Ensemble Learning

Business Analytics (IME654)

2022. 11. 21

Team: 동기사랑

Member: 김창현, 정진용

Ensemble Learning

해당 발표자료는

고려대학교 산업경영공학과

강필성 교수님: 비즈니스 애널리틱스(IME654)
김성범 교수님: 다변량 통계분석 및 데이터 마이닝(IME567)

의 강의자료를 사용했음을 미리 밝힙니다.

Ensemble Learning

❖ Ensemble learning

✓ Model ensemble은 여러 모델들을 함께 사용하여 기존보다 성능을 더 올리는 방법을 말함

1. Bagging

- Bootstrap Aggregating의 약자이며 bootstrap을 이용하는 방법
  - ✓ Bootstrap: 주어진 데이터셋에서 random sampling을 거쳐 새로운 데이터셋을 만들어내는 과정
  - ✓ 만들어진 여러 데이터셋을 바탕으로 결과를 voting
- ex) Random Forest

2. Voting

- Voting은 크게 Hard voting과 soft voting으로 나눌 수 있음
  - ✓ Hard voting: 각 하위 학습 모델(weak learner)들의 예측 결과값을 바탕으로 다수결 투표하는 방식
  - ✓ Soft voting: 각 하위 학습 모델(weak learner)들의 예측 확률값의 평균 또는 가중치 합을 사용하는 방식

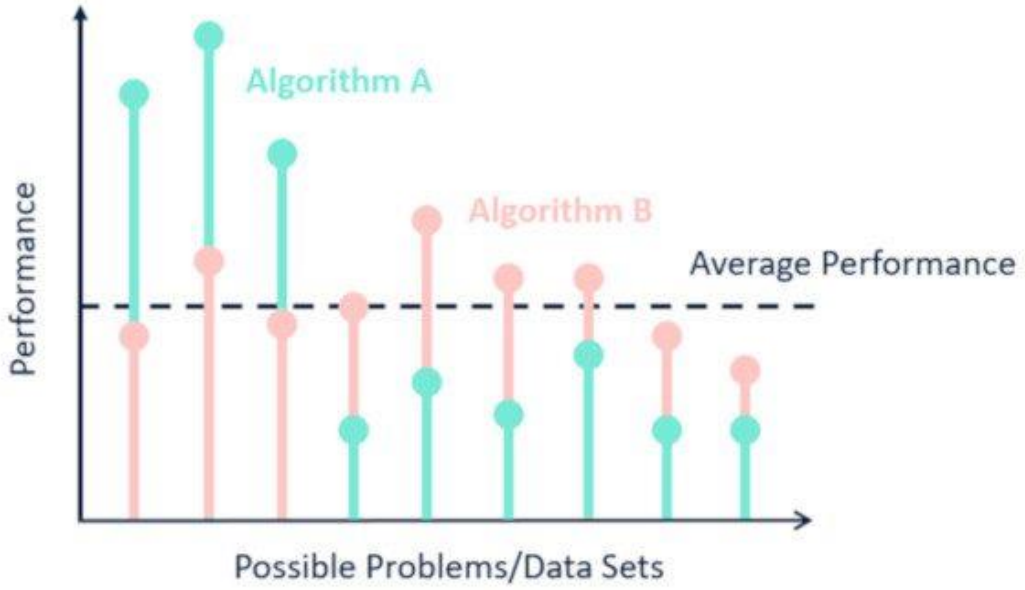
3. Boosting

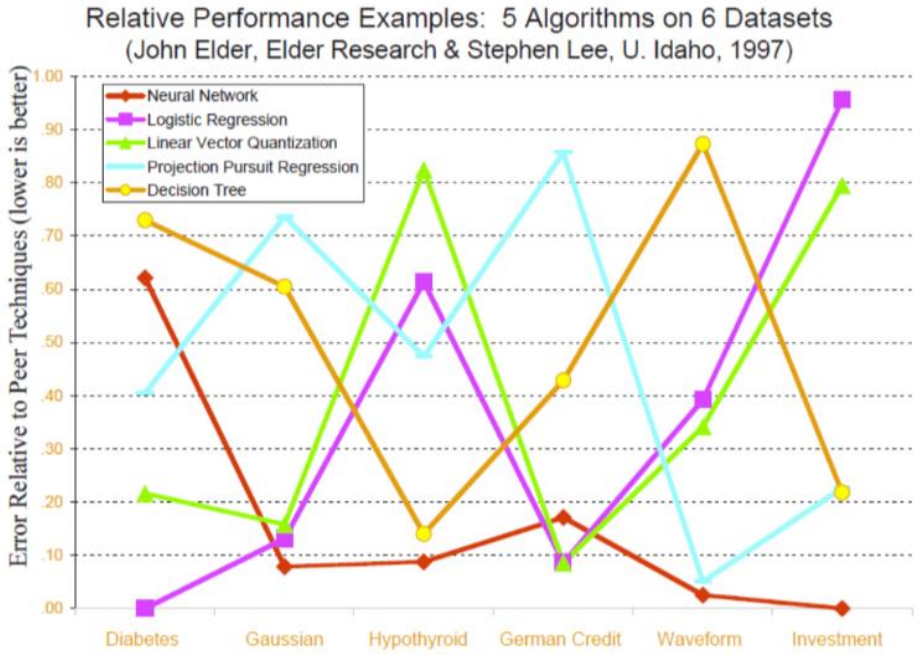
- 모델 iteration의 결과에 따라 데이터셋 샘플에 대한 가중치를 부여하며 모델을 업데이트하는 방식
- 반복할 때마다 각 샘플의 중요도에 따라 다른 분류기가 만들어지고 최종적으로는 모든 iteration에서 생성된 모델의 결과를 votin함
- Adaptive Boosting(AdaBoost)와 Gradient Boosting Model(GBM) 계열로 나눌 수 있음

Background

❖ No Free Lunch Theorem?

- 머신러닝은 다양한 샘플 데이터에 학습(fitting)을 시킴으로써 일반화되기를 목적으로 함
- '모델이 학습을 한다'라는 의미는 샘플 데이터로 구성된 가설 공간속에서 데이터에 알맞은 가설을 채택하는 것이라 볼 수 있음
- 다양한 가설이 존재하는 만큼 귀납적 편향의 문제에 마주침





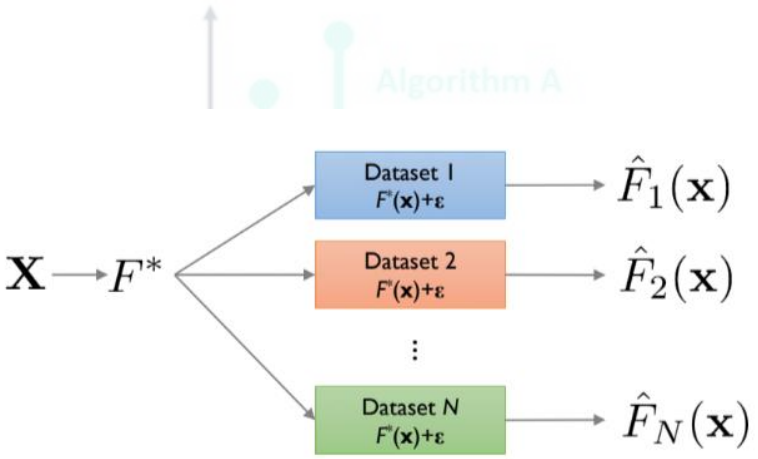
다양한 데이터셋에 대한 각 알고리즘들 성능 비교

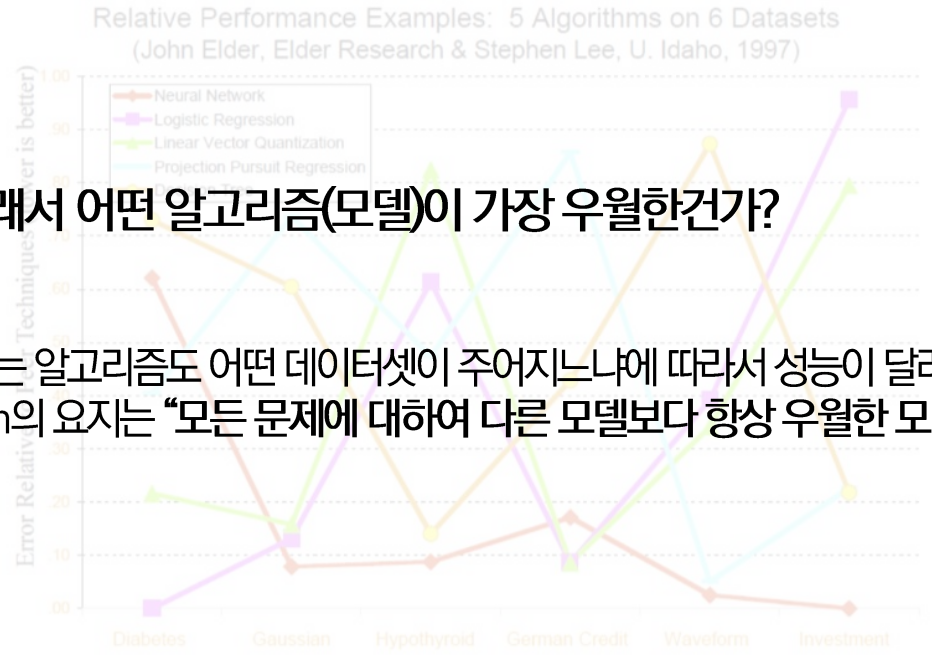
https://www.kdnuggets.com/2019/09/no-free-lunch-data-science.html

Background

❖ No Free Lunch Theorem?

- 머신러닝은 다양한 샘플 데이터에 학습(fitting)을 시킴으로써 일반화되기를 목적으로 함
- '모델이 학습을 한다'라는 의미는 샘플 데이터로 구성된 가설 공간속에서 데이터에 알맞은 가설을 채택하는 것이라 볼 수 있음
- 다양한 가설이 존재하는 만큼 귀납적 편향의 문제에 마주침





다양한 데이터셋에 대한 각 알고리즘들 성능 비교

그래서 어떤 알고리즘(모델)이 가장 우월한건가?

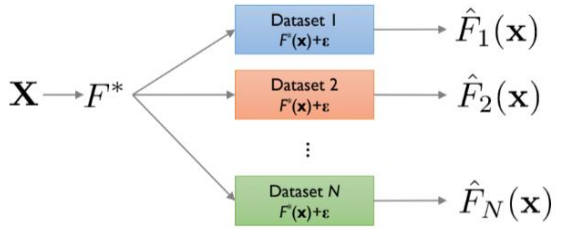
성능이 좋다고 하는 알고리즘도 어떤 데이터셋이 주어지느냐에 따라서 성능이 달라짐
No Free Lunch Theorem의 요지는 **"모든 문제에 대하여 다른 모델보다 항상 우월한 모델은 없다"**

https://www.kdnuggets.com/2019/09/no-free-lunch-data-science.html

Background

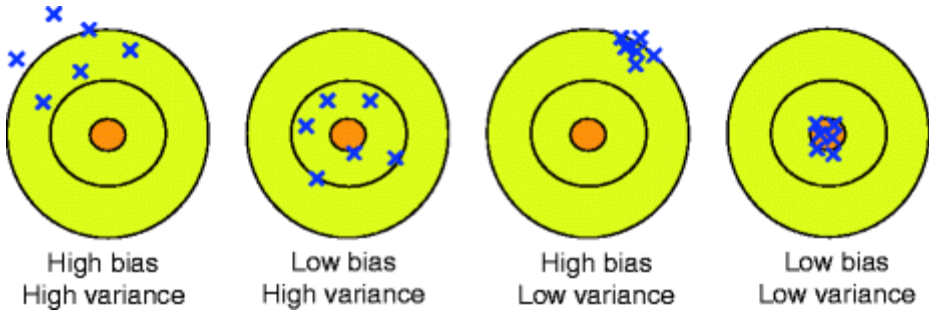
❖ Bias-Variance Decomposition

- 특정 데이터에 대한 오차를 편향과 분산에 의한 에러로 나눌 수 있음
- 편향이 높으면 과소적합이 발생하며 분산이 높으면 과적합이 발생함
- Bias는 정답과 평균 추정치 차이, Variance는 평균 추정치와 특정 데이터셋에 대한 추정치 차이



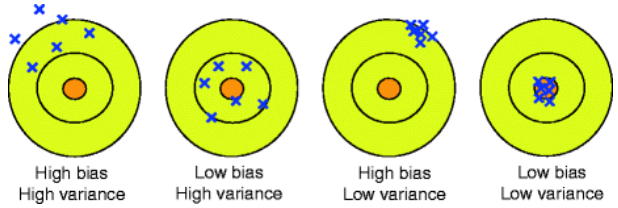
$$y = F^*(\mathbf{x}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

$$\bar{F}(\mathbf{x}) = E[\hat{F}_D(\mathbf{x})]$$



$$\begin{aligned}
Error(X_0) &= E[(y - \hat{F}(X)|X = X_0)^2] \\
&= E[(F^*(X) + \varepsilon - \hat{F}(X))^2] \qquad (\because y = F^*(X) + \varepsilon) \\
&= E[(F^*(X) - \hat{F}(X))^2] + \sigma^2 \\
&= E[(F^*(X) - \bar{F}(X) + \bar{F}(X) - \hat{F}(X))^2] + \sigma^2 \\
&= E\left[\left(F^*(X) - \bar{F}(X)\right)^2 + \left(\bar{F}(X) - \hat{F}(X)\right)^2 + 2\left(F^*(X) - \bar{F}(X)\right)\left(\bar{F}(X) - \hat{F}(X)\right)\right] + \sigma^2 \\
&= E\left[\left(F^*(X) - \bar{F}(X)\right)^2\right] + E\left[\left(\bar{F}(X) - \hat{F}(X)\right)^2\right] + \sigma^2 \\
&= Bias^2\left(\hat{F}(X_0)\right) + Var^2\left(\hat{F}(X_0)\right) + \sigma^2
\end{aligned}$$

Background



❖ Ensemble Learning

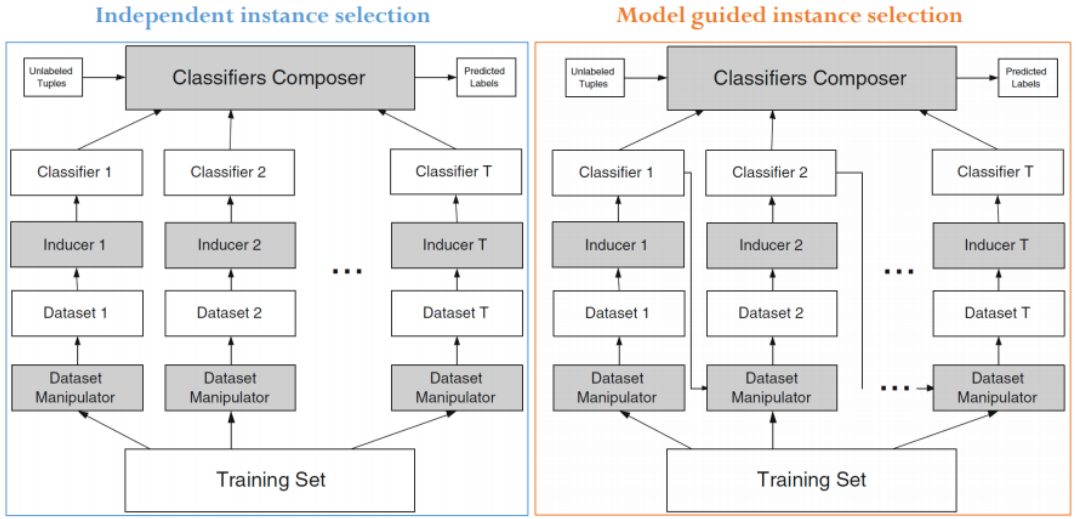
- 앙상블의 목적은 각 단일 모델의 좋은 성능을 유지하면서 다양성(diversity)을 확보하는 데 있음
  - ✓ Implicit diversity를 확보: 전체 데이터셋의 부분집합에 해당하는 여러 데이터셋을 준비한 뒤 따로 학습
  - ✓ Explicit diversity를 확보: 먼저 생성된 모델의 측정값으로부터 새로운 모델을 생성하여 학습

단일 모델 : Decision Tree, ANN, SVM, k값이 작은 K-NN

Bagging이나 Random Forest 등을 통해서 분산을 줄이자!

단일 모델 : Logistic Regression, k값이 큰 K-NN

Boosting을 통해서 분산을 줄이자!



Background

❖ Ensemble Learning

- 다양성을 확보한다면 단일 모델보다 앙상블 모델이 좋은 성능을 보이는 건지?
  - ✓ 수식 증명

Background

- ❖ Ensemble learning VS single algorithm learning
  - 실제 앙상블 모델들이 단일 모델보다 좋은 성능을 보임

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Getting Started

We've built a few resources to help you get started with the dataset.
Download a copy of the dataset (distributed under the CC BY-SA 4.0 license):

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899
6 Apr 18, 2021	TransNets + SFVerifier + SFEnsembler (ensemble) *Senseforth AI Research* https://www.senseforth.ai/	90.487	92.894
6 Dec 01, 2020	EntitySpanFocusV2 (ensemble) *RICOH_SRCB_DML*	90.521	92.824

https://rajpurkar.github.io/SQuAD-explorer/

✓2016

Object detection (DET)[top]

Task 1a: Object detection with provided training data

Ordered by number of categories won

Team name	Entry description	Number of object categories won	mean AP
CUImage	Ensemble of 6 models using provided data	109	0.662751
Hikvision	Ensemble A of 3 RPN and 6 FRCN models, mAP is 67 on val2	30	0.652704
Hikvision	Ensemble B of 3 RPN and 5 FRCN models, mean AP is 66.9, median AP is 69.3 on val2	18	0.652003

Object localization (LOC)[top]

Task 2a: Classification+localization with provided training data

Ordered by localization error

Team name	Entry description	Localization error	Classification error
Trimps-Soushen	Ensemble 3	0.077087	0.02991
Trimps-Soushen	Ensemble 4	0.077429	0.02991
Trimps-Soushen	Ensemble 2	0.077668	0.02991
Trimps-Soushen	Ensemble 1	0.079068	0.03144

✓2017

Object detection (DET)[top]

Task 1a: Object detection with provided training data

Ordered by number of categories won

Team name	Entry description	Number of object categories won	mean AP
BDAT	submission4	85	0.731392
BDAT	submission3	65	0.732227
BDAT	submission2	30	0.723712
DeepView(ETRI)	Ensemble_A	10	0.593084
NUS-Qihoo_DPNs (DET)	Ensemble of DPN models	9	0.656932
KAISTNIA_ETRI	Ensemble Model5	1	0.61022
KAISTNIA_ETRI	Ensemble Model4	0	0.609402
KAISTNIA_ETRI	Ensemble Model2	0	0.608299
KAISTNIA_ETRI	Ensemble Model1	0	0.608278
KAISTNIA_ETRI	Ensemble Model3	0	0.60631

Object localization (LOC)[top]

Task 2a: Classification+localization with provided training data

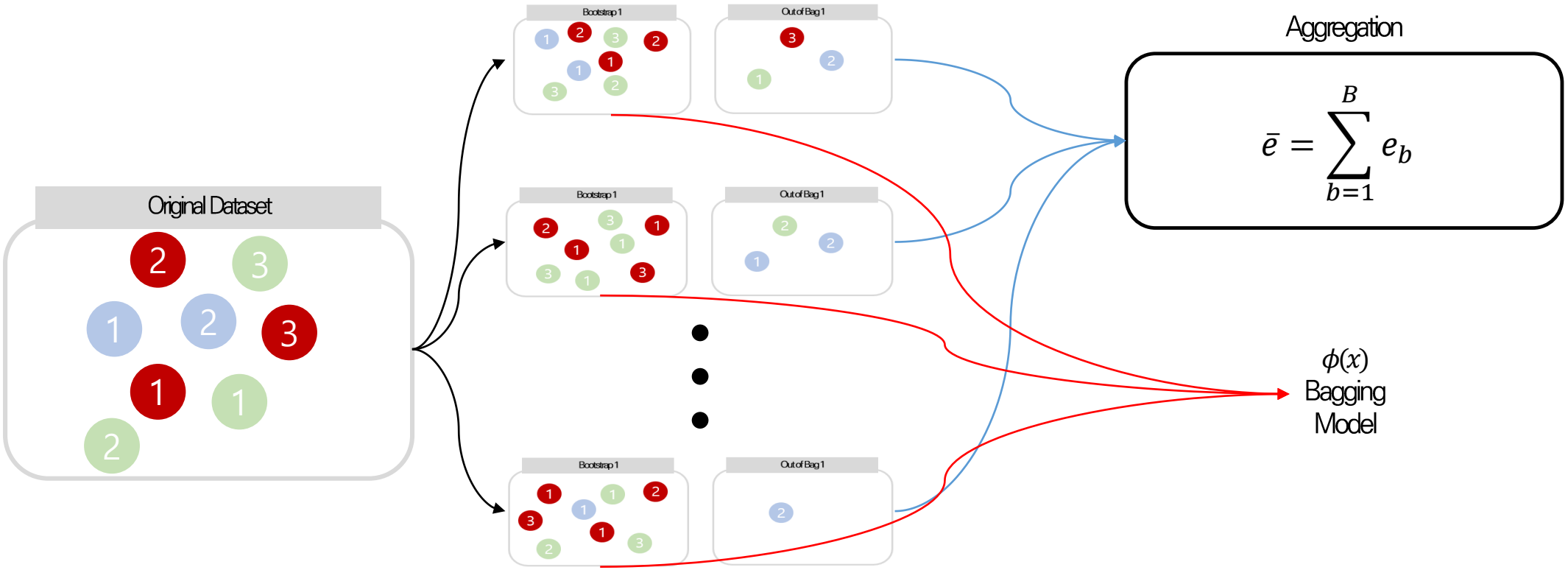
Ordered by localization error

Team name	Entry description	Localization error	Classification error
NUS-Qihoo_DPNs (CLS-LOC)	[E3] LOC: Dual Path Networks + Basic Ensemble	0.062263	0.03413
Trimps-Soushen	Result-3	0.064991	0.02481
Trimps-Soushen	Result-2	0.06525	0.02481
Trimps-Soushen	Result-4	0.065261	0.02481
Trimps-Soushen	Result-5	0.065302	0.02481
Trimps-Soushen	Result-1	0.067698	0.02481

Ensemble Learning

❖ Booststrap Aggregating(Bagging)이란?

- 배깅은 주어진 데이터셋에 대해 bootstrap 샘플링을 이용하여 단일 알고리즘 모델보다 더 좋은 모형을 만들 수 있는 앙상블 기법임
- 각 데이터셋은 복원추출을 통해 기존 데이터셋만큼의 크기를 갖도록 샘플링됨
- 개별 샘플링 된 데이터셋은 bootstrap이라 함



Ensemble Learning

❖ Booststrap Aggregating(Bagging)이란?

- 배깅은 주어진 데이터셋에 대해 bootstrap 샘플링을 이용하여 단일 알고리즘 모델보다 더 좋은 모형을 만들 수 있는 앙상블 기법임
- 각 데이터셋은 복원추출을 통해 기존 데이터셋만큼의 크기를 갖도록 샘플링됨
- 개별 샘플링 된 데이터셋은 bootstrap이라 함

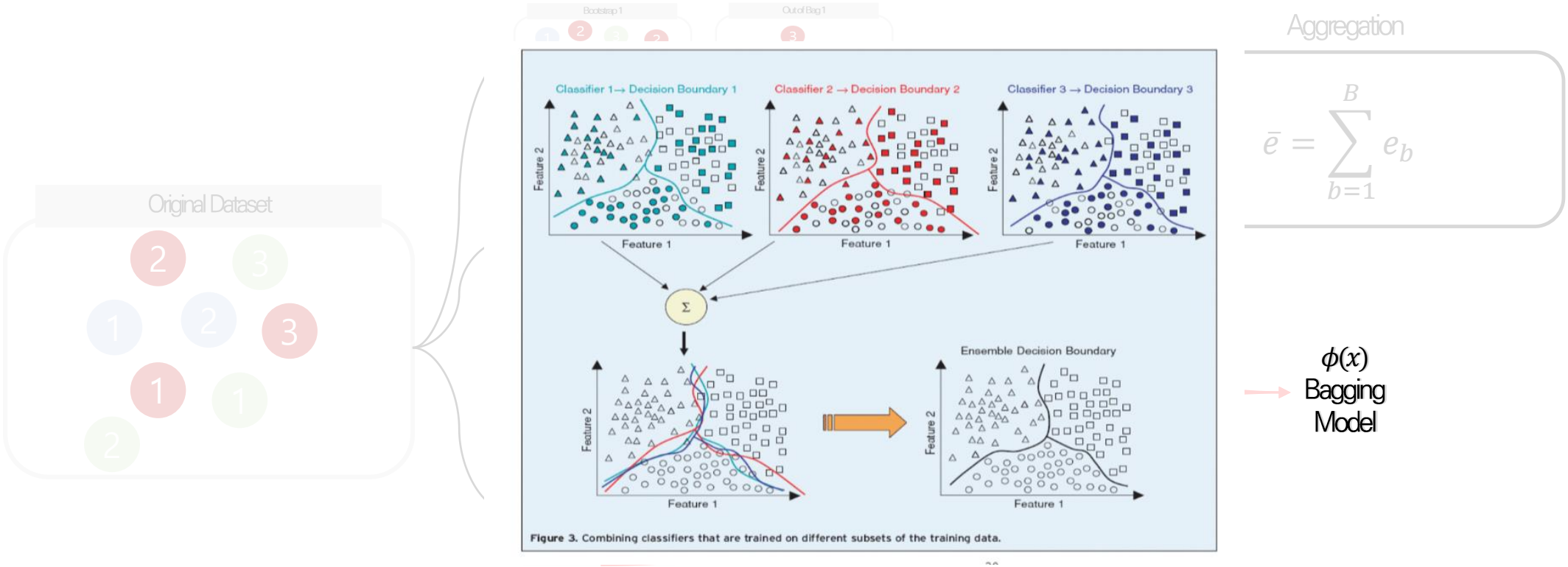
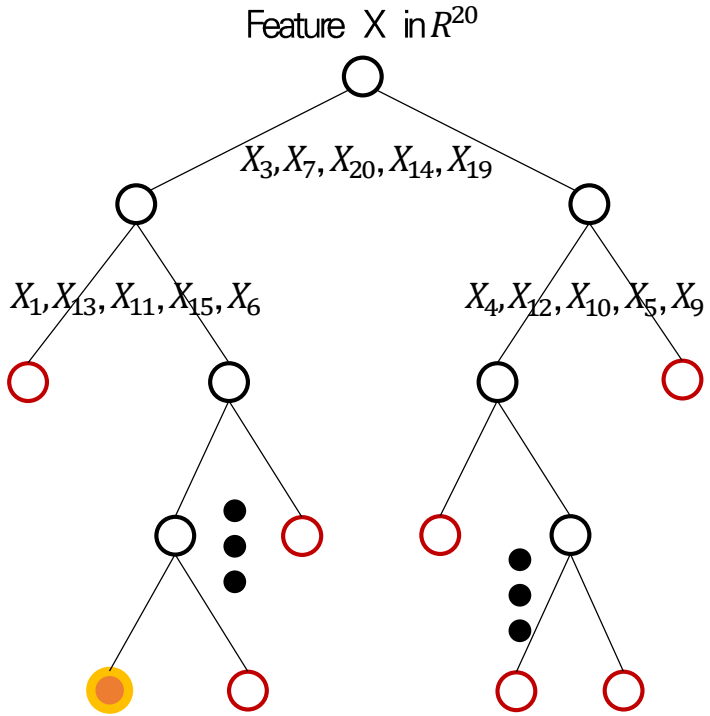


Figure 3. Combining classifiers that are trained on different subsets of the training data.

Ensemble Learning

❖ Randomly chosen predictor variables이란?

- 앙상블의 diversity를 확보하기 위한 기법
- 각 decision tree 분기점을 탐색할 때, 기존 변수 수보다 적은 변수 수를 임의로 선택하여 분기함



(1) 변수 X_i가 tree split에 한번도 사용되지 않았다면,
OOB Error of the original Data e_i = OOB Error of the Permuted Data p_i

(2) 변수 X_i가 tree split에 중요하게 사용되었다면,
OOB Error of the original Data e_i < OOB Error of the Permuted Data p_i

- m번째 tree에서 변수 i에 대한 Random permutation 전후 OOB error의 차이

$$d_i^m = p_i^m - e_i^m$$

- 전체 Tree들에 대한 OOB error 차이의 평균 및 분산

$$\bar{d}_i = \frac{1}{m}\sum_{i=1}^{m} d_i^m, \quad s_i^2 = \frac{1}{m-1}\sum_{i=1}^{m}(d_i^m - \bar{d}_i)^2$$

- i번째 변수의 중요도: $v_i = \frac{\bar{d}_i}{s_i}$

Ensemble Learning

❖ Random Forest

- Decision tree으로 ensemble을 하는 기법
- Ensemble을 할 때, 다양성을 확보하기 위하여 bagging과 randomly chosen predictor variables 두 가지 기법을 사용함