복잡한 기계 학습

임 경 덕

다양한 데이터 분석

- ❖ 데이터 요약(aggregation)
- ❖ 검정(test)
- ❖ 모의실험(simulation)
  - 확률 분포 등을 활용하여 관심 대상에 대한 기댓값을 확률적으로 예측

- ❖ 데이터 기반 의사결정(data-driven decision)
  - ▪ 규칙기반 시스템(rule-based system)
    - 확인된 사실, 가설, 분석 결과를 바탕으로 조건이나 규칙을 설정
    - 금융 상품 가입 심사 등 정보가 제한적일 때 주로 활용
  - ▪ 기계학습 알고리즘(machine learning algorithm)
    - 다양한 변수의 관계 속에서 의미 있는 정보와 패턴을 파악, 활용
    - 신용등급 등 활용가능한 정보가 많을 때 활용

기계 학습의 이해

❖ 기계 학습(machine learning)

주어진 문제를 해결하기 위해 알고리즘과 기계의 저장, 연산 능력을 활용
데이터 속에 존재하는 다양하고 복잡한 패턴을 컴퓨터로 파악하는 과정

- 데이터 속 차이를 확인하고 설명

❖ 지도학습(supervised learning) : 관심변수 y의 차이를 설명

- 회귀(regression) : 연속형(continuous) 관심변수 y (1, 100의 차이)
- 분류/판별(classification) : 범주형(categorical) 관심변수 y (0, 1의 차이)

❖ 비지도학습(unsupervised learning) : 관측치 간 차이를 일반화

- 군집화(clustering) : 가까운 관측치들끼리 묶어 군집화

지도학습과 변수의 구분

❖ 설명변수와 반응변수

변수의 관계를 설명할 때 영향을 주는 변수와 영향을 받는 변수를 구분

- **설명변수(explanatory variable)**
  - 독립변수(independent variable)
  - 결과에 영향을 줄 수 있는 변수

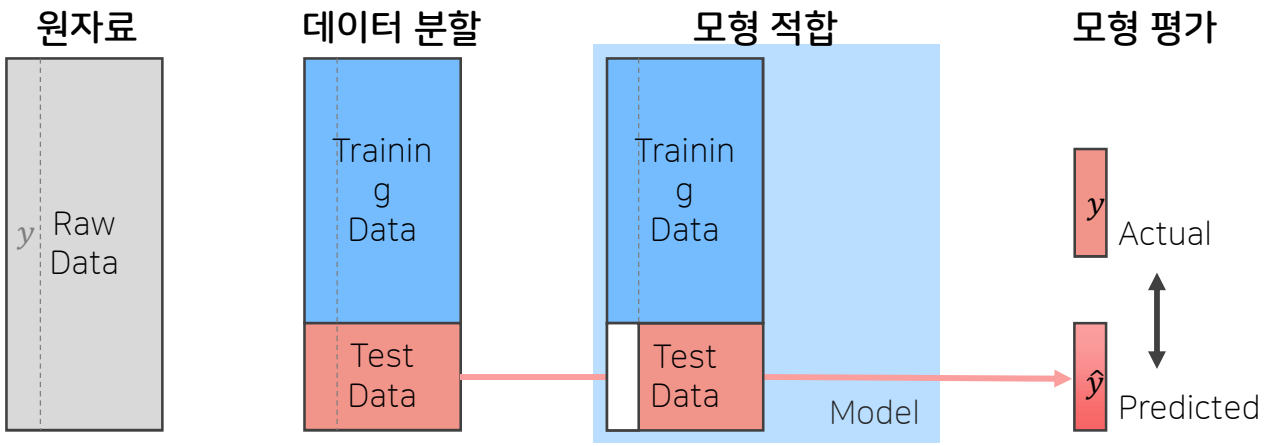
- **관심변수**
  - 종속변수(dependent variable), 반응변수(response variable)
  - 설명 변수의 변화에 따라 값이 결정되는 변수

❖ 모형 적합 : 설명변수로 관심변수의 차이를 설명하는 과정

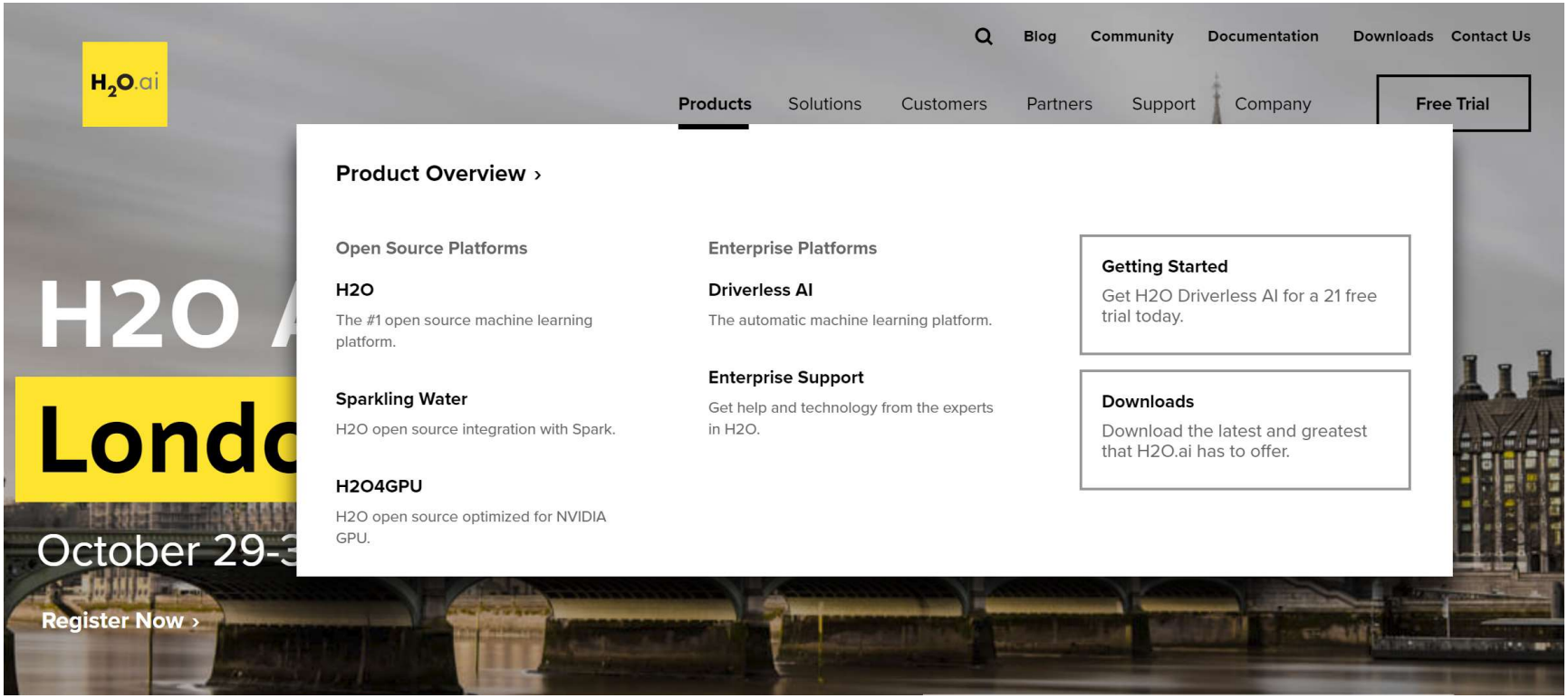
교차 검증을 통한 모형 평가

❖ 데이터 분할을 활용한 모형 평가 과정

1. Training(훈련) 데이터와 Test(검증) 데이터 분할
2. Training 데이터를 활용한 모형 적합(fitting)
3. 적합된 분류모형을 Test 데이터를 활용해서 평가



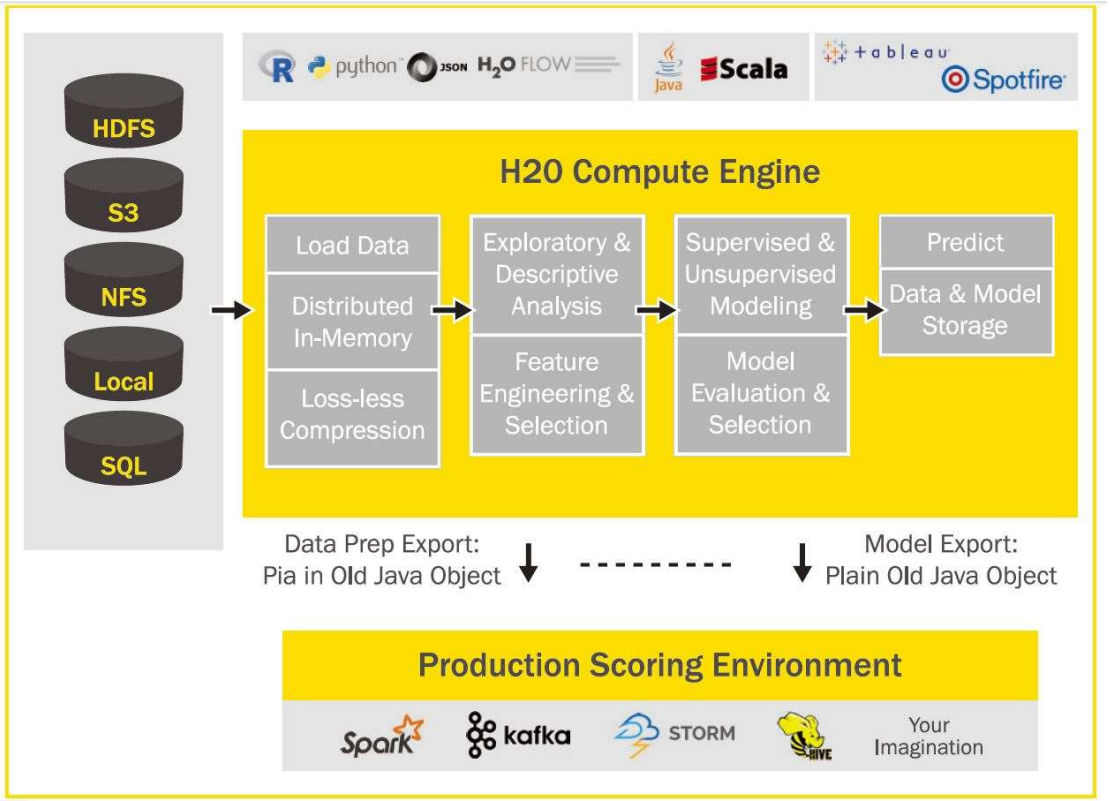
H2O를 활용한 앙상블 모형



“#1 open source machine learning platform”

- h2o
  - 기계 학습 방법론 중심의 알고리즘 솔루션
  - Spark 활용 분산처리, GPU 활용 병렬처리 등의 기술과 접목
  - Driverless AI 등 비즈니스 수요 창출

- h2o 프레임워크



앙상블의 개념

❖ 앙상블(Ensemble)

일반적으로 판별의 문제에서 복수의 모형을 활용하는 기법

복수 모형의 평균 혹은 다수결로 예측에 활용

- Bagging : 관측치/변수를 랜덤해서 선택하는 붓스트랩(Bootstrap)을 활용(복수 모형의 다수결로 판별)
- Boosting : 앞 모형의 오차를 줄이는 보조모형을 계속해서 추가 (점진적 모형 개선)

❖ 앙상블 모형의 장점과 단점

- 단점 : 연산량의 증가
- 장점 : 복수 모형 활용을 통한 모형 안정화

의사결정 나무의 앙상블

❖ 의사결정 나무 모형의 특성

- 장점 : 교호작용 등 비선형(Nonlinear) 관계를 비교적 잘 설명
- 단점 : 모형의 불안정성, 낮은 정확도(Accuracy)
  - 새로운 관측치, 변수의 유입 및 유출에 따른 모형의 변동이 큼

앙상블 기법을 활용하여 여러 그루의 나무를 활용하여 보완

- Random Forest : 의사결정 나무의 bagging
- Gradient Boosting Machine : 의사결정 나무의 boosting

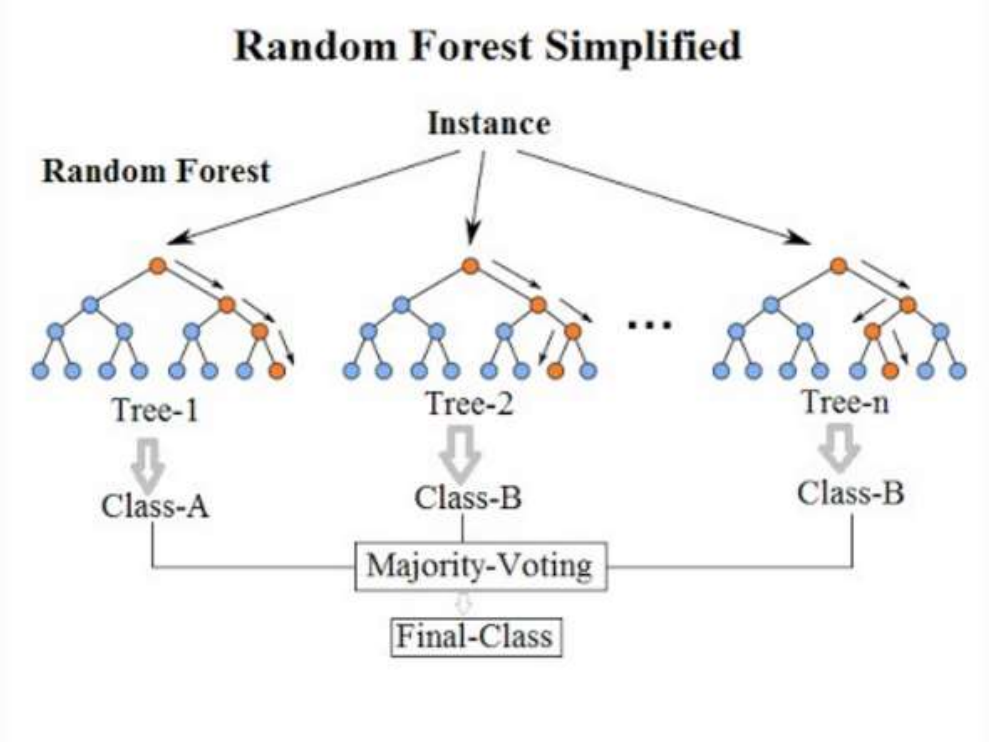
랜덤 포레스트

❖ 랜덤 포레스트(random forest)

Bagging을 활용한 의사결정 나무의 앙상블 기법

관측치/변수를 랜덤으로 선택한 데이터로 하나의 모형 생성

복수의 모형의 평균 혹은 다수결을 활용하여 예측



GBM

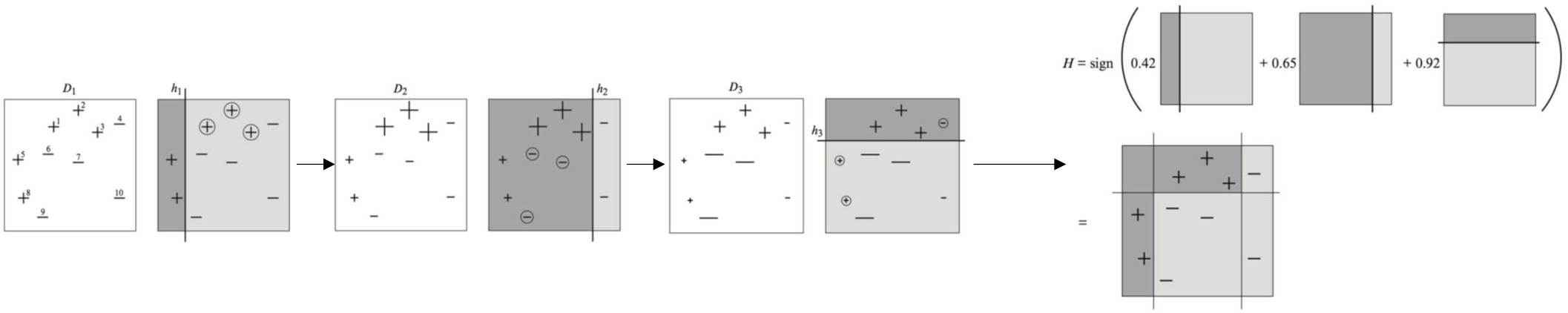
❖ GBM(Gradient Boosting Machine)

의사결정나무의 Boosting

오분류되거나 오차가 큰 관측치에 가중치를 두고 보조 모형 재적합

정확도가 높지만 과적합 발생 가능성도 높음

❖ 알고리즘 도식화 예제



h2o를 활용한 모형 적합

- **h2o 설치**

  java 기반의 어플리케이션으로 R과 별도로 구동
  R에서 'h2o' 패키지 설치를 통해 구동 가능

- **데이터의 형태 변환**

  as.h2o( )를 활용해 데이터를 h2o 객체로 변환

- **모형 적합**

  변환된 데이터를 활용하여 모형 적합
  h2o.randomForest( ) : RF 모형 적합 / h2o.gbm( ) : GBM 모형 적합

- **모형 최적화 및 성능 평가**

  옵션을 활용한 각 모형 모수(parameter) 변경 가능
  h2o.grid( ) : 각 모형의 모수 조합에 대한 최적 모수 조합 계산
  모형 적합 시 오차행렬, 오분류율 등 주요 지표 자동 계산

(참고) 모형의 비교

❖ 지표를 활용한 모형 비교

데이터에 다양한 알고리즘을 적용하고 다양한 모형 적합 가능

비교 지표 등을 통해 모형 간 비교 후 최적 모형 선택

❖ 연속형 관심변수의 비교지표

- 실제 값과 예측 값의 차이를 기반으로 계산

❖ 범주형 관심변수의 비교지표

- 실제 범주와 예측 범주의 오차 행렬을 기반으로 계산

연속형 관심변수의 평가 지표(오차의 절댓값)

❖ 평균 절댓값 오차(MAE ; mean absolute error)

각 관측치의 오차(실제 값과 예측 값 차이)의 절댓값에 대한 평균

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

❖ 평균 백분위 절대값 오차(MAPE ; … percentile error)

각 관측치의 실제 값 대비 오차의 절댓값에 대한 평균

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

연속형 관심변수의 평가 지표(오차의 제곱)

❖ 평균 제곱 오차(MSE ; mean squared error)

각 관측치의 오차의 제곱에 대한 평균

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

❖ 평균 제곱근 오차(RMSE ; root MSE)

각 관측치의 실제 값과 예측 값 차이의 제곱에 대한 평균의 절대값

$$\sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

범주형 관심변수의 평가 지표 (예측 확률)

❖ Log Loss(logarithmic loss)

실제 값과 예측 확률을 활용한 지표

- 계산 : $\text{logloss} = -\{y\log(p) + (1-y)\log(1-p)\}$

  y : 실제 값(0 또는 1)

  p : 관측치에 대한 예측 확률 $(0 \le P(Y=1) \le 1)$

<u>값이 작을 수록 더 좋은 모형을 의미</u>

실제 1에 대해 확률을 1로 예측한 경우

- $\text{logloss} = 0 : (1-y) = \log(p) = 0$

실제 0에 대해 확률을 0로 예측한 경우

- $\text{Logloss} = 0 : y = \log(1-p) = 0$

실제 1에 대해 확률을 0에 가깝게 예측한 경우

- $\text{logloss} \uparrow : -\log(p) \to -\infty$

실제 0에 대해 확률을 1에 가깝게 예측한 경우

- $\text{logloss} \uparrow : -\log(1-p) \to -\infty$

범주형 관심변수의 평가 지표 (오차 행렬)

❖ 오차 행렬(confusion matrix)

실제 값과 예측 값에 따른 경우의 수를 분류한 표

각종 지표 계산에 활용

		Actual	
		Positive(+)	Negative(−)
Predicted	Positive(+)	True positive (a)	False positive (c, Type I error)
	Negative(−)	False negative (b, Type II Error)	True negative (d)

- Negative : 음성, 가설 검정의 H_0에 해당
- Positive : 양성, 가설 검정의 H_1에 해당
- True : 적중
- False : 오분류

범주형 관심변수의 주요 평가 지표

❖ 정확도(accuracy)

- 오차행렬을 활용한 계산 : $\frac{a+d}{a+b+c+d}$

  전체 중에서 적중한 것의 비중

❖ 오분류율(error rate)

- 오차행렬을 활용한 계산 : $\frac{b+c}{a+b+c+d}$

  전체 중에서 오분류한 것의 비중

  오분류율 = 1 – 정확도

범주형 관심변수의 주요 평가 지표(2)

❖ 민감도(sensitivity, recall, true positive rate)

- 오차행렬을 활용한 계산 : $\frac{a}{a+b}$

  감염자의 검사결과가 양성일 확률, 문제가 있는 사람을 잘 찾아낼 확률

❖ Precision

- 오차행렬을 활용한 계산 : $\frac{a}{a+c}$

  양성으로 판단한 사람 중 실제 감염자의 비중

❖ F1 score

- 정의 : $\left(\frac{recall^{-1}+precision^{-1}}{2}\right)^{-1} = 2\frac{precision \cdot recall}{precision+recall}$

  precision과 recall의 조화 평균

  감염자를 잘 찾아내면서 동시에 많은 사람을 양성으로 예측하는 것을 경계

감사합니다.