Machine Learning

- 데이터 학습
- 목표 : 일반화
- 대표예시 : 스팸 필터링

Learning from Experience

1. Supervised Learning

- Labeled된 입력과 출력에서 학습
- 정답 존재

2. Unsupervised Learning

- 패턴 존재

3. Semi-Supervised Learning

- Supervised, Unsupervised Learning 모두 사용
- 예시 : Reinforcement Learning

4. Keyword

- 출력 = Response Variable
- 입력 = Features

- Training set - Supervised를 위함
- Test set - Performance 평가를 위함

Machine Learning Tasks

1. Supervised Machine Learning

1. Classification
   - 이산 값 예측

2. Regression
   - 연속 반응 변수 예측

2. Unsupervised Machine Learning

1. Cluster
   - 유사한 그룹끼리 묶을 때

2. Dimensionality Reduction
   - 반응 변수에 큰 영향을 주는 설명 변수를 찾는 것
   - 설명 변수가 천개 이상 될 경우

Training Data and Test Data

1. Test set

- Performance 평가
- Training set의 자료는 포함되지 않음

2. Validation

- 3번째 관측 값(Training, Test set)
- hyperparameter를 조정 할 때 필요

3. Common Allocation

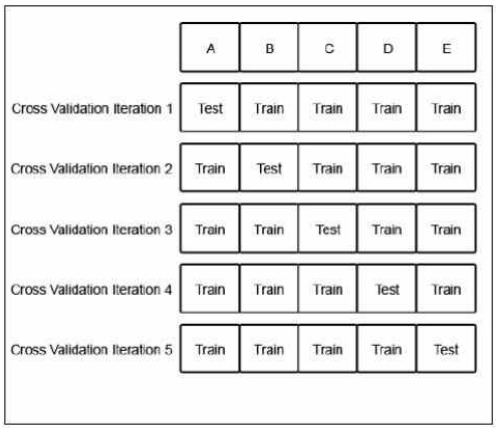
- 50% : Training set
- 25% : Test set
- 나머지 : Validation set

4. Cross-Validation

- 같은 데이터에서 Train과 Algorithm 적용
- 기존의 방법보다 더 정확함

Cross-Validation Example

- 5개의 동일한 크기의 subset으로 분할
- A, B, C, D, E로 Label

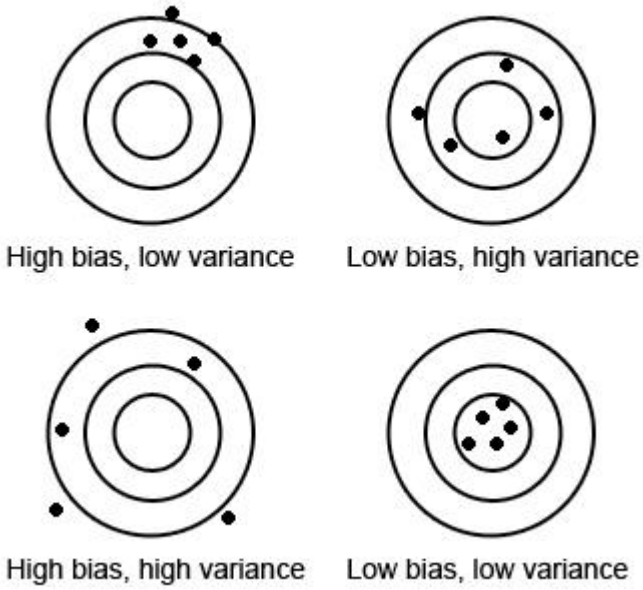


Performance Measures, Bias, and Variance

1. Supervised Learning

Bias, Variance – Fundamental cause of prediction error

- 예측 오류의 근본적인 이유



Bias-Variance trade-off

- 한쪽을 줄이면 다른 한쪽이 올라감

2. Unsupervised Learning

- 'Error Signal'을 측정할 척도가 따로 없음
- Performance Metrics – 데이터 구조의 특성을 측정함

Performance Measures

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

1. Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- 정확도
- 전체 중에 정답

2. Precision

$$P = \frac{TP}{TP + FP}$$

- 검출 한 것 중에 정확도
- True라 분류한 것 중, 진짜 True일 확률

3. Recall

$$R = \frac{TP}{TP + FN}$$

- 검출율

Summary

- Machine Learning은 경험에서 학습하여 업무의 Performance를 높이는 것이다.
- Supervised Learning은 Labeled된 것이다.
- Unsupervised Learning은 Unlabeled된 것이다.
- Classification은 이산 반응 변수를 예측 한다.
- Regression은 연속 반응 변수를 예측 한다.
- Clustering은 비슷한 그룹끼리 묶는 것이다.
- Dimensionality Reduction은 설명 변수의 수를 줄이는 것이다.
- Bias와 Variance는 Trade-off관계이며, 가장 흔하게 사용하는 Performance Measure이다.