Machine Learning 2

Ensemble, Bagging, RandomForest

Dept. SW and Communication Engineering

Prof. Giseop Noh (kafa46@hongik.ac.kr)

Contents

Recap: Decision Tree

Recap: Information Entropy

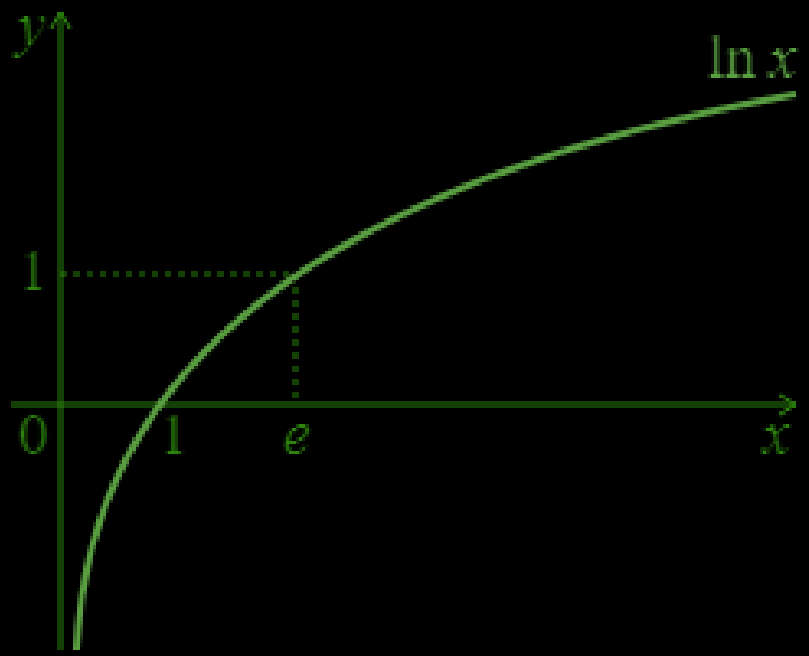
- **Information, *I***
  - 도대체 정보를 어떻게 표현할까?
  - 어떤 정보가 가치 있을까?
  - 내일은 해가 동쪽에서 뜬다.
  - 내일은 해가 서쪽에서 뜬다.
  - 교수님은 강의가 있는 날 출근하신다.
  - 교수님은 내일 퇴직하신다.

    :

$$Information\ (I) \propto \frac{1}{p(x)} = p(x)^{-1}$$

$$x: random\ variable$$

Recap: Which Function we choose in Information



$$I(x) \propto \frac{1}{p(x)} = p(x)^{-1}$$

x: *random variable*

$$I(x) = \log_a \frac{1}{P(x)} = \log_a P(x)^{-1} = -\log_a P(x) \propto -\ln P(x)$$

Recap: Information Entropy

Information Entropy

- Expected Information of Individual Events
- It becomes easier if you think in terms of the formula for average.
- Average

  → Expected Value: Multiply each outcome by its probability

  → Then sum up everything.

Claude Shannon (1916~2001)
새넌에 의해 제안되어
'새넌 엔트로피'라고 불리기도 함

$$H(P) = H(x) = E_{x\sim P}[I(x)] = E_{x\sim P}[-\log P(x)]$$

$$= -\sum_{x} P(x) \cdot \log P(x) = \sum_{i=1}^{n} p_i \cdot \log p_i$$

Recap: Information Gain

Information Gain

- A metric used to evaluate the effectiveness of each attribute when selecting a splitting criterion for nodes in the training data
- If data is split based on the attribute that maximizes information gain at each node,

  ➔ resulting tree structure will be the most efficient in terms of classification

The greater the Information Gain, the more informative the attribute

$$Gain(D, A) = H(D) - H_A(D)$$

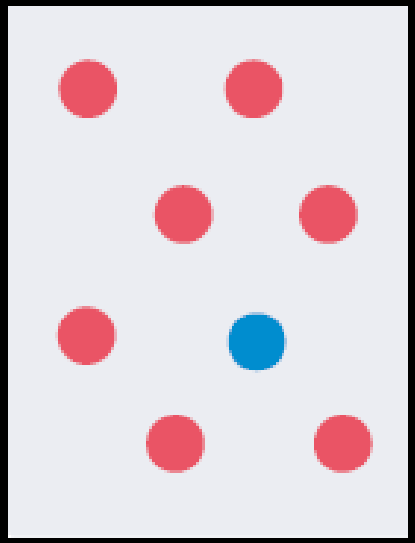
- $H(D)$: 전체 데이터 집합 D 에 대한 엔트로피
  - 데이터를 분할하기 전, 현재 전체 데이터가 얼마나 불확실하고 섞여 있는지를 수치로 표현
- $H_A(D)$: 속성 A를 기준으로 데이터를 분할했을 때의 전체 엔트로피
  - 속성 A로 데이터를 나눠본 뒤, 나눠진 각 그룹의 엔트로피를 계산하고,
  - 그 그룹들의 크기를 고려해 전체 평균 엔트로피를 구한 것!

Recap: Impurity

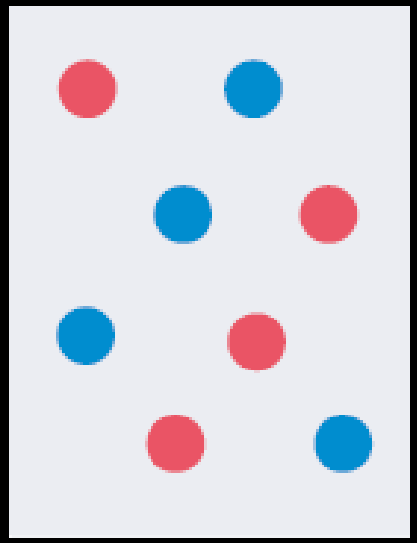
■ **Impurity (불순도)**

- If the data contains only one color (i.e., one class), the impurity is low
- If the data contains a mix of different colors (i.e., multiple classes evenly mixed)

➔ Impurity is high



Low impurity
(mostly one class)



High impurity
(balanced class mix)

Gini Index

Gini Index

- A metric that measures the impurity of a dataset
- Gini index ranges from 0 to 1
  - Gini = 0 → Pure node (perfectly classified)
  - Gini = 1 → Maximum impurity (completely mixed classes)
- The lower the Gini index, the purer the node

$$GINI(D) = 1 - \sum_j p(j)^2 \text{ , where } p(j): Propotion\ of\ class\ j\ in\ dataset\ D$$

CART Algorithm

- Uses the Gini index to split nodes
- Grows the tree by selecting the split that minimizes the Gini index

Ensemble

Ensemble Learning

Meaning of “Ensemble”

Online wiki:

전체적인 어울림이나 통일. '조화'로 순화한다는 의미의 프랑스어이며
음악에서 2인 이상이 하는 노래나 연주를 말하며 흔히 뮤지컬에서
주, 조연 배우들 뒤에서 화음을 넣으며 춤을 추고 노래를 부르면서
분위기를 돋구는 역할을 말한다.

Ensemble Learning

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

Core Insight in Ensemble

How do we make decisions in our life?

- Decision made by a single expert?

Which is the better approach?

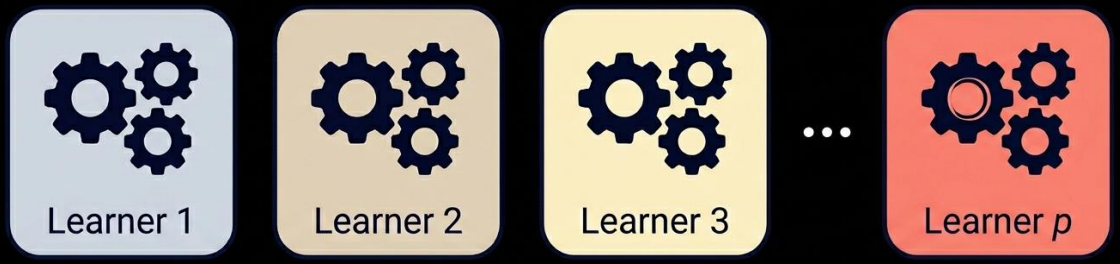
- Or by considering diverse opinions from many people?

⇨ Importance of collective intelligence

Collective Intelligence in ML

Machine learning follows a similar idea!

- Build multiple learners using different learning algorithms
- Combine their outputs to make better decisions



Assumption:

Aggregating multiple opinions leads to

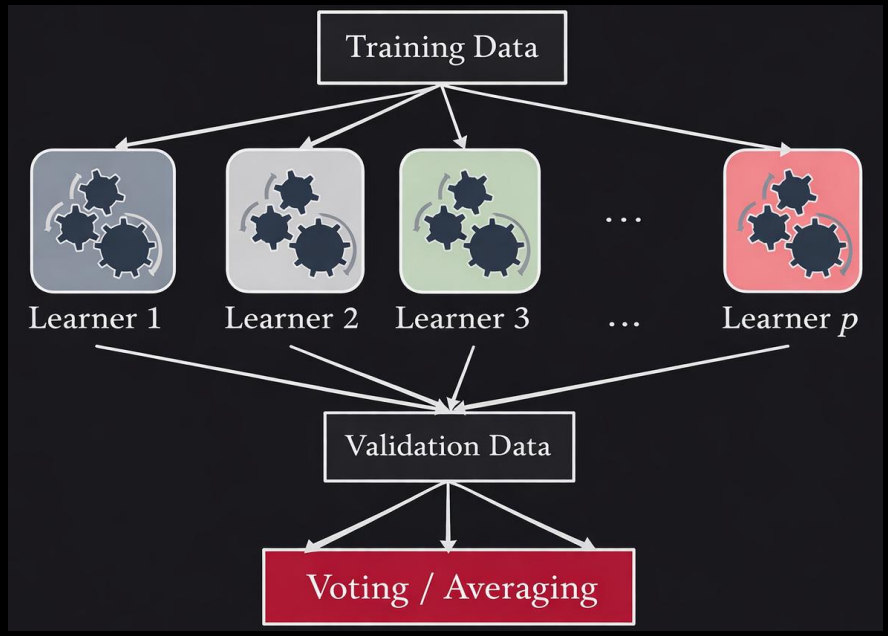
more accurate predictions

Ensemble Learning

Combine multiple models to create a new model

Base Model

- A model used as a basic component when combining multiple models
- The final model is built by combining several base models
  - Also referred to as:
    - ✓ Weak Learner
    - ✓ Classifier
    - ✓ Base Learner
    - ✓ Single Learner



Mathematical Principle of Ensemble Learning

Law of Large Numbers

The average of randomly sampled values from a large population is likely to be close to the true population mean

Mathematical Principle of Ensemble Learning

Experiment: Coin Toss

Toss 10 times

Estimate probability of heads & tails

Toss 1,000,000 times

Estimate probability of heads & tails

Which experiment is more reliable?

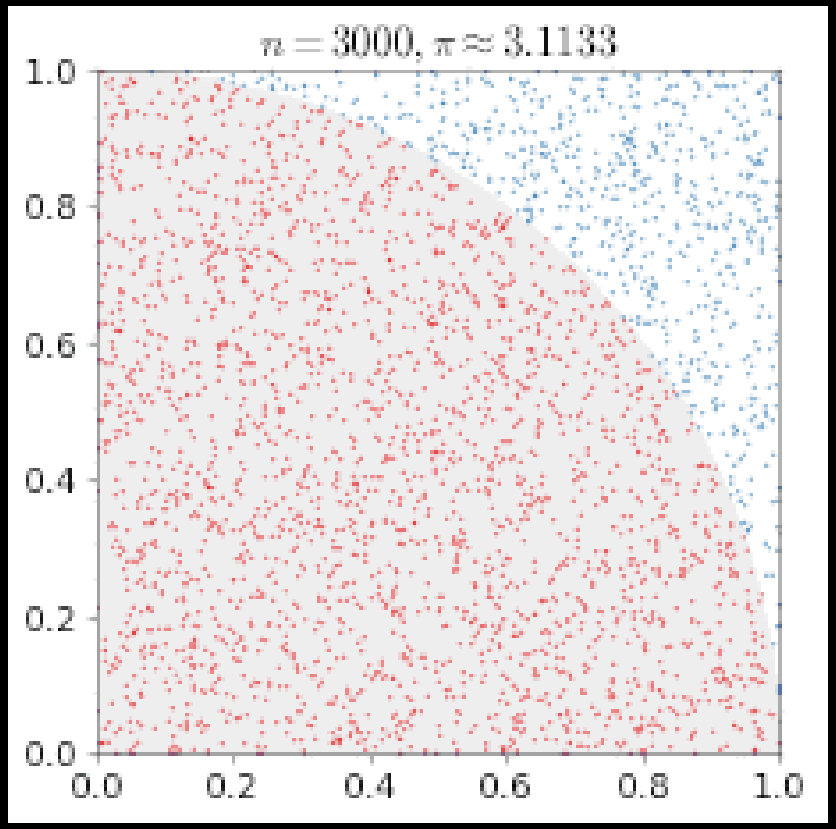
As the number of samples increases,
the sample mean converges
to the true mean!

In Practice: Monte Carlo Method

Law of Large Numbers

As the number of samples increases, the sample mean converges to the true mean!

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mathbb{E}[X]$$



Monte Carlo Method

$$\mathbb{E}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

- Use random sampling to approximate expectations
- Useful when exact computation is difficult

Binomial Distribution

Describes events with two possible outcomes

Examples:

- ✓ 0 or 1
- ✓ True or False,
- ✓ Success or Failure

⋮

Can be used to model the probability of correct predictions when combining multiple learners!

Random Variable: k

When an experiment is repeated n times,

and a specific event occurs k times

$$P(K = k)$$

$$= f(k; n, p)$$

$$= \binom{n}{k} p^k (1-p)^{n-k}$$

$$where \ \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Mathematical Principle of Ensemble Learning

Interpreting Binomial Distribution for Ensemble Learning

- L: number of learners
- k: number of correct predictions
- p: probability of error (error rate)

Ensemble Probability

$$P(k) = f(k; L, p) = \binom{L}{k} p^k (1-p)^{L-k}$$

If the error probability $p_{error} > \frac{1}{2}$

Increasing # of learners $L \rightarrow$ Decrease overall performance!

Each base learner should perform better than random guessing i.e., Accuracy $\geq \frac{1}{2}$

Also, base learners should be independent!

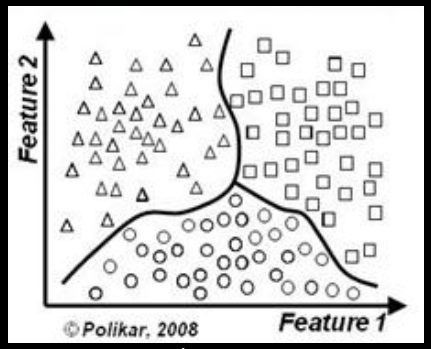
Simple Example: Better Learners & More Performance

실험횟수	정답	Learner 1	Learner 2	Learner 3	투표 결과
1	1	1	1	0	1
2	0	1	0	0	0
3	1	1	1	0	1
4	1	0	0	1	0
5	1	1	1	1	1
6	0	1	1	0	1
7	0	0	0	1	0
8	0	0	1	0	0
정확도		5/8 = 0.625	5/8 = 0.625	5/8 = 0.625	6/8 = 0.75

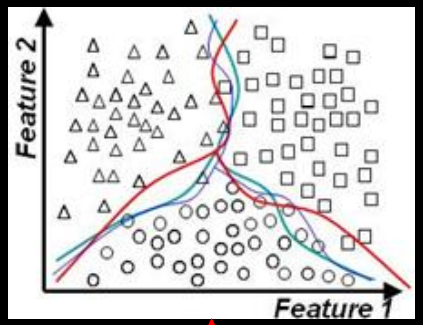
Ensemble Learning Approach

Approach:

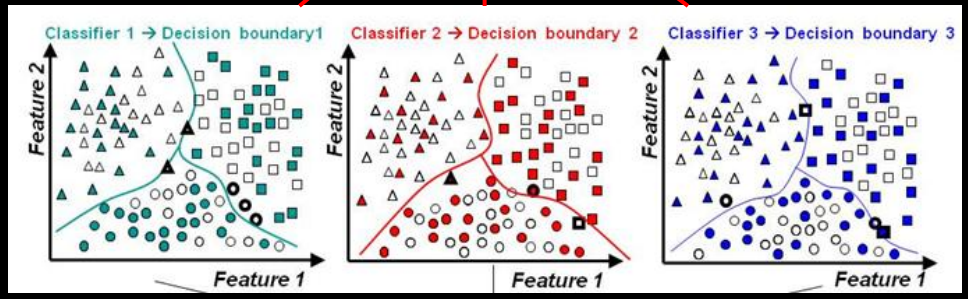
- Use weak learners with low individual performance
- Combine multiple learners to achieve higher overall performance
- **Apply majority voting** among learners
  - ✓ Aggregate multiple classifiers
  - ✓ Individual errors are compensated
- Leverage the power of collective intelligence



Ensemble-based Decision Boundary

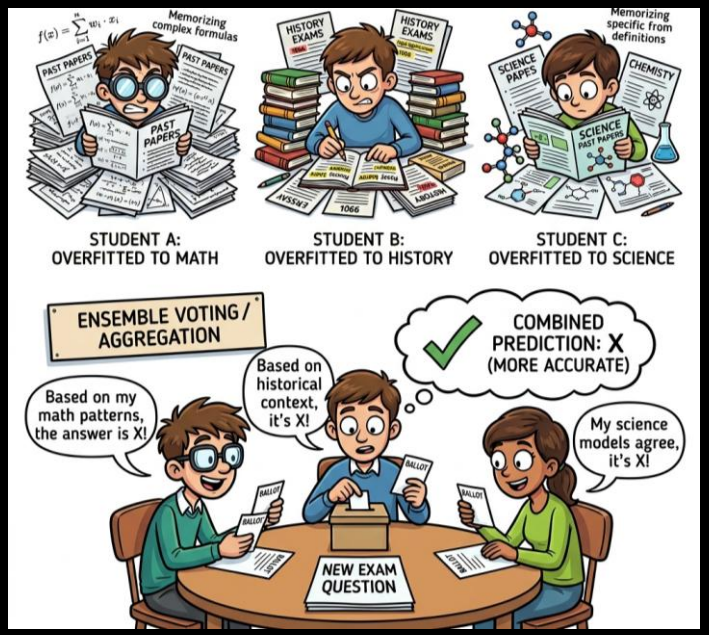


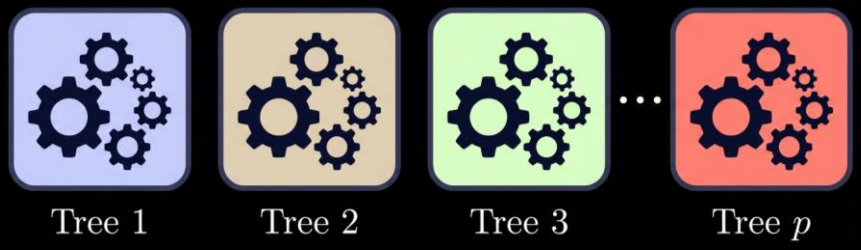
Merge & Voting



Characteristics on Ensemble Learning

1. Strategies to reflect Base Learner's opinions.
2. Overfitting model is applied by default.
3. Mainly tree-based models.





Note:

- Ensemble refers to a combination of different learners.
- Not necessarily using trees

Types of Ensemble Learning

Problems on Pure Tree based

- Features of basic tree models
- Same Data → Same Results
- Combining multiple tree models results in the same result
- No Ensemble Learning Effect

Extreme Performance

- Slightly better performance than previous best approach
- Lots of computation
- High complexity

Bagging

Reconfiguring Data

→ Configuring Models in Variety

RandomForest

Reconfigure Data

+ Reconfigure Variables

→ Configuring Models in Variety

Boosting

More weights on unpredictable data

Studying wrong answer again

- Adaboost
- GradientBoost:

Xgboost, LightGBM, Catboost

Exercise with Codes

Exercise:

1. Voting System

2. Numerical Analysis

Bagging

Bagging: Bootstrap

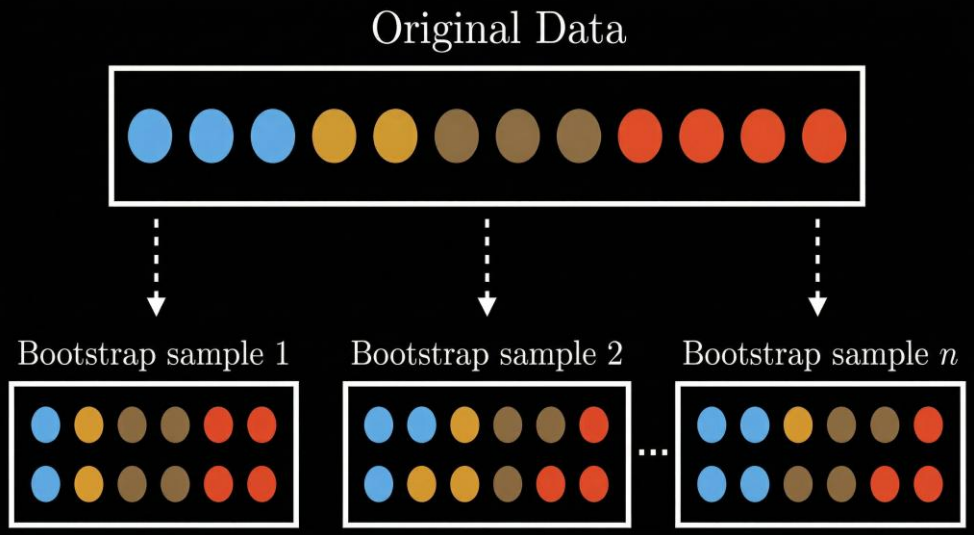
Bootstrap

Sampling with replacement

The same data point can be selected multiple times

Repeat:

Select one sample → Put it back → Select again → Put it back ⋯



Property:

- Some data points are not selected in each sample on average: 36%

Bootstrap Aggregating

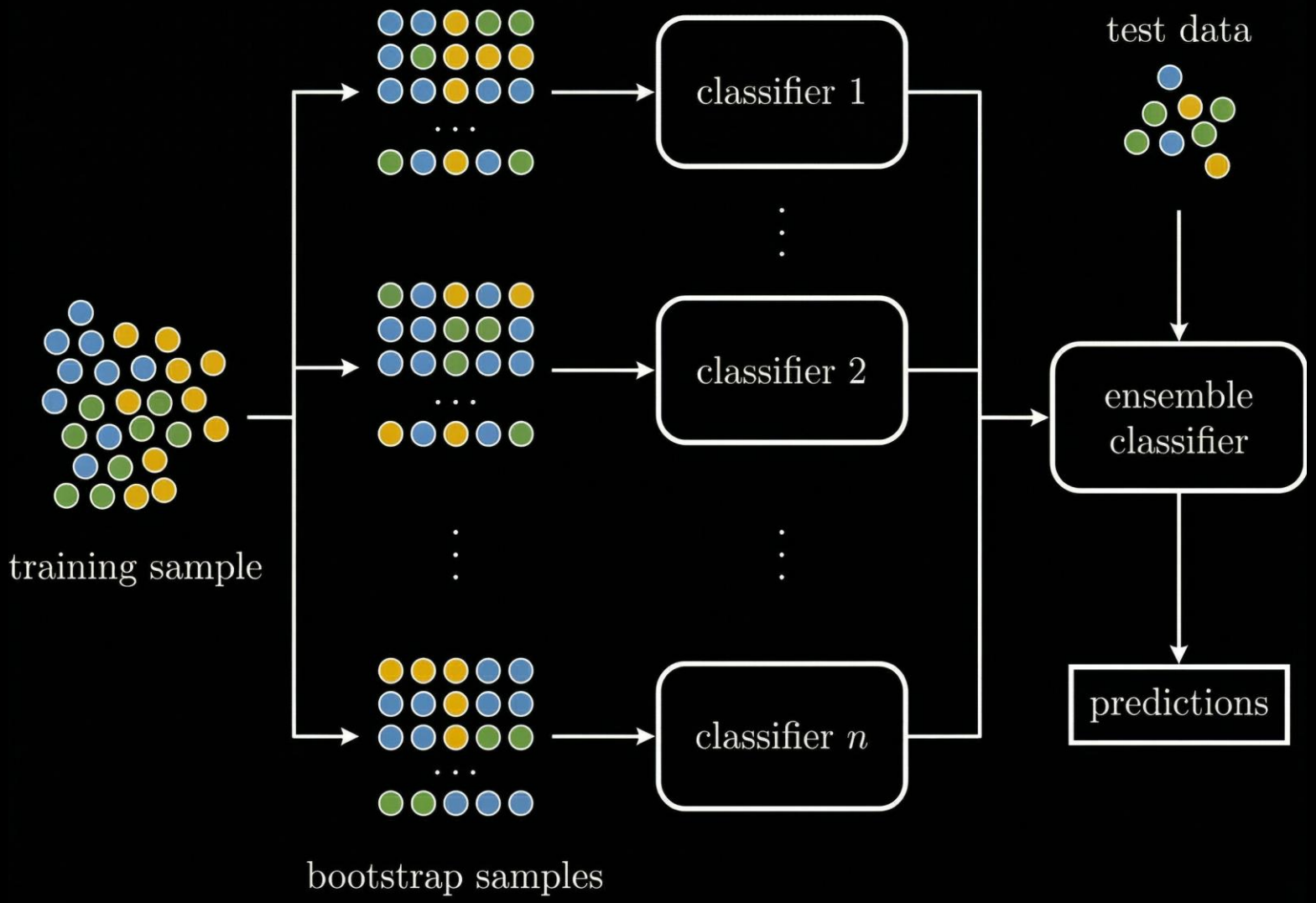
Training Process

- Generate multiple bootstrap samples from the original training data
- Train a separate classifier on each sample
  - classifier 1
  - classifier 2
  - ⋮
  - classifier n

Out-of-Bag (OOB) data

- Data not selected during bootstrap sampling
- Used as test data for each classifier

Conceptual Understanding



Exercise with Codes

Exercise:

1. Bagging System

RandomForest

Motivation of RandomForest

잠깐 복습, 공분산(Covariance)

■ **확률변수: 각각의 근원사권들에 실수 값을 대응시키는 함수**

- 두 개의 동전을 던지는 실험 → 동전 앞면의 개수 (X라는 확률변수) $f(x) \longrightarrow X(x)$와 비슷한 개념
  - X(앞, 앞) = 2, X(앞, 뒤) = 1, X(뒤, 앞) = 1, X(뒤, 뒤) = 0

■ **확률분포: 확률 변수에서 확률 값의 함수로 표현, 대부분 $f(x)$로 표기함**

- $f(2) = P(X = 2) = P(H, H) = {}^1/_4$, $f(1) = P(X = 1) = P\big((H, T), (T, H)\big) = {}^2/_4 = {}^1/_2$,
  $f(0) = P(X = 0) = P(T, T) = {}^1/_4$

■ **확률변수의 기대값(중심 경향값, 평균): $E(X) = \mu = \sum_{i=1}^{n} x_i f(x_i)$**

■ **확률변수의 분산: $Var(X) = E(X - \mu)^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)$**

■ **공분산: 두개의 확률변수 X, Y가 어떤 관계를 가지면서 변화하는지에 대한 척도**

- $Cov(X, Y) = E(X - \mu_X)(Y - \mu_Y) = \sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y) f(x_i, y_i)$
- X, Y가 독립이면 $Cov(X, Y) = 0$

Bagging Pros vs. Cons

Advantage of bagging:

- Preserves the bias of individual trees
- Significantly reduces variance
- More robust to noisy data

Reduce covariance between trees

→ Motivation for Random Forest

Disadvantage of bagging:

Depends on both the variance and covariance of individual trees

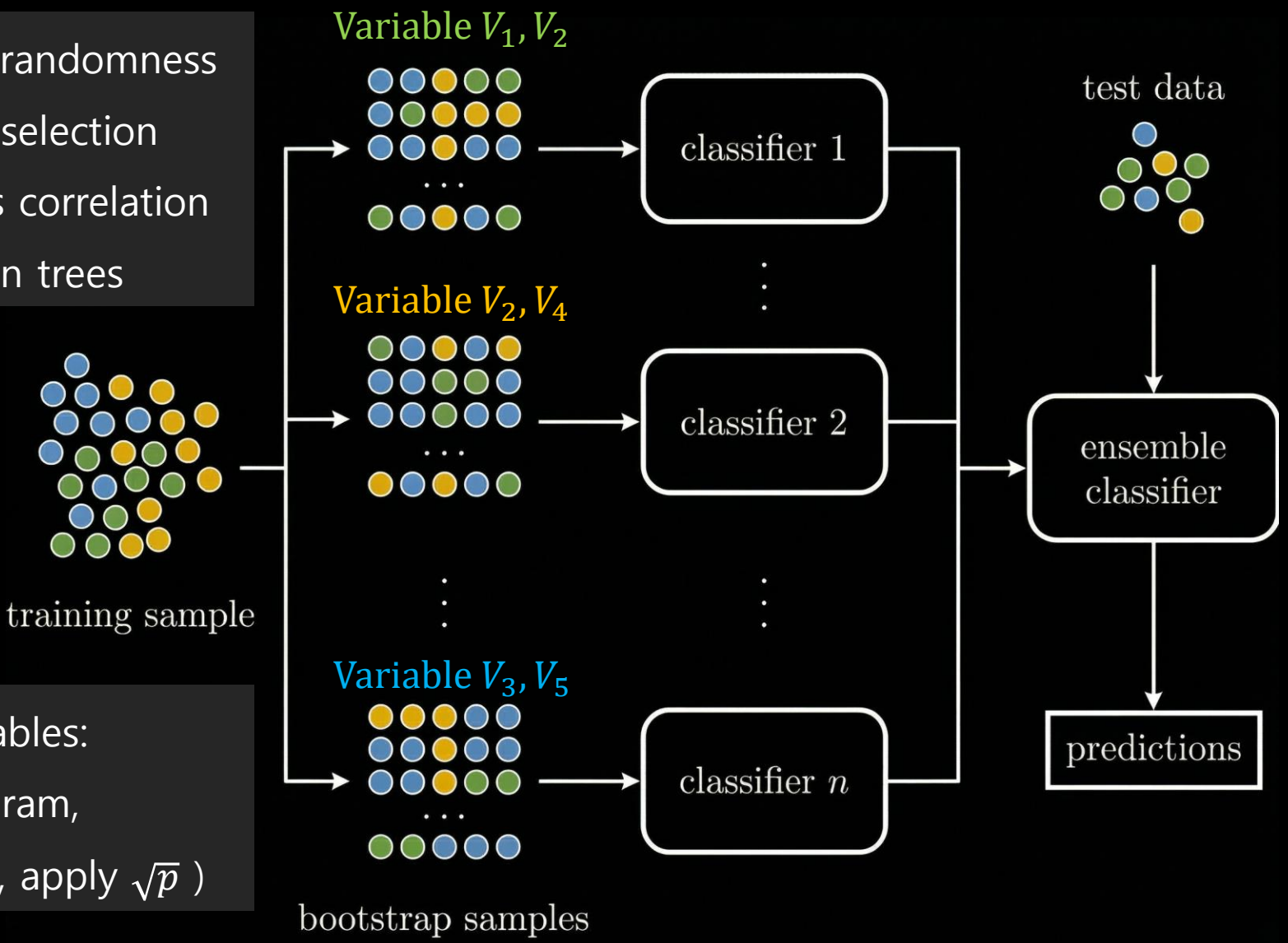
$$Var(X,Y) = Var(X) + Var(Y)Cov(X,Y)$$

Sampling with replacement → duplicates are inevitable

- Base learners are not guaranteed to be independent
- No guarantee that $Cov(X,Y) = 0$

Random Selection of Variables

Introduce randomness
in feature selection
→ reduces correlation
between trees



# of variables:
Hyper param,
generally, apply $\sqrt{p}$)

Exercise with Codes

Exercise:

1. RandomForest

Thank you!