Unit 1

Bagging & Random Forest

11주차. 앙상블(Ensemble)

앙상블
(Ensemble)

» 여러 알고리즘 또는 모형을 평균화하여 성능을 향상

"성능을 증명 받아 금융 AI에서 많이 사용되는 알고리즘"

학습 내용

- 앙상블 서론
- Bagging
- Random Forest

학습 목표

- 앙상블에 대해 알 수 있다.
- 앙상블의 알고리즘을 이해하고 실제 적용할 수 있다.

앙상블 서론

☑ 알고리즘 성능 개관

? 다른 모든 알고리즘을 능가하는 절대 알고리즘이 있을까 ✕

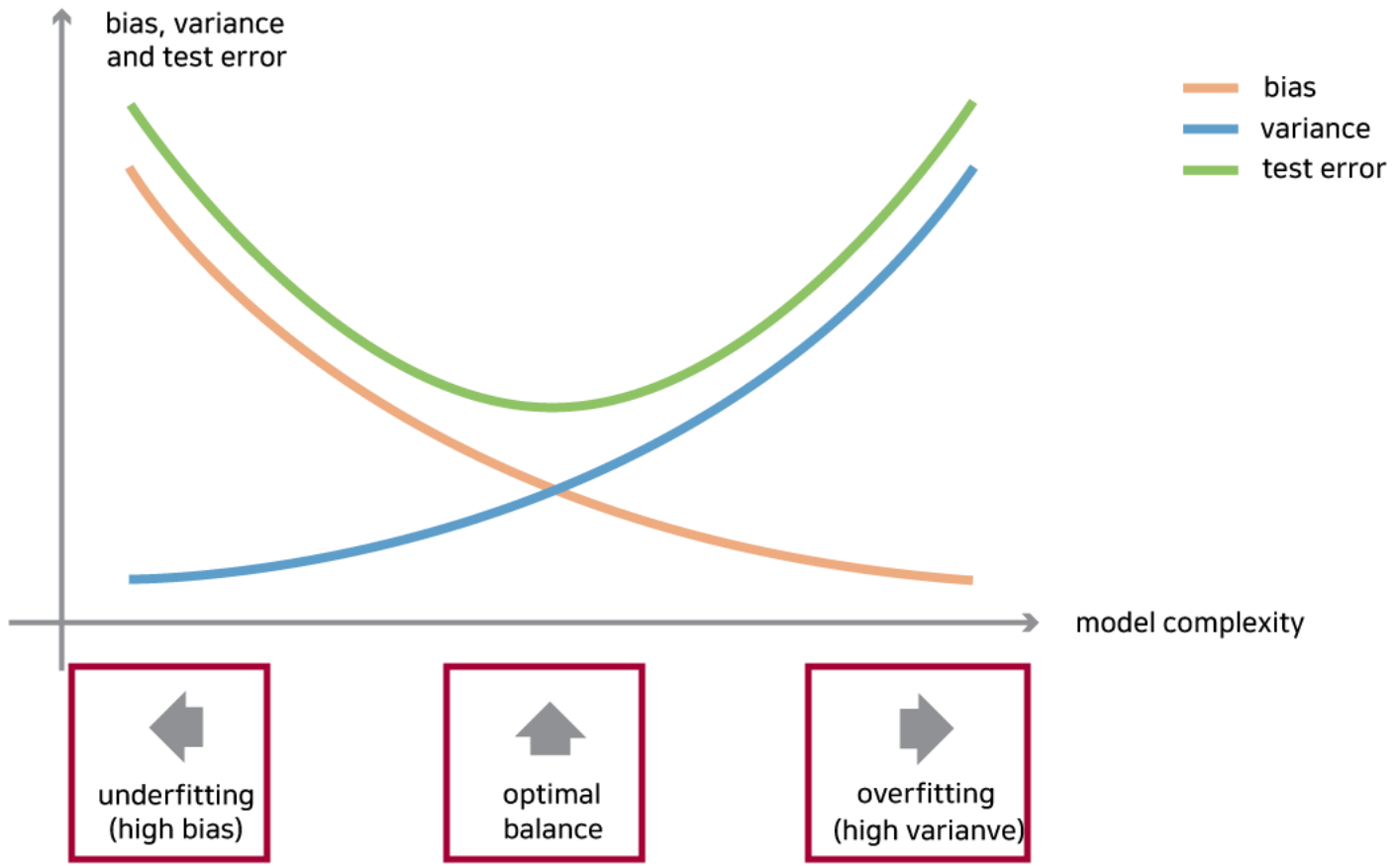
"No free lunch theorems for optimization"

- 다양한 문제를 풀어나갈 때 평균적으로 보면 거의 동일

앙상블 서론

☑ Bias-variance dilemma(Trade-off)

절대 알고리즘은 없지만
부분적으로 성능을
개선할 수 있는
알고리즘 존재



➤ 적절하게 Variance와 Bias를 줄일 수 있는 Feature의 수를 결정

앙상블 서론

Bias-variance dilemma(Trade-off)

Error

- Variance
- Bias
- Irreducible Error

- 백색잡음
- 절대 줄일 수 없는 에러
- 완벽하게 설명되고 남은 부분

완벽한 모형이 없기 때문에 대부분 백색잡음이 아니다!

앙상블 서론

Bias-variance dilemma(Trade-off)

Variance

- 추정 값들의 흩어진 정도
- 추정 값의 평균과 추정값의 차이에 대한 것
- Parameter Estimate

Bias

- 추정값의 평균과 참 값들의 차이
- 참 값과 추정 값의 거리를 의미

앙상블 서론

☑ Bias-variance dilemma(Trade-off)

분산 감소 (Variance Reduction)

» Training set이 완전히 독립적인 경우
앙상블을 평균화하는 것이
Bias에 영향을 주지 않고 분산을 감소시킴

편향 감소 (Bias Reduction)

» 단순모형보다 단순모형들의 평균이 훨씬 더 큰
Capacity를 갖음으로써 Bias 감소시킴

» 모형 평균화 방식 : Boosting에서의 기법

혼성 모델(Hybrid Model)

- 여러 가지 알고리즘을 결합하는 모델
- 가장 좋은 단일 알고리즘보다 결과가 좋음

앙상블 서론

☑ 혼성 모델

혼성 모델의 도입 배경

- 다른 모든 방법을 능가하는 보편적으로 우수한 알고리즘의 존재에 대한 의문
- 특정 문제를 가장 높은 성능으로 풀 수 있는 알고리즘에 대한 필요성

앙상블 서론

Resampling

Resampling

데이터의 양이 충분치 않을 때,
같은 샘플을 여러 번 사용하는 것

- 통계적 신뢰도 향상
- Bias-variance tradeoff를 통해 샘플 수가 충분히 클 경우 분산 감소
- 여러 모델을 활용하면 Bias 감소
- 분산이 줄어들면 MSE도 감소

이러한 필요성과 장점으로 인해 혼성 모델이 활용

앙상블 서론

Weak Learner

Weak Learner(약한 학습자)

- Decision stumps
- Shallow decision trees
- Naïve Bayes
- Logistic regression

→ 0과 1로 분류하는 단순한 학습자

앙상블 서론

Weak Learner

약한 학습자 = Base Model

→ Bias와 분산이 매우 큰 경향

Weak learner의 결합 → Strong Learner (Ensemble Model) → Bias / Variance 감소

➤ Weak Learner의 Gathering

앙상블 서론

☑ Weak Learner

Weak Learner 결합 방식

1 Bagging

- 동종의 weak learner를 독립적으로 학습, 평균화
- 분산 축소 효과

2 Boosting

- 동종의 weak learner를 적응적 방식으로 순차적 학습
- 확정적인 전략에 따라 결합
- Bias 축소

앙상블 서론

☑ Ensemble(앙상블)

Ensemble(앙상블)

- 같은 문제에 대해 서로 다른 여러 알고리즘이 해를 구하고, 결합 알고리즘이 그들을 결합하여 최종 해를 만드는 방식
- 유사한 여러 문제들에 대해 하나의 알고리즘이 해를 구하고, 결합 알고리즘이 그들을 결합하여 최종 해를 만드는 방식

앙상블 서론

☑ Ensemble(앙상블)

혼성 모델의 필요성

Weak learner : 간단하지만 오류가 많음 → **발육부진 분류기 생성** → **혼성 모델 사용**

- Noisy 또는 Bushy
- Weak 또는 Stunted

앙상블 서론

- [x] Ensemble(앙상블)

혼성 모델의 선호도

Boosting Approach > Bagging Approach > Single Tree (weak learner)

앙상블 서론

☑ 분류기 앙상블 시스템

분류기 앙상블 시스템

➡ **Resampling을 통해 생성된 샘플 집합들을 이용하여 분류기를 각각 훈련**

Random Forest

- 특징 벡터의 부분 공간을 이용하여,
  샘플 부분 집합을 생성 후 분류기를 훈련

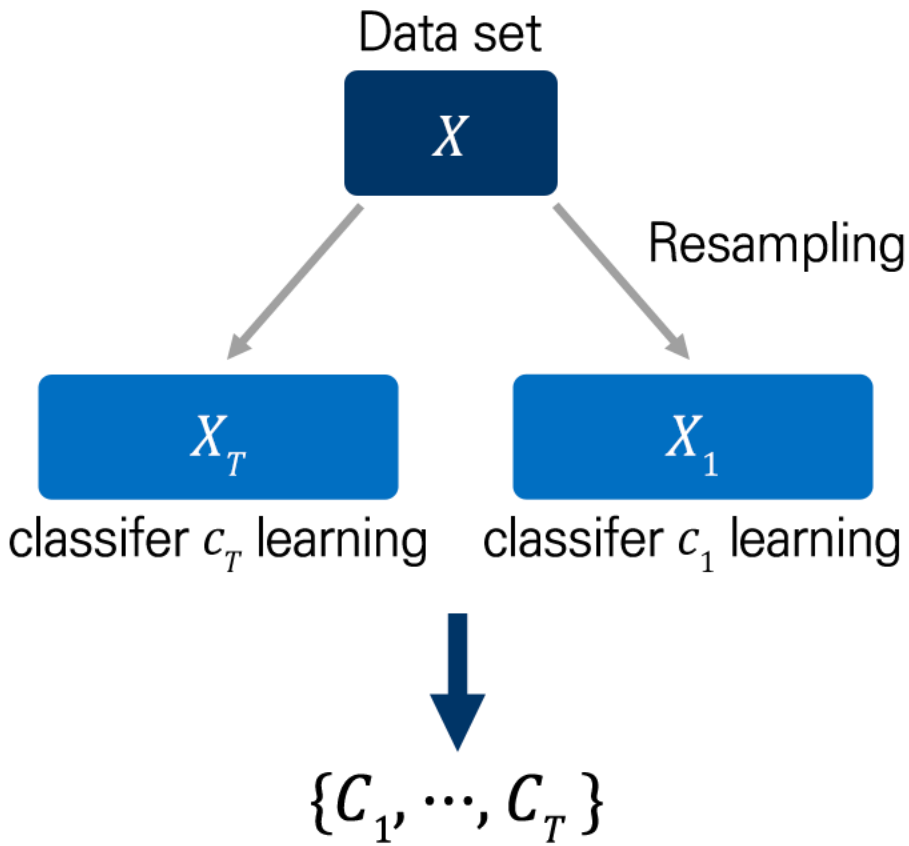
앙상블 서론

☑ 분류기 앙상블 시스템

앙상블을 구성하는 분류기

1. 요소 분류기(Component Classifier)
2. 기초 학습기(Base Learner)

요소 분류기들의 출력을 결합해서
하나의 분류 결과를 만드는 과정



Bagging

☑ Bagging 소개

Bagging(Breiman,1996)

➡ Bootstrap Aggregating

» Bootstrapping된 샘플 집합에서 훈련 후,
입력 값에 대하여 분류기들의 평균값
또는 다수결 투표를 취함

Bootstrap
(반복적인 복원 추출) ➡ **Aggregation**
(결과를 모두 종합)

Bagging

☑ Bagging 효과

Bagging의 **효과**

1. 분산이 높은 분류기에 사용되면 이를 축소시켜주는 효과
2. 트리 분류기와 같이 불안정성을 보이는 분류기에 큰 효과
3. 훈련 집합이 달라지면 차이가 큰 트리 생성 → 다양성 확보
4. Bias를 변화시키지 않고 variance를 감소

Bagging

Bagging 알고리즘

- 입력 : 훈련 집합 $X=\{(X_1, t_1),(X_2),\cdots,(X_N, t_N)\}$ 샘플링 비율 $\rho(0\leq\rho\leq1)$
- 출력 : 분류기 앙상블(Classifier Ensemble)

$t = 0, \quad C = \emptyset$
Repeat{
$t = t + 1$
X에서 임의로 ρN개의 샘플을 뽑아 X_t 라 한다.(with replacement)
X_t 로 분류기 $\boldsymbol{c}_i$ 를 학습한다.
$C = C \cup \{c_i\}$, $\boldsymbol{c}_i$ 는 서로 독립
} until (멈춤조건)
Return C

샘플이 반복되면서 여러 번 뽑히거나 안 뽑히는 샘플이 있을 수도 있음

Repetition이 중단되고 최종적으로 분류기에 대해서 Majority 룰에 의해 선택

Bagging

☑ Bagging 활용

Bagging이 도움이 되는 경우

1. Over-fitted Base Model을 사용할 때
2. Training Data에 높은 의존성을 가진 모델을 사용할 때

Bagging

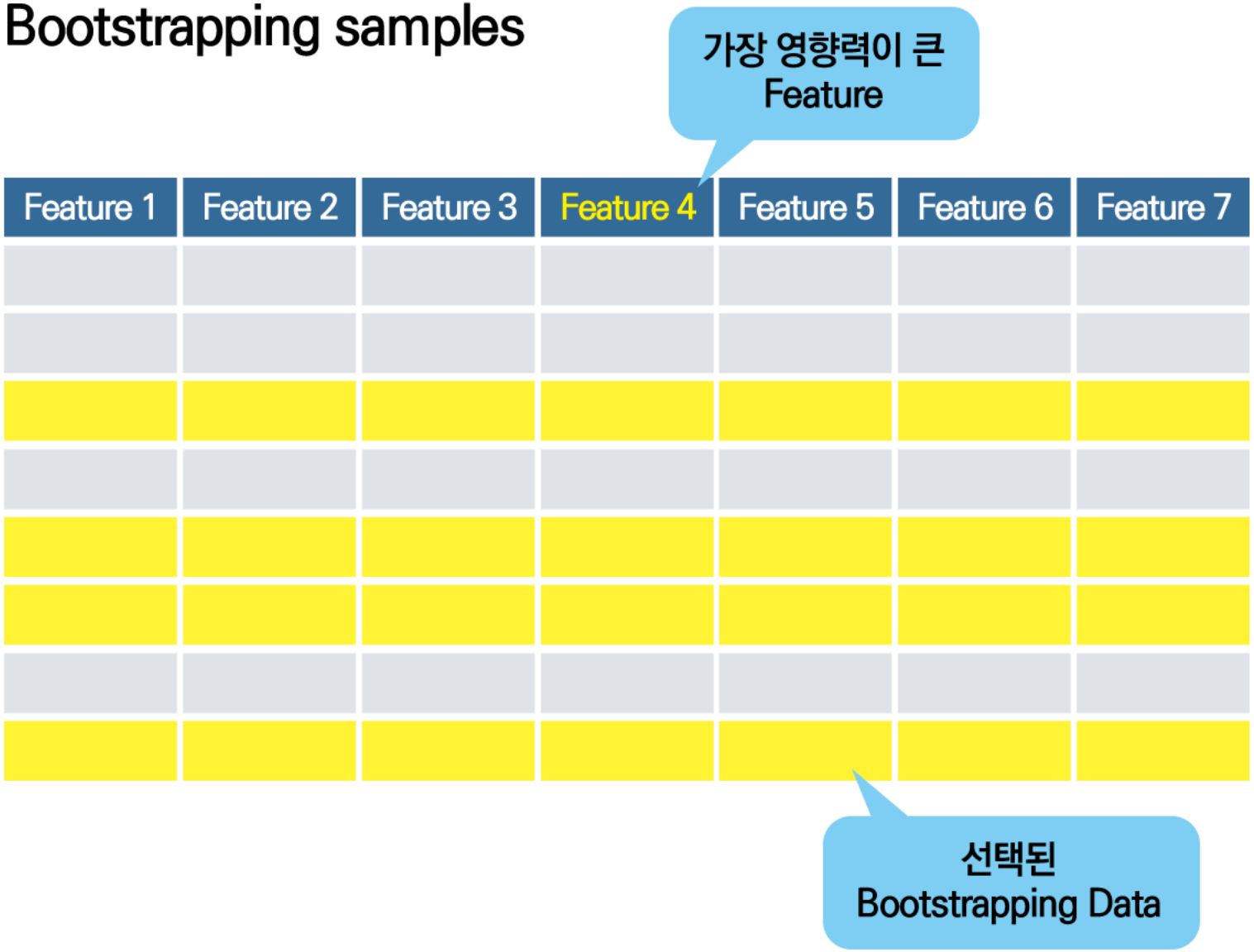
☑ Bagging 활용

Bagging이 **도움이 안되는 경우**

1. High Bias Base Model을 사용할 때
2. Base model이 Training Data를 변경하는 것에 대해 Robust 할 때

Bagging

☑ Bootstrapping samples



- Observation에 대해 샘플링
- 반복해서 샘플링이 되거나,
  한번도 샘플링 되지 않을 수 있음
- 모든 Feature들이 동시에 샘플링

Bagging

☑ Bootstrapping samples의 단점

Bootstrapping samples의 단점

1. 열을 모두 선택하고, 행을 랜덤하게 선택
2. 열을 모두 선택하게 되면,
   대다수 Tree의 결과가 유사
3. Tree 간의 상관관계가 형성되어
   분산 감소 효과에 부정적

Bagging

☑ Bootstrapping samples의 단점

독립인 경우

$x_1, \cdots, x_n$ iid : $E[X] = \mu$, $Var[X] = \sigma^2$ ⬅ $var(x_i) = \sigma^2$; $Var\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{\sigma^2}{n}$

➤ n이 커지면 분산이 줄어듦

상관관계가 있는 경우

$$corr(x_i, x_j) = \rho$$

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

➤ n이 커지면 $\rho\sigma^2$ 이 커져 분산감소효과가 사라지고, 오히려 분산이 커짐

Random Forest

Random Forest 소개

Random Forest (Brieman, 2000)

Randomly Selected Feature

- 일정 Feature에 대해서 랜덤하게 선택
- Tree 모델에 Bagging과 Subspace Sampling을 적용

Random Forest

☑ Random Forest 활용

? 최적 Feature의 개수는

전체 변수가 p개 이고, p=m이면

Bagging에서
Bootstrapping 하는 것과 동일

Random Forest

☑ Random Forest 활용

? 최적 Tree의 개수는

OOB(Out-of-Bag) observations의
Estimate of Test Error를 계산

부스트랩 샘플을 이용해 개별 학습기 학습

↓

OOB에 속하는 샘플들에 대해 적용 → 에러 도출

↓

에러가 안정적인 값을 보일 때 Tree의 최적 개수 결정

Random Forest

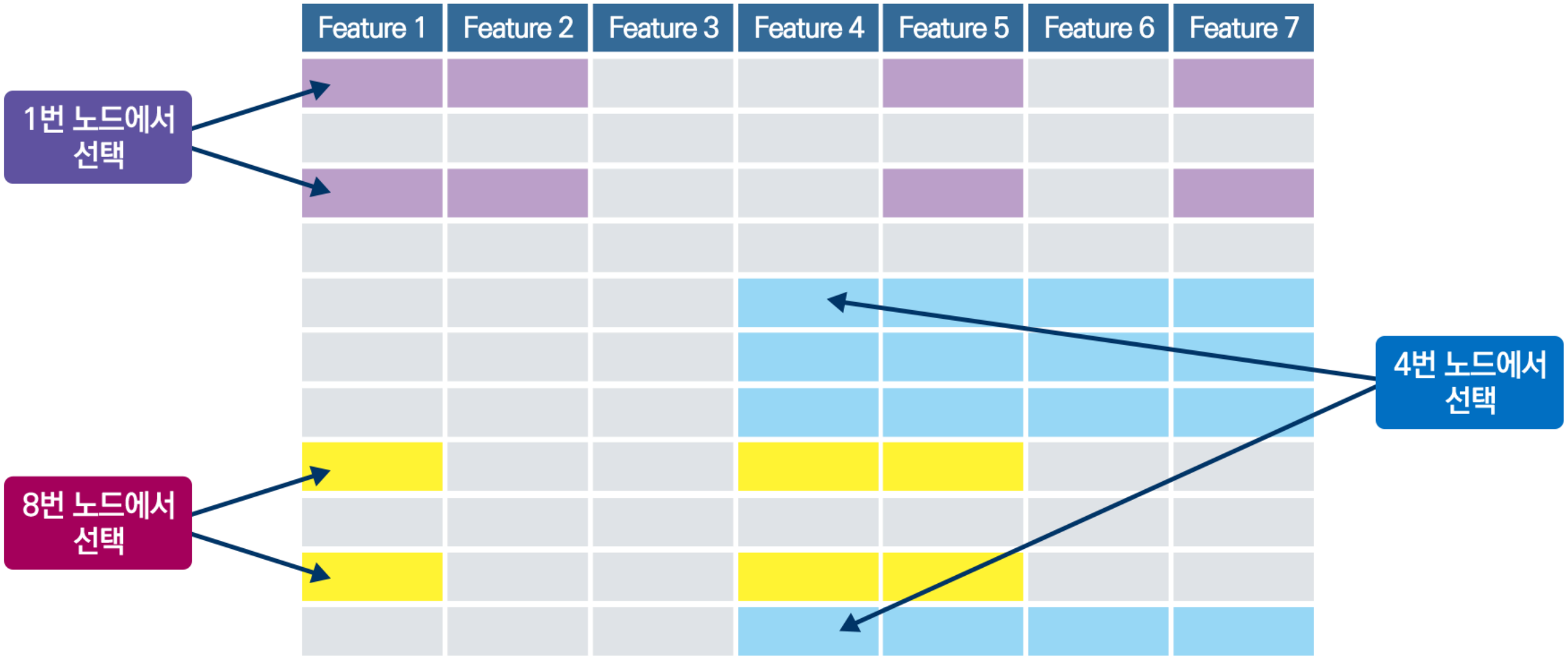
☑ Random Forest 역할

Random Forest의 역할

- Tree 간의 Correlation 축소
- 평균을 구할 때 분산을 줄여주는 역할
- 결과값들의 평균을 통해 Regression 예측
- Classification은 Majority Vote

Random Forest

☑ Training Set Sampling



Random Forest

Random Forest 알고리즘

- 입력 : 훈련 집합 $X=\{(\mathbb{X}_1, t_1),(\mathbb{X}_2, t_2),\cdots,(\mathbb{X}_N, t_N)\}$,, 샘플링 비율 p $(0<p\leq 1)$)
- 출력 : 분류기 앙상블(Classifier Ensemble)

$t = 0, \quad C = \emptyset$

Repeat{

$t = t + 1$

X에서 임의로 pN개의 샘플을 뽑아 X_t 라 한다.(replacement)로 분류기 c_i 를 학습한다.

노드에서 split할 때, sample의 p feature 중 random하게 m feature를

선택된 feature들로 생성된 부분 공간 샘플을 후보로 하여 학습한다.

$C = C \cup \{c_i\}$

} until (멈춤조건)

Return C

Random Forest

☑ Random Forest 알고리즘

" 일부 Feature만 샘플링하는
Random Forest "