

1	2	3	4	5
A	B	A	B	A

60%

영희

1	2	3	4	5
A	A	B	A	A

80%

진수

1	2	3	4	5
B	A	A	A	B

60%

제출

문제	1	2	3	4	5
개수	2A 1B	2A 1B	2A 1B	2A 1B	2A 1B
제출	A	A	A	A	A

100%

1

앙상블

즉, 한 사람의 정답률은 좋지 않을 수 있다.

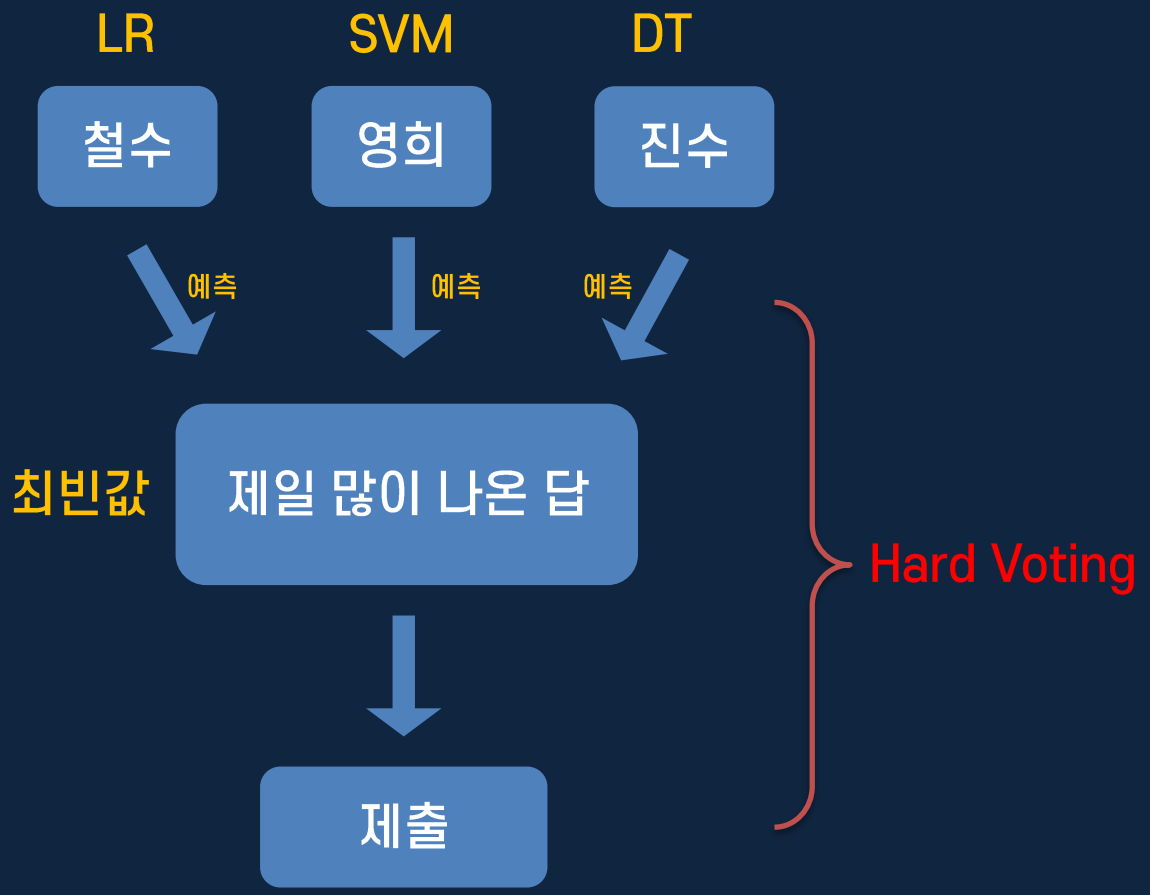
하지만 다 모아놓으면 한 사람이 하는 것보다 정답률이 더 좋아질 수 있다.

-> 집단지성

1

앙상블

ML에서는 ?



1

앙상블

√ Hard Voting은 앙상블의 한 종류

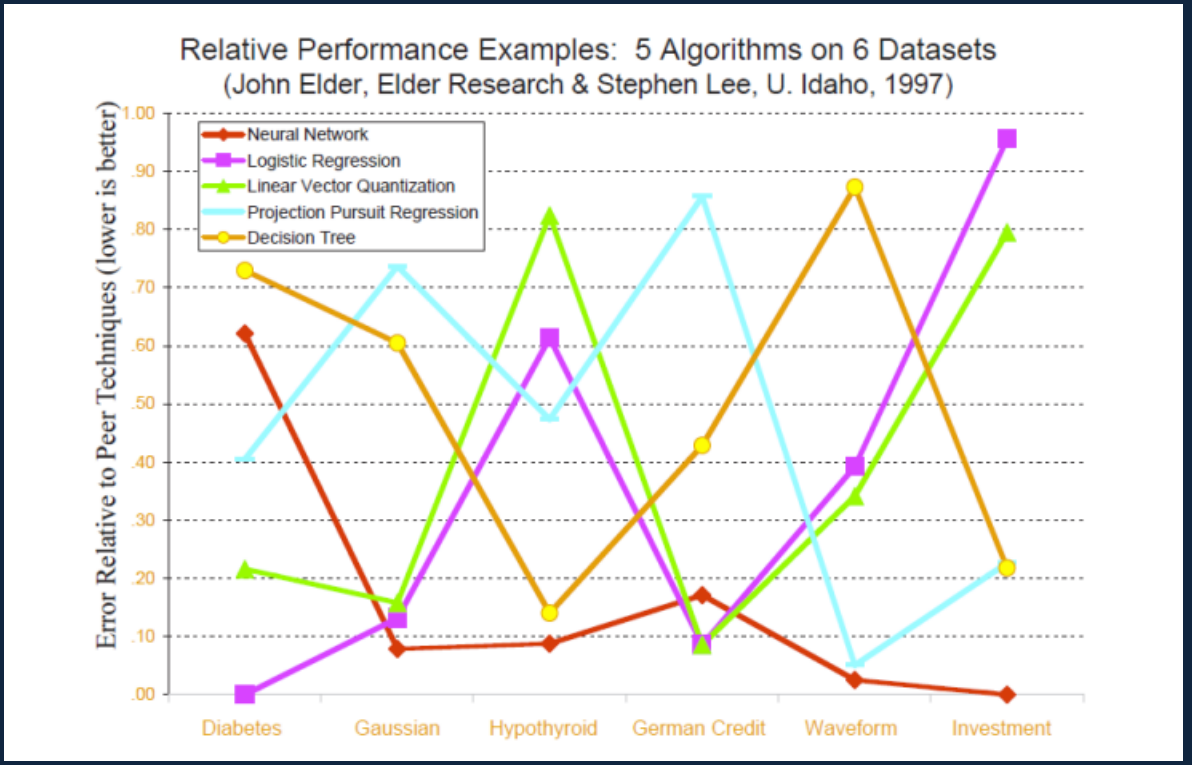
√ 앙상블은 알고리즘이 아니라 테크닉에 가깝다

1

앙상블

· 단일 모델의 한계

√ 모든 데이터셋에 대해 항상 최고인 알고리즘이 있을까 ?



1

앙상블

· 공짜 점심 이론

√ 모든 데이터셋에 항상 최고인 알고리즘은 존재할 수 없다.

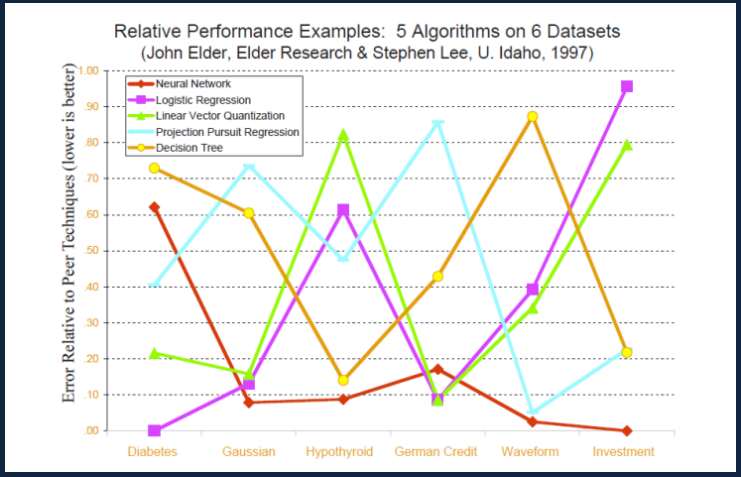
√ 좋은 일반화 성능을 기대한다면, 여러 알고리즘을 실험해서 가장 좋은 성능의 모델을 찾아야 한다.

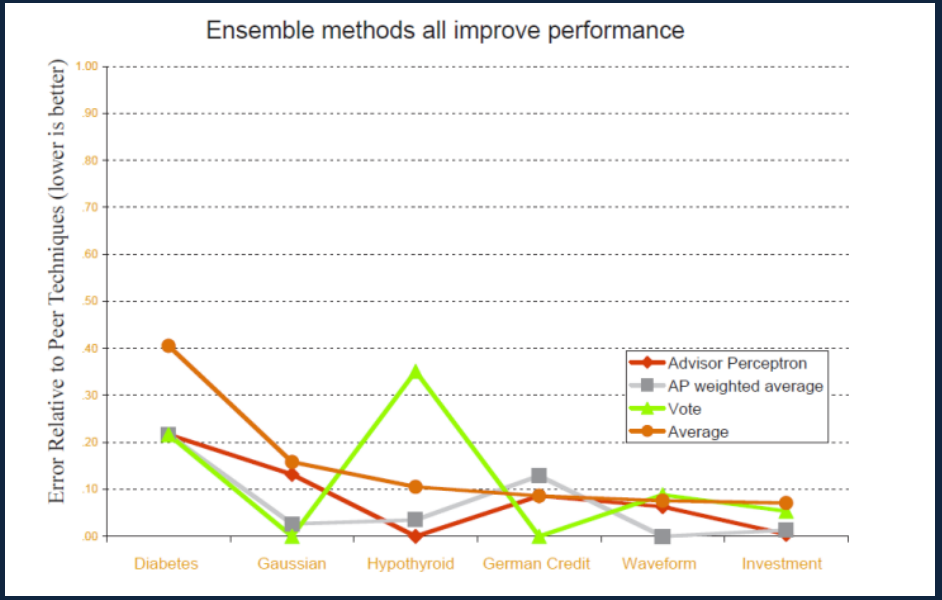
1

앙상블

· 여러 모델을 조합하면 ?

√ 단일 모델보다 성능이 확연히 좋아진다



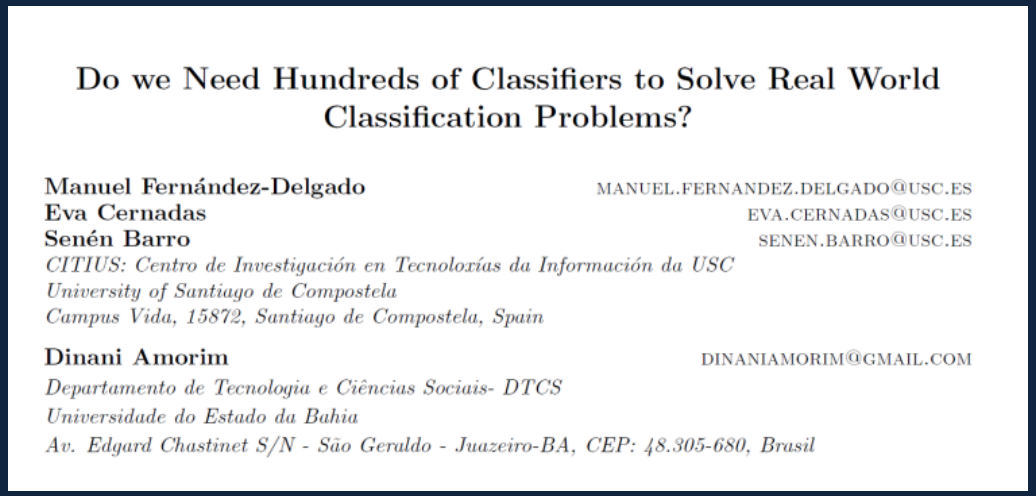


1

앙상블

경험적 연구 소개 (2014)

참고 : delgado14a.pdf (jmlr.org)



Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil

· 121개의 데이터셋에 대해

179개 알고리즘으로 성능 검증

1 앙상블

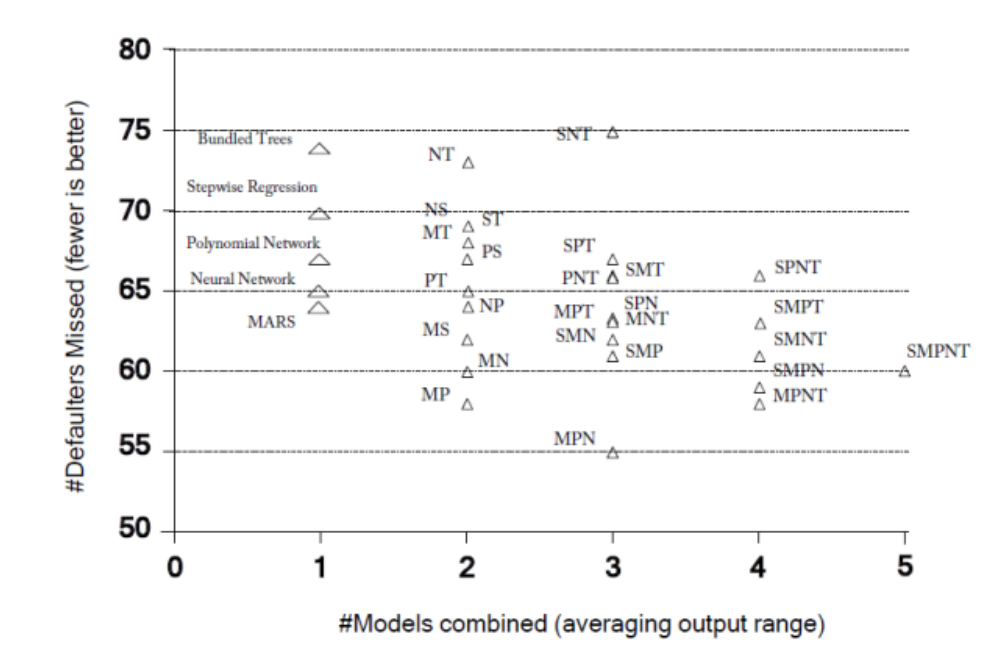
DO WE NEED HUNDREDS OF CLASSIFIERS TO SOLVE REAL WORLD CLASSIFICATION PROBLEMS?

Rank	Acc.	κ	Classifier	Rank	Acc.	κ	Classifier
32.9	82.0	63.5	parRF_t (RF)	67.3	77.7	55.6	pda_t (DA)
33.1	**82.3**	**63.6**	rf_t (RF)	67.6	78.7	55.2	elm_m (NNET)
36.8	81.8	62.2	svm_C (SVM)	67.6	77.8	54.2	SimpleLogistic_w (LMR)
38.0	81.2	60.1	svmPoly_t (SVM)	69.2	78.3	57.4	MAB_J48_w (BST)
39.4	81.9	62.5	rforest_R (RF)	69.8	78.8	56.7	BG_REPTree_w (BAG)
39.6	82.0	62.0	elm_kernel_m (NNET)	69.8	78.1	55.4	SMO_w (SVM)
40.3	81.4	61.1	svmRadialCost_t (SVM)	70.6	78.3	58.0	MLP_w (NNET)
42.5	81.0	60.0	svmRadial_t (SVM)	71.0	78.8	58.23	BG_RandomTree_w (BAG)
42.9	80.6	61.0	C5.0_t (BST)	71.0	77.1	55.1	mlm_R (GLM)
44.1	79.4	60.5	avNNet_t (NNET)	71.0	77.8	56.2	BG_J48_w (BAG)
45.5	79.5	61.0	nnet_t (NNET)	72.0	75.7	52.6	rbf_t (NNET)
47.0	78.7	59.4	pcaNNet_t (NNET)	72.1	77.1	54.8	fda_R (DA)
47.1	80.8	53.0	BG_LibSVM_w (BAG)	72.4	77.0	54.7	lda_R (DA)
47.3	80.3	62.0	mlp_t (NNET)	72.4	79.1	55.6	svmlight_C (NNET)
47.6	80.6	60.0	RotationForest_w (RF)	72.6	78.4	57.9	AdaBoostM1_J48_w (BST)
50.1	80.9	61.6	RRF_t (RF)	72.7	78.4	56.2	BG_IBk_w (BAG)
51.6	80.7	61.4	RRFglobal_t (RF)	72.9	77.1	54.6	ldaBag_R (BAG)
52.5	80.6	58.0	MAB_LibSVM_w (BST)	73.2	78.3	56.2	BG_LWL_w (BAG)
52.6	79.9	56.9	LibSVM_w (SVM)	73.7	77.9	56.0	MAB_REPTree_w (BST)
57.6	79.1	59.3	adaboost_R (BST)	74.0	77.4	52.6	RandomSubSpace_w (DT)
58.5	79.7	57.2	pnn_m (NNET)	74.4	76.9	54.2	lda2_t (DA)
58.9	78.5	54.7	cforest_t (RF)	74.6	74.1	51.8	svmBag_R (BAG)
59.9	79.7	42.6	dkp_C (NNET)	74.6	77.5	55.2	LibLINEAR_w (SVM)
60.4	80.1	55.8	gausspr Radial_R (OM)	75.9	77.2	55.6	rbfDDA_t (NNET)
60.5	80.0	57.4	RandomForest_w (RF)	76.5	76.9	53.8	sda_t (DA)
62.1	78.7	56.0	svmLinear_t (SVM)	76.6	78.1	56.5	END_w (OEN)
62.5	78.4	57.5	fda_t (DA)	76.6	77.3	54.8	LogitBoost_w (BST)
62.6	78.6	56.0	knn_t (NN)	76.6	78.2	57.3	MAB_RandomTree_w (BST)
62.8	78.5	58.1	mlp_C (NNET)	77.1	78.4	54.0	BG_RandomForest_w (BAG)
63.0	79.9	59.4	RandomCommittee_w (OEN)	78.5	76.5	53.7	Logistic_w (LMR)
63.4	78.7	58.4	Decorate_w (OEN)	78.7	76.6	50.5	ctreeBag_R (BAG)
63.6	76.9	56.0	mlpWeightDecay_t (NNET)	79.0	76.8	53.5	BG_Logistic_w (BAG)

- 공짜 점심 이론이 틀리지는 않았다.
- 순위가 통계적으로 유의미한 차이가 있음

1

앙상블


80
75
70
65
60
55
50
0
1
2
3
4
5
#Defaulters Missed (fewer is better)
#Models combined (averaging output range)
Bundled Trees
Stepwise Regression
Polynomial Network
Neural Network
MARS
NT
NS
MT
ST
PS
PT
NP
MS
MN
MP
SNT
SPT
PNT
SMT
MPT
SPN
MNT
SMN
SMP
MPN
SPNT
SMPT
SMNT
SMPN
MPNT
SMPNT

1

앙상블

· 정리

√ 앙상블은 '거의' 모든 경우에서 단일 모델보다 좋은 성능을 보일 수 있다.

√ 모든 상황에서 최고의 성능을 내는 단일 모델은 없다.

√ 분류기가 독립적일수록, 오차에 상관관계가 없을수록 더 효과적

√ 개별 분류기의 성능이 어느정도 보장되어야 함

1

앙상블

· 앙상블을 어떻게 할까 ?

√ 단일 모델들을 조합해서 사용

√ 편향과 분산을 고려

$$Err(x_0) = E\left[y - \hat{F}(x) | x = x_0\right]$$

. . . . . .

$$= Bias^2\left(\hat{F}(x_0)\right) + Var\left(\hat{F}(x_0)\right) + \sigma^2$$

Bias-Variance Decomposition : https://ratsgo.github.io/machine%20learning/2017/05/19/biasvar/

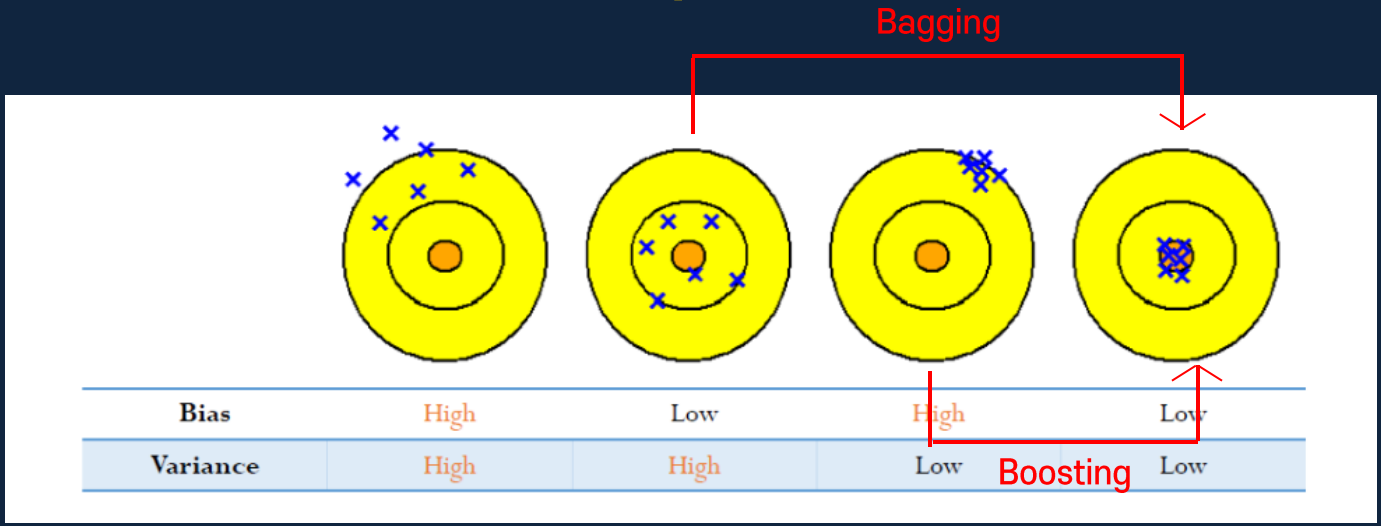
1

앙상블

· 편향과 분산

√ 편향 : 실제 레이블(정답값)과 여러 모델들이 예측한 값의 평균의 차이

√ 분산 : 개별 모델들의 확산 정도



· Low Bias, High Variance Models : DT, SVM, ANN

· High Bias, Low Variance Models : LR

2

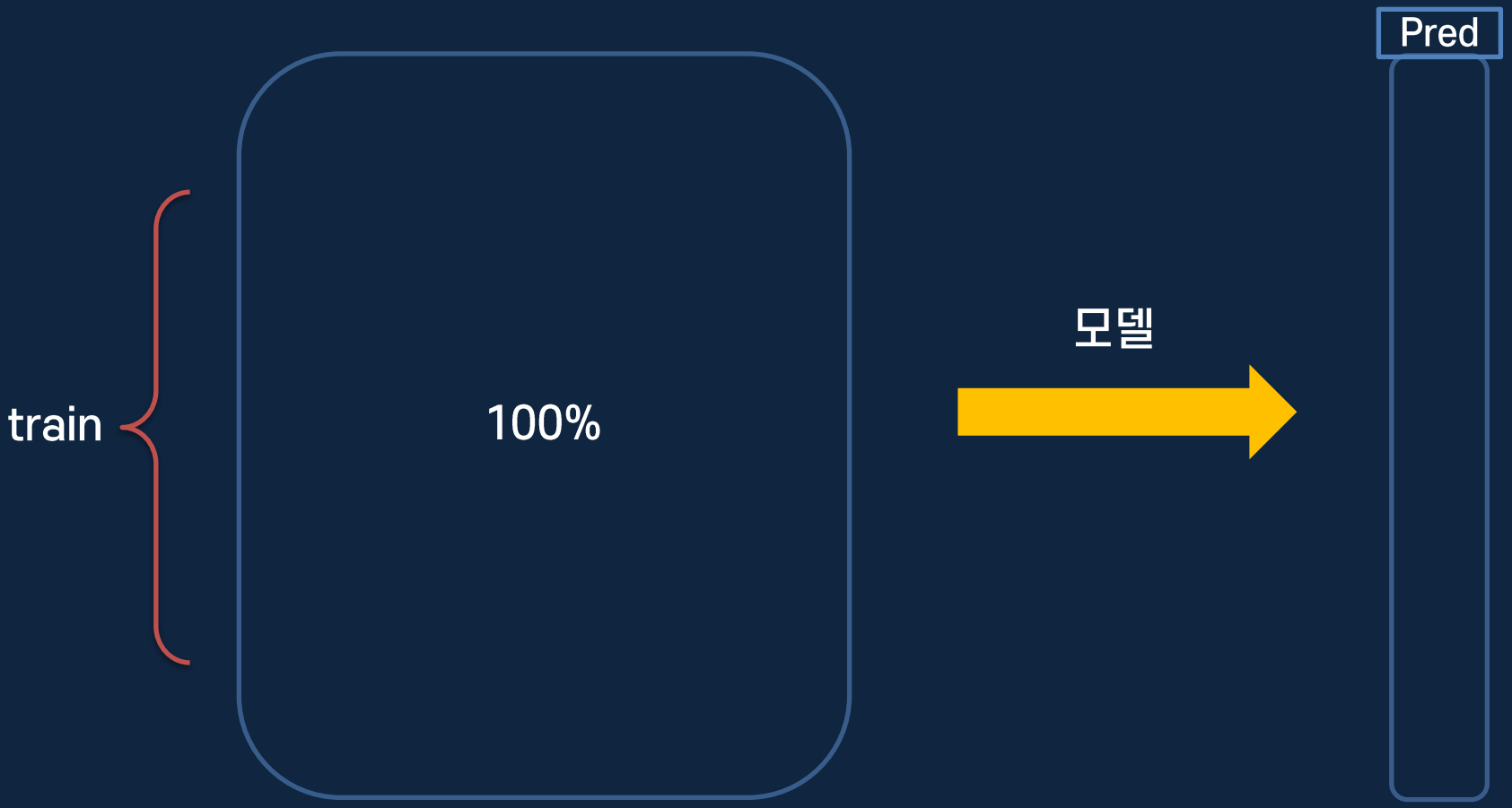
배깅과 페이스팅

- 배깅(Bagging) : Bootstrap Aggregating

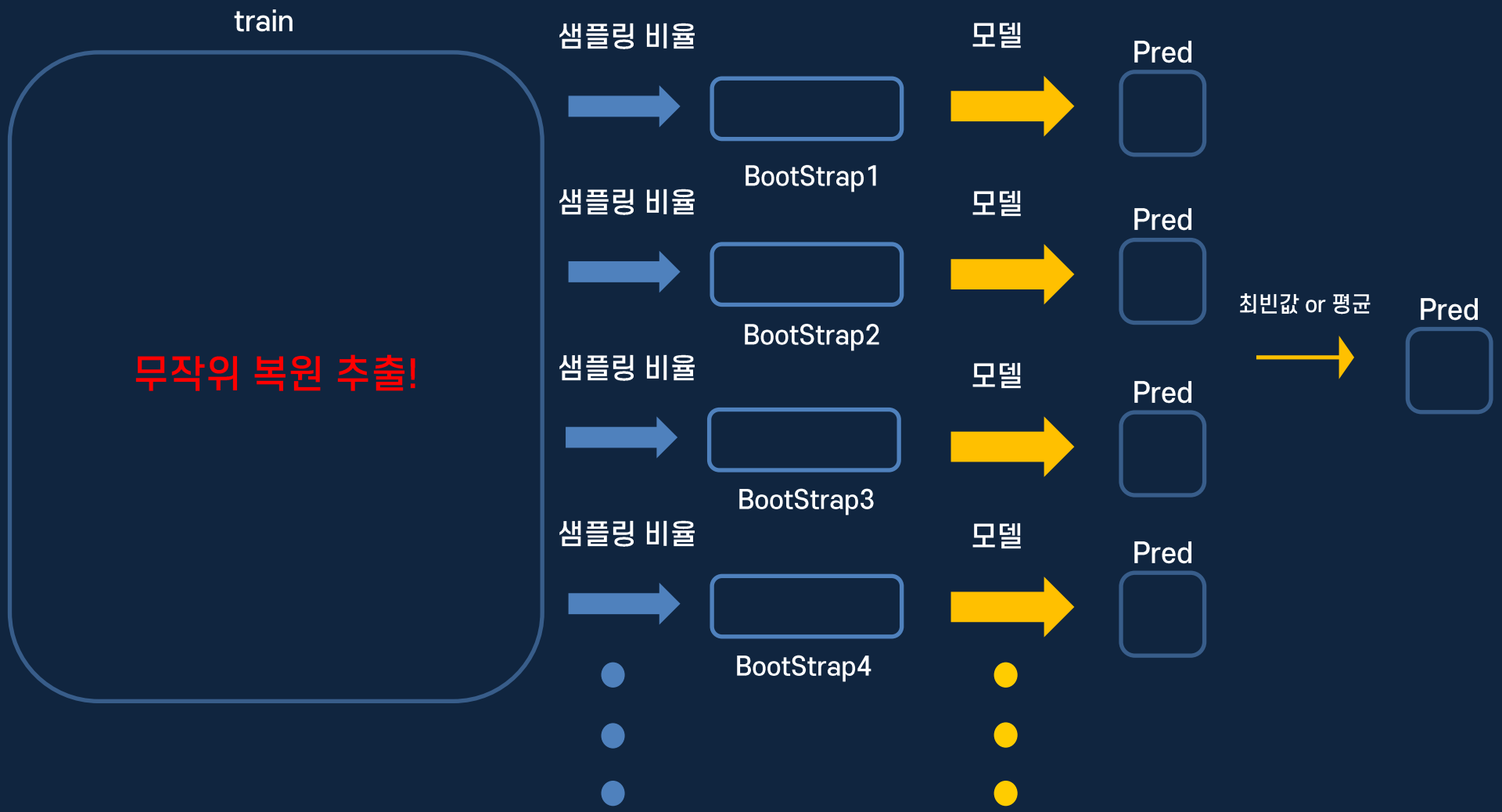
  -> 다양한 Bootstrap을 만들어 단일 모델을 돌리고 결과를 취합하는 방식

2

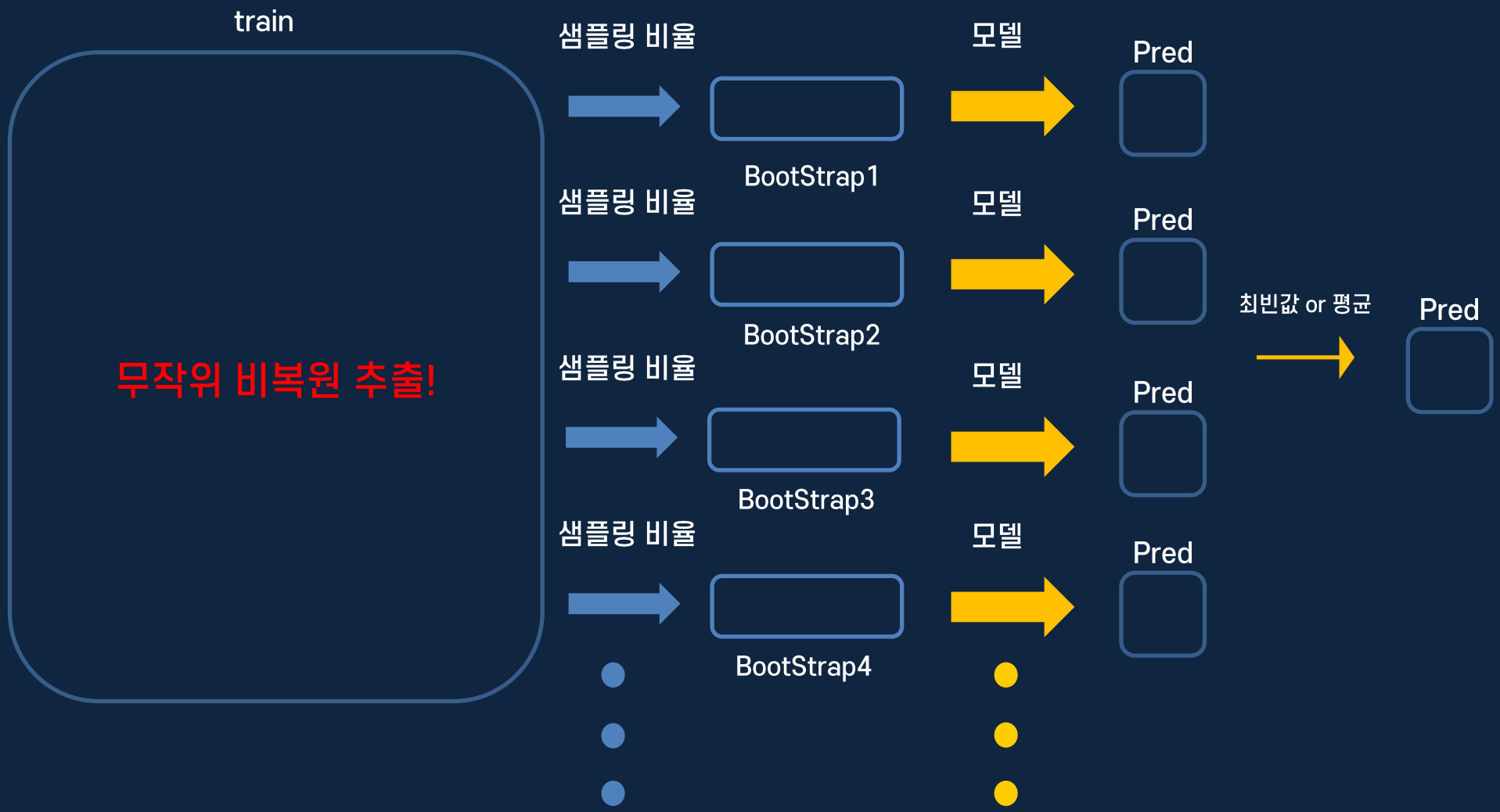
배깅과 페이스팅



- 배깅


train
무작위 복원 추출!
샘플링 비율
BootStrap1
모델
Pred
샘플링 비율
BootStrap2
모델
Pred
샘플링 비율
BootStrap3
모델
Pred
샘플링 비율
BootStrap4
모델
Pred
최빈값 or 평균
Pred

- 페이스팅


train
무작위 비복원 추출!
샘플링 비율
BootStrap1
모델
Pred
샘플링 비율
BootStrap2
모델
Pred
샘플링 비율
BootStrap3
모델
Pred
샘플링 비율
BootStrap4
모델
Pred
최빈값 or 평균
Pred

2

배깅과 페이스팅

배깅과 페이스팅

· 특징

√ 개별 예측기의 편향은 train 100% 다 써서 훈련 했을 때보다는 높음

√ 경험적으로, 앙상블의 결과는 원본 데이터셋으로 하나의 예측기를 훈련시킬 때와 비교해 편향은 비슷하나, 분산은 줄어듬.

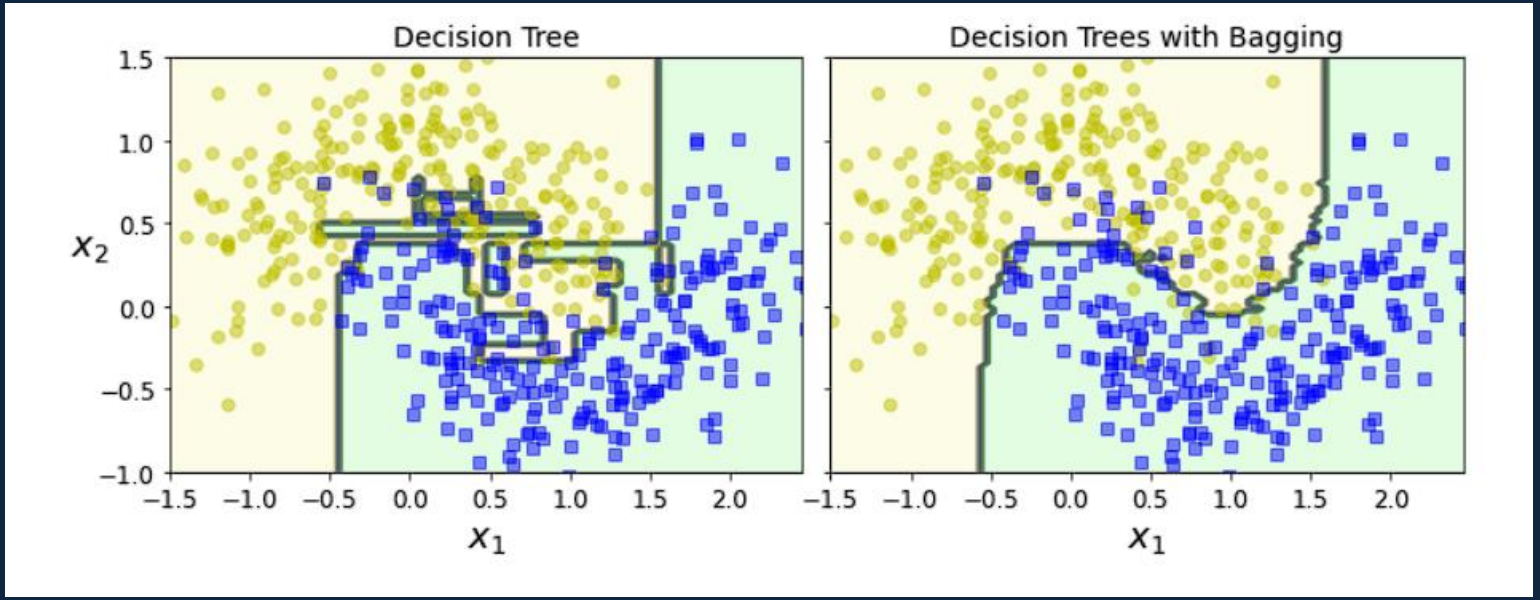
-> 모델의 편향은 낮으나 분산이 큰 모델들에 적합 !

(DT, SVM, ANN..)

2

배깅과 페이스팅

배깅과 페이스팅



- 결정 트리 하나의 예측보다 훨씬 일반화 잘 되어 있음
- 배깅은 일반적으로 비슷한 편향에서 더 작은 분산을 만듦

2

배깅과 페이스팅

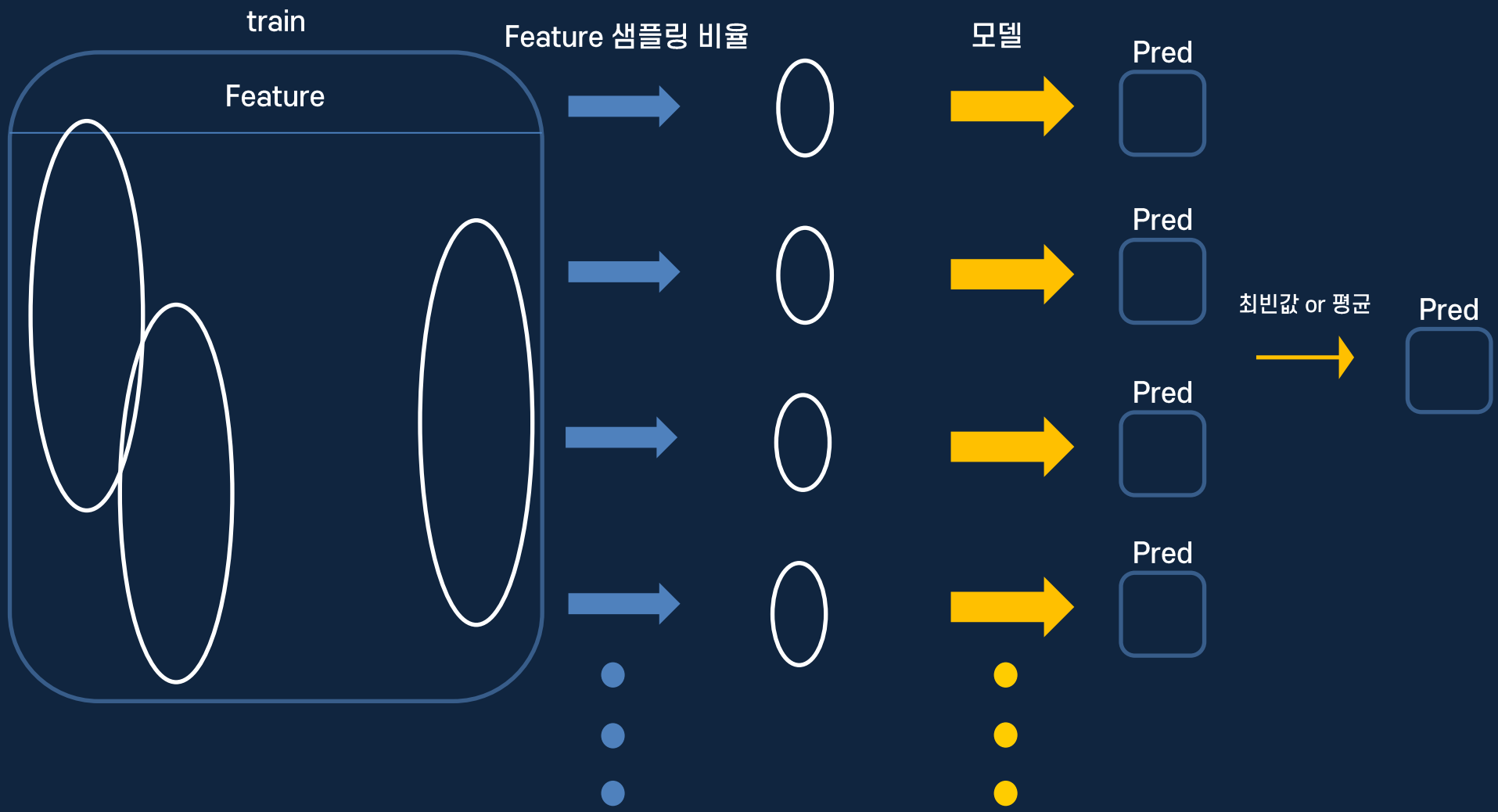
랜덤 패치와 서브 스페이스

- 반대로, 특성(Feature) 샘플링도 가능하다 !

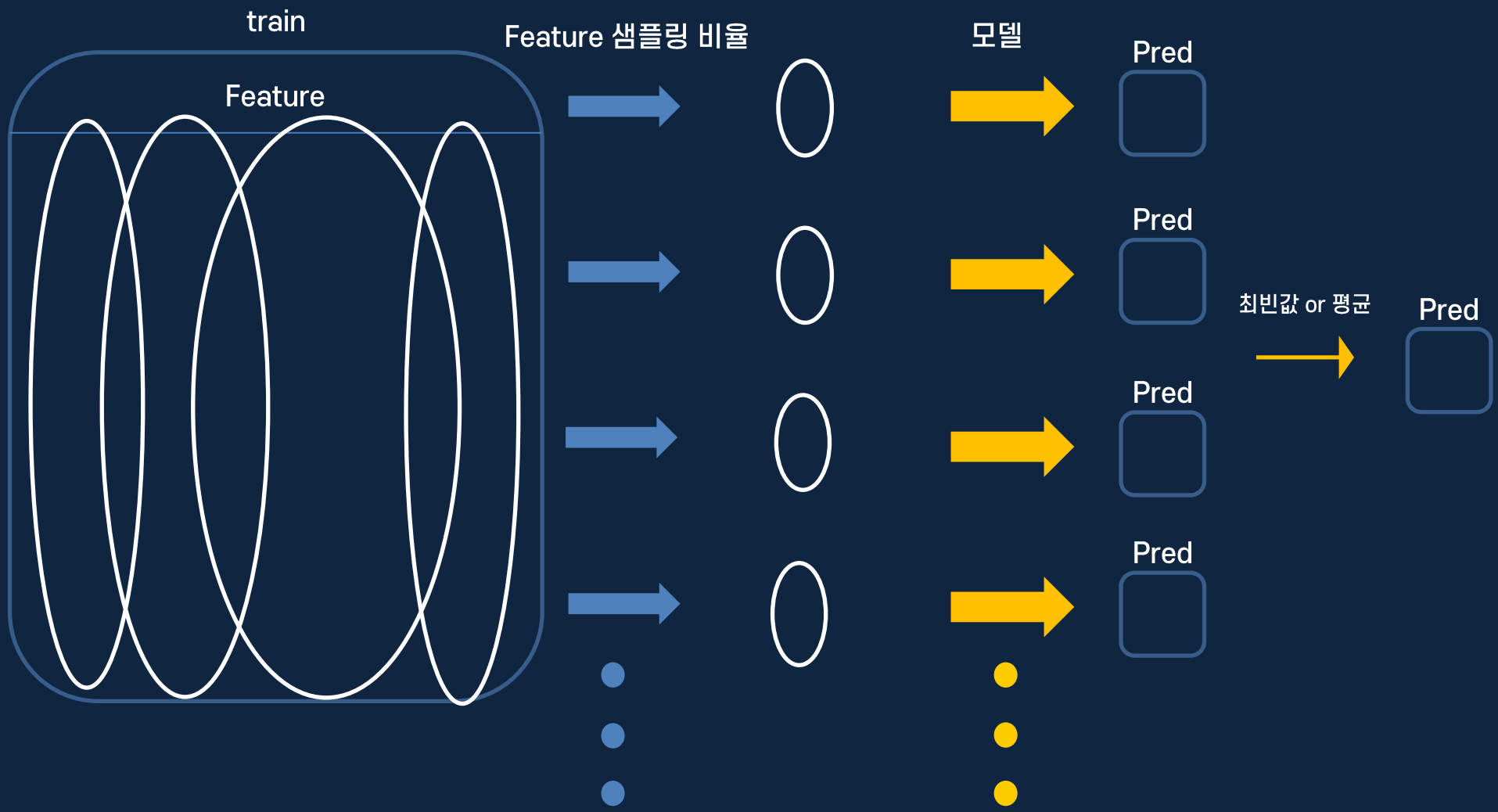
- 서브스페이스 방식 : 훈련 샘플은 모두 사용하고 특성만 샘플링 하는 방식
- 랜덤패치방식 : 훈련 샘플과 특성을 모두 샘플링 하는 방식

- 고차원 데이터셋에서 유용

- 랜덤 패치 방식


train
Feature
Feature 샘플링 비율
모델
Pred
Pred
Pred
Pred
최빈값 or 평균
Pred

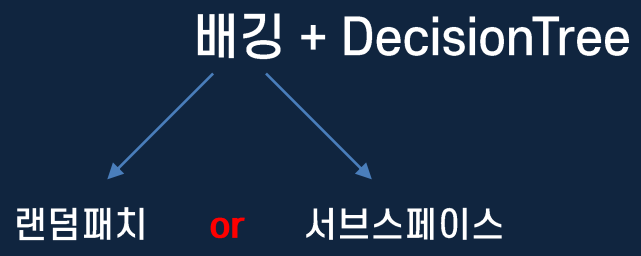
- 서브스페이스방식


train
Feature
Feature 샘플링 비율
모델
Pred
Pred
Pred
Pred
최빈값 or 평균
Pred

3

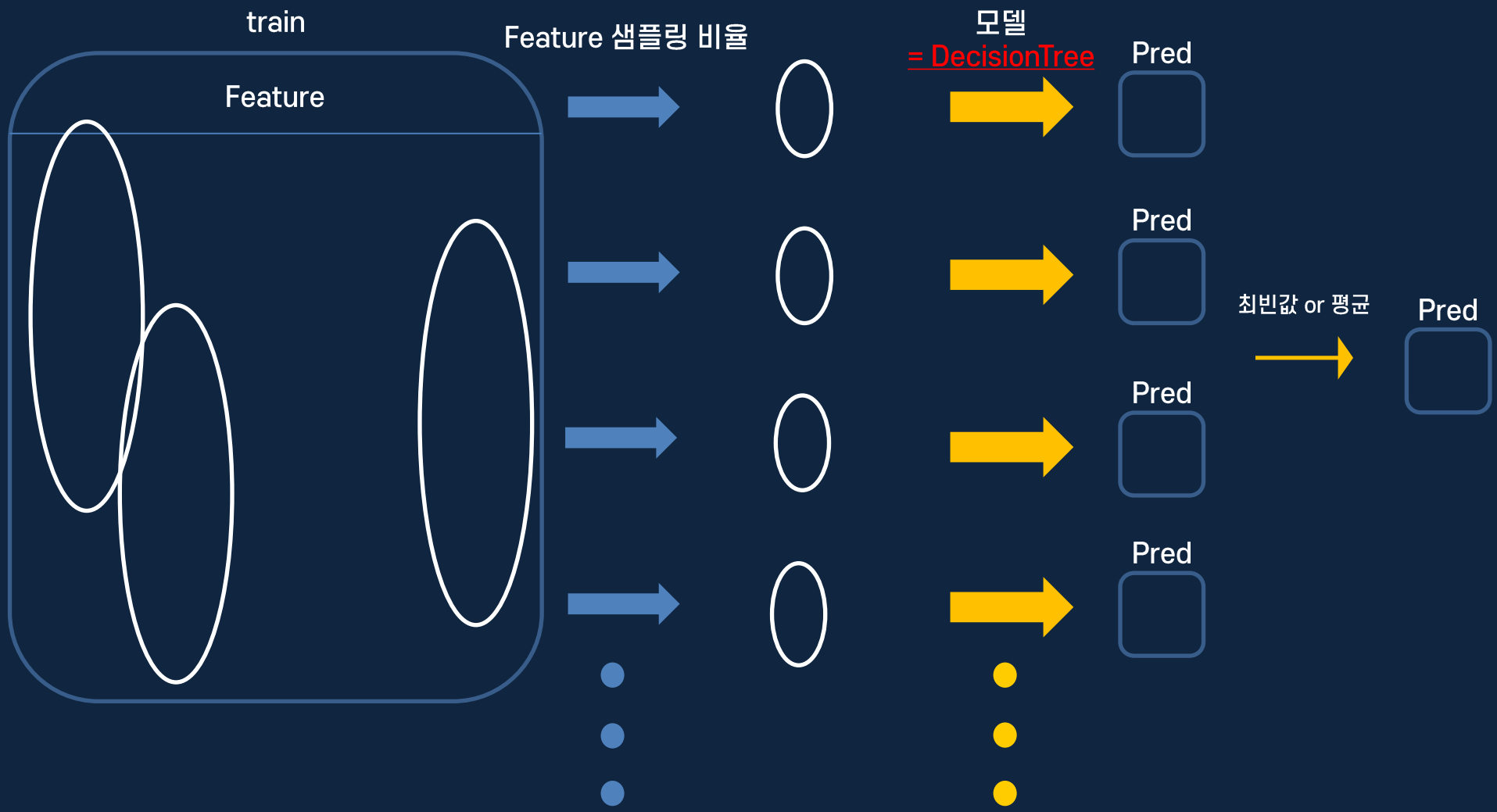
랜덤포레스트

랜덤 포레스트



3

랜덤 포레스트


train
Feature
Feature 샘플링 비율
모델
= DecisionTree
Pred
Pred
Pred
Pred
최빈값 or 평균
Pred

3

랜덤 포레스트

랜덤포레스트

- 수많은 의사 결정 트리로 학습된 결과를 앙상블 하여 새로운 예측을 내는 모델
- DT보다 훈련 샘플링의 무작위성 + 특성 샘플링의 무작위성 추가
- 과적합 방지 + 앙상블 효과

4

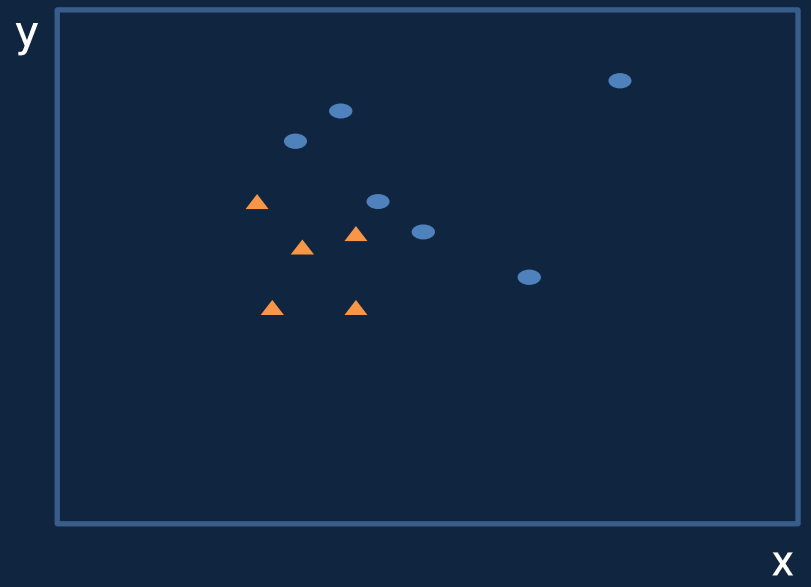
엑스트라트리

엑스트라트리

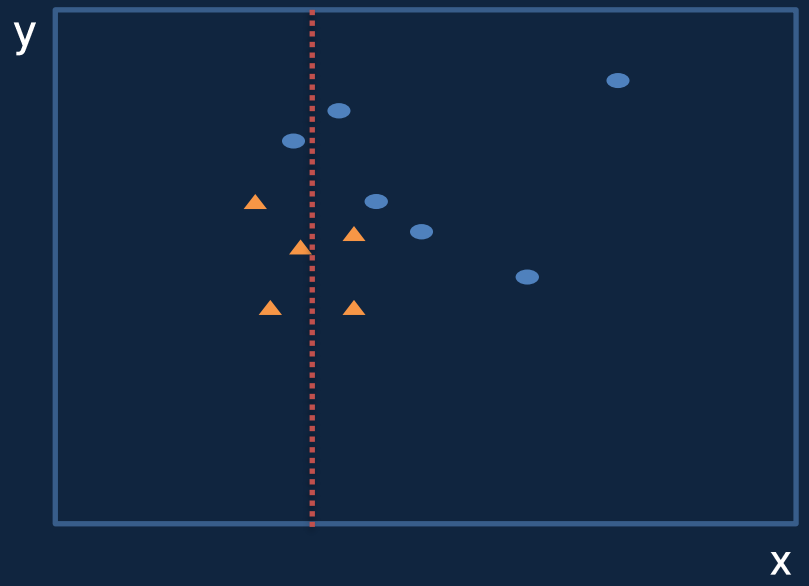
- 더 극단적으로 Random한 트리 모델

4

엑스트라트리



4 엑스트라트리

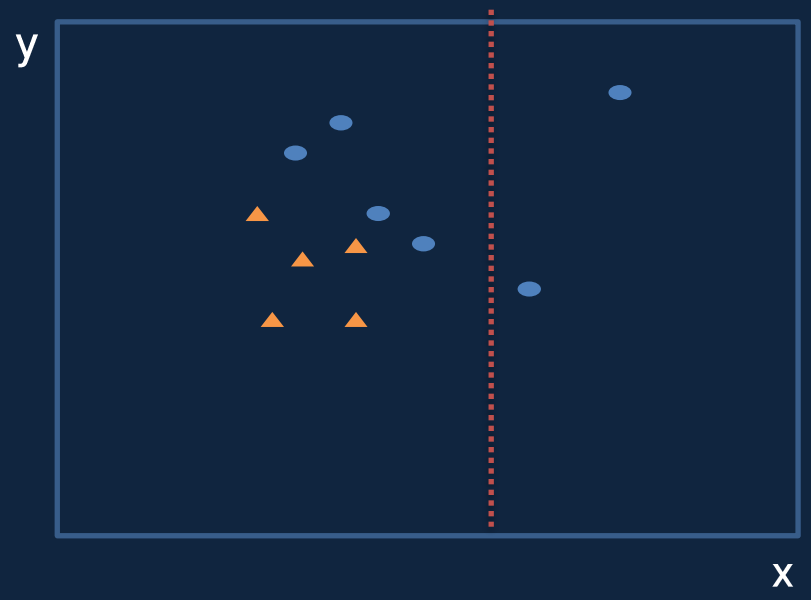


$$G_{left} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = \frac{6}{16} \qquad G_{right} = 1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = \frac{20}{49}$$

$$J(k, t_K) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

$$= \frac{4}{11} * \frac{6}{16} + \frac{7}{11} * \frac{20}{49}$$

4 엑스트라트리



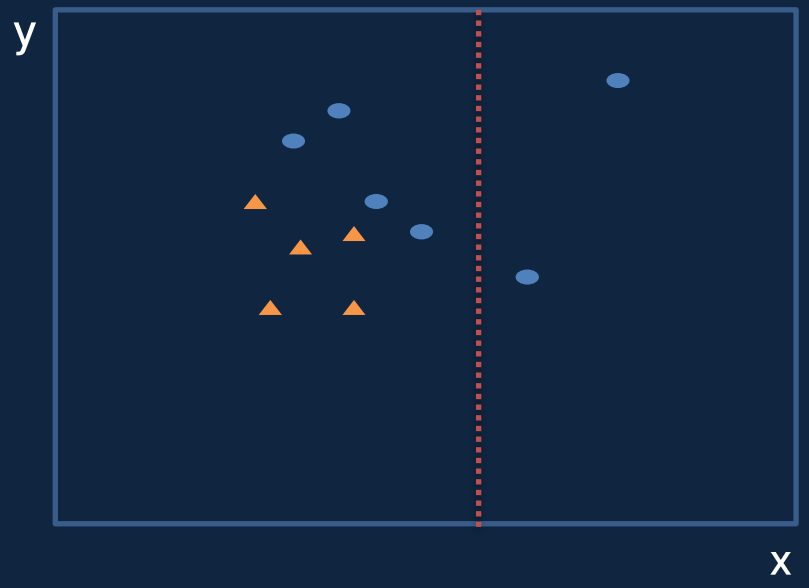
$$G_{left} = 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2 = \frac{40}{81} \qquad G_{right} = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0$$

$$J(k, t_K) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

$$= \frac{9}{11} * \frac{40}{81} + \frac{2}{11} * 0$$

4

엑스트라트리



$$\min\_samples\_split = 2$$

· DT에서 최적의 임계값을 찾는 과정을 생략하고 무작위로 분할

√ 계산속도 Up, 무작위성 Up

5

베이지안 최적화

1. Manual Search
2. Grid Search → 노하우 + 오래걸림 + 완전 무작위
3. Random Search

→ Bayesian Optimization 등장 !

5

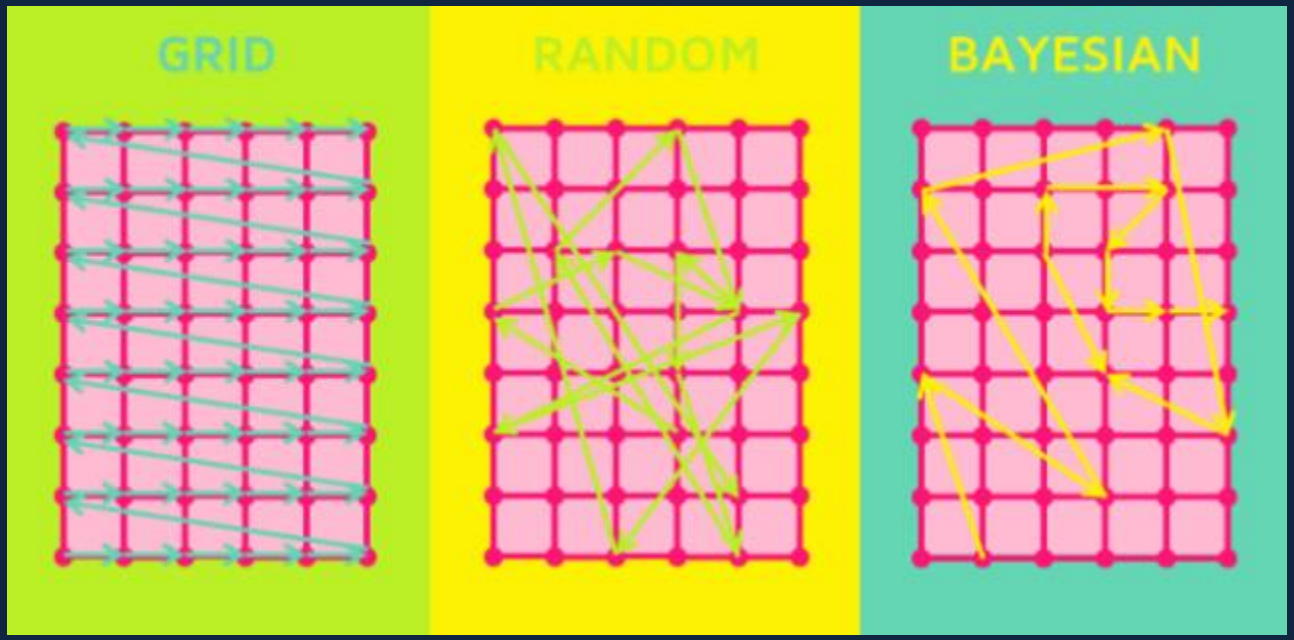
베이지안 최적화

베이지안 최적화란 ?

- ‘사전 정보’를 기반으로 ‘최적해’를 탐색
- 파라미터 튜닝 과정에서 얻는 정보를 반영하여 다음 하이퍼파라미터를 추정하는 것
- 최적의 탐색값을 찾는 속도가 빠르다.

5

베이지안 최적화



Q n A

<과제>

이번 주차에 배운 다양한 앙상블 모델을 통해 저번 과제 성능 더 올려보기

THANK YOU