2. Model Building and Assessment

금융 데이터마이닝

2022 Summer

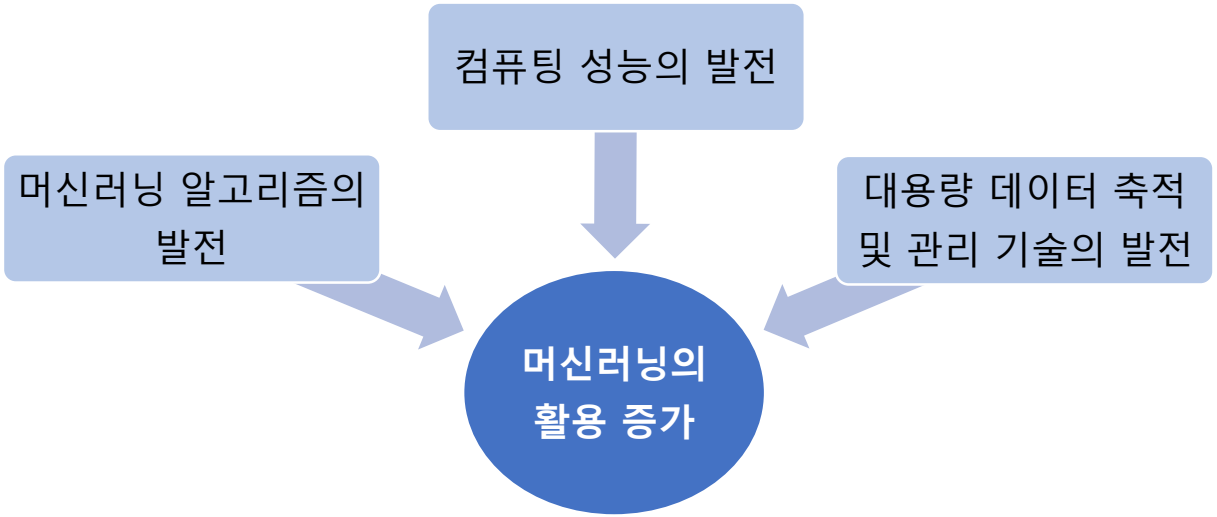
김아현

머신러닝 기법 개요 및 분류

⊙데이터 마이닝을 위한 머신러닝

- **머신러닝이란**
  - "컴퓨터 시스템에 명시적으로 프로그래밍 하지 않더라도 데이터를 스스로 학습하여 문제를 해결할 수 있게 하는 기술" – 1959년 Arthur Samuel
  - 데이터에서 패턴을 찾아내는 연산과정들의 통칭이며, 패턴을 찾아내기 위한 컴퓨터 연산을 필요로 함.
  - 사람이 인지하기 어려운 복잡한 규칙과 패턴을 파악하여 의미있는 결과를 얻을 수 있음.
  - 인공지능 기술을 위한 알고리즘에 해당.

- **머신러닝의 발전**



머신러닝 기법 개요 및 분류

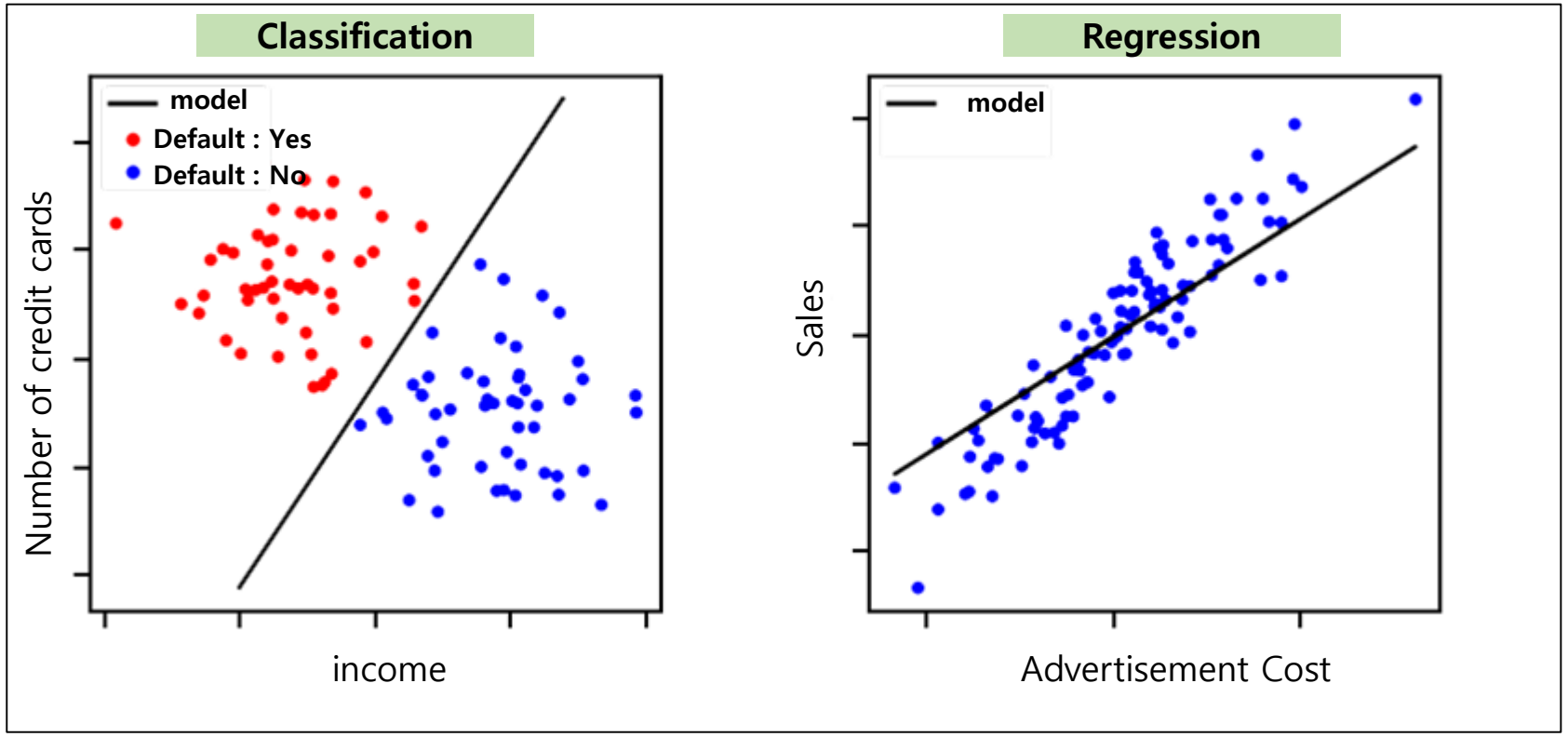
⊙머신러닝 방법론의 분류

- **지도학습 (Supervised Learning)**
  - 라벨이 있는 훈련용 데이터에서, 여러 특징변수 (feature variables)를 이용하여 목표변수 (target variable)의 라벨 (label)을 예측하도록 모델을 학습함.
  - 라벨의 데이터 타입에 따라 라벨이 연속형이면 **회귀 (regression)** 알고리즘, 라벨이 범주형이면 **분류 (classification)** 알고리즘으로 구분함.
  - 대표 알고리즘.
    - Linear Regression
    - k-nearest Neighbors
    - Logistic Regression
    - SVM
    - Decision Tree
    - Random Forest
    - Boosting

머신러닝 기법 개요 및 분류

⊙머신러닝 방법론의 분류

- **지도학습 (Supervised Learning)**
  - 분류(classification) VS 회귀(regression)



머신러닝 기법 개요 및 분류

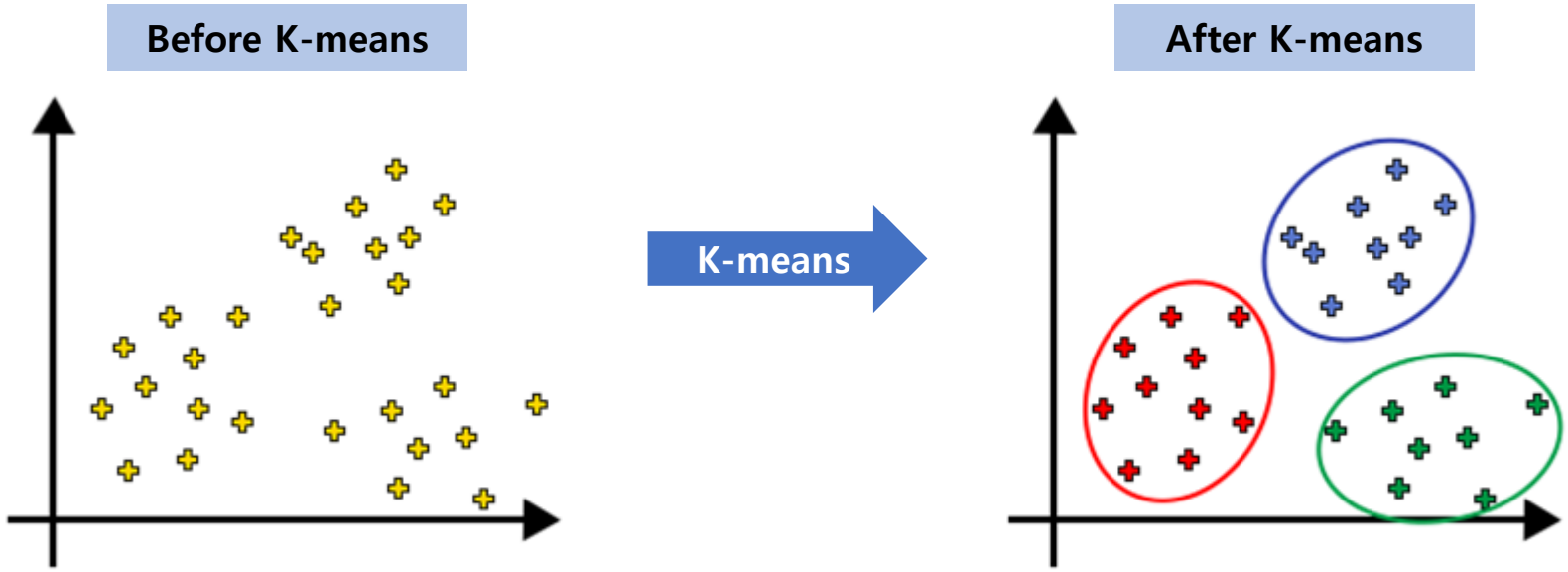
⊙머신러닝 방법론의 분류

- **비지도학습 (Unsupervised Learning)**
  - 라벨이 없는 훈련용 데이터에서 특징 변수들 간의 관계나 유사성을 기반으로 의미있는 패턴을 추출하는 학습 방법으로 자율학습 이라고도 함.
  - **군집화 (clustering)**, **차원축소 (dimension reduction)**, **추천시스템 (recommendation)** 등에 활용됨.
  - 대표 알고리즘.
    - K-means Clustering
    - Hierarchical Clustering
    - PCA
    - Apriori
    - Collaborative Filtering

머신러닝 기법 개요 및 분류

⊙머신러닝 방법론의 분류

- **비지도학습 (Unsupervised Learning)**
  - 군집화(clustering)



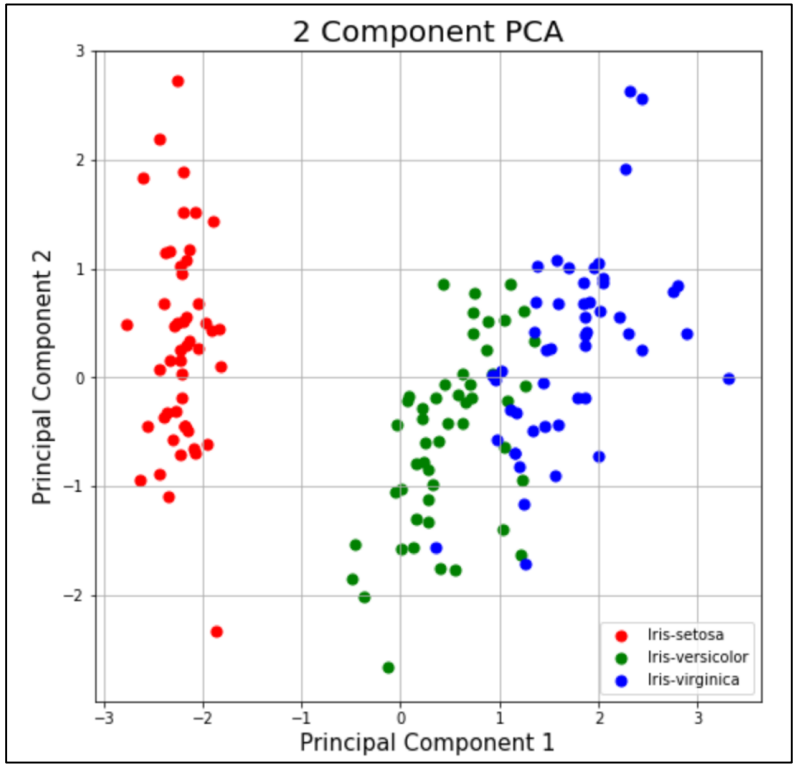
머신러닝 기법 개요 및 분류

⊙머신러닝 방법론의 분류

- **비지도학습 (Unsupervised Learning)**
  - 차원축소(dimension reduction)

	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

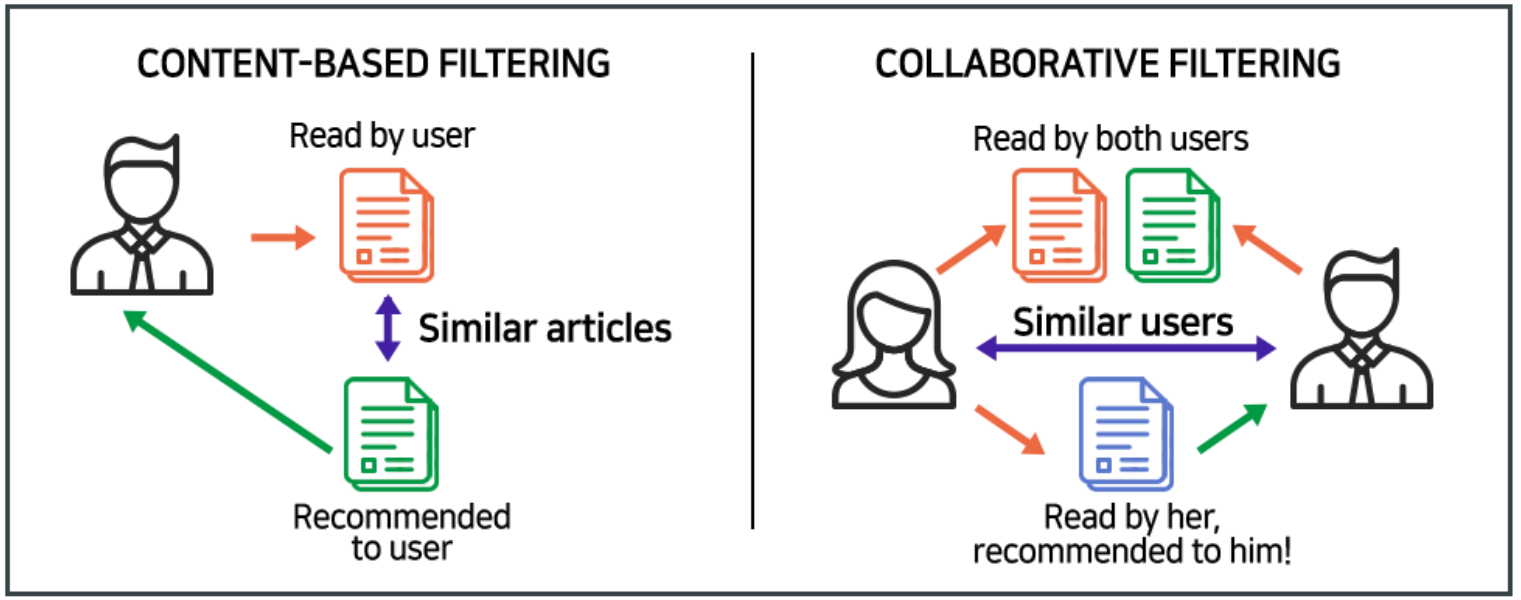
PCA



머신러닝 기법 개요 및 분류

⊙머신러닝 방법론의 분류

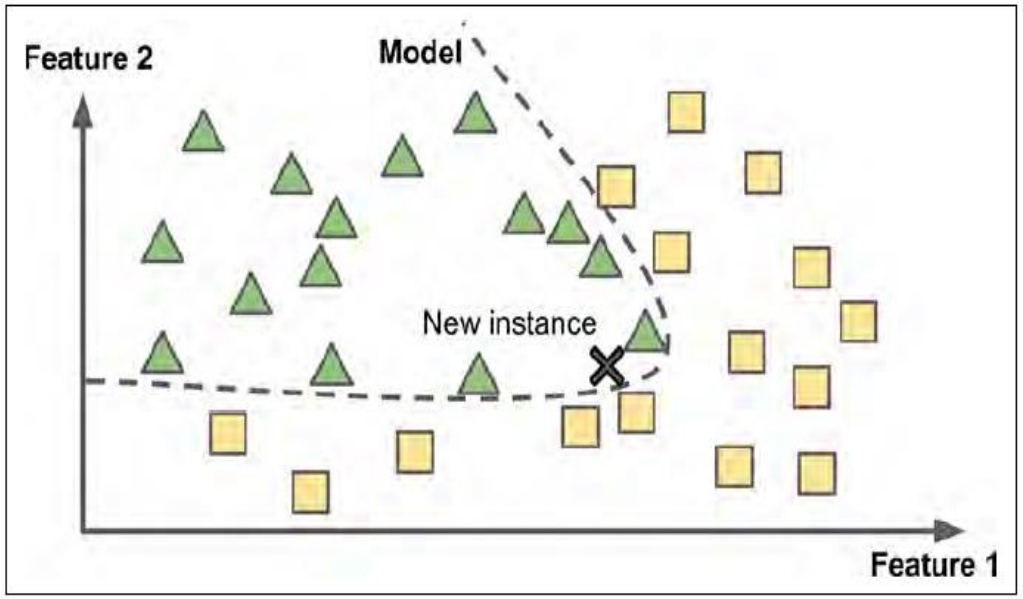
- **비지도학습 (Unsupervised Learning)**
  - 추천 시스템(recommendation)



머신러닝 모델 분석방법

⊙머신러닝 모델의 분석 절차

- **모델 기반 지도학습 알고리즘의 일반적인 분석 절차**
  - 주어진 데이터 전처리 및 탐색
  - 적절한 모델을 선택
  - 주어진 데이터로 모델을 훈련시킴
  - 훈련된 모델을 적용하여 새로운 데이터에 대한 예측을 수행

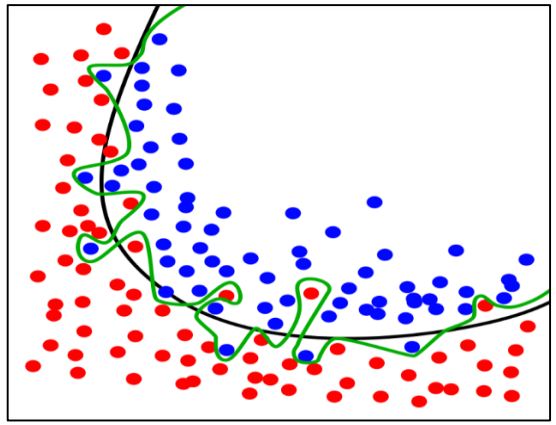


머신러닝 모델 분석방법

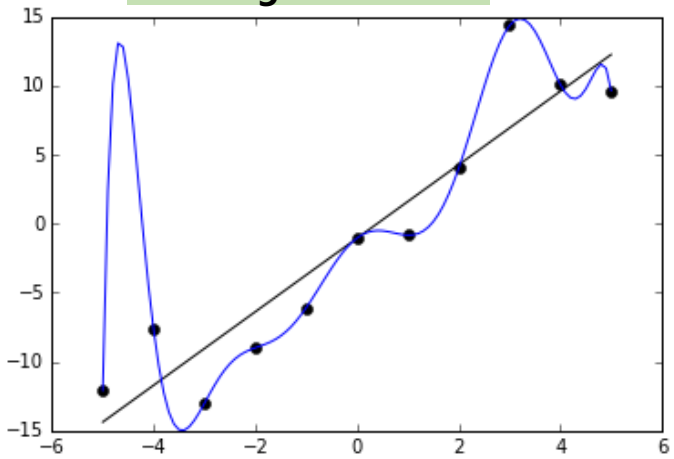
⊙머신러닝 모델의 검증 및 평가

- **과대적합(overfitting)의 문제**
  - 주어진 자료는 거의 완벽한 예측이 가능하지만, 미래의 새로운 자료에 대한 예측력이 떨어지는 문제.
  - 복잡한 알고리즘을 사용하여 데이터를 훈련하는 경우 과대적합 문제를 항상 염두에 두어야 함.

Classification



Regression



- **모델 평가의 필요성**
  - 과대적합을 막기 위해서는, 모델이 새로운 데이터에 얼마나 잘 일반화될지를 파악해야 함.
  - 모델 적합에 사용된 자료를 평가를 위해 재활용하지 않고, 평가만을 위한 데이터를 확보할 필요있음.

머신러닝 모델 분석방법

⊙머신러닝 모델의 검증 및 평가

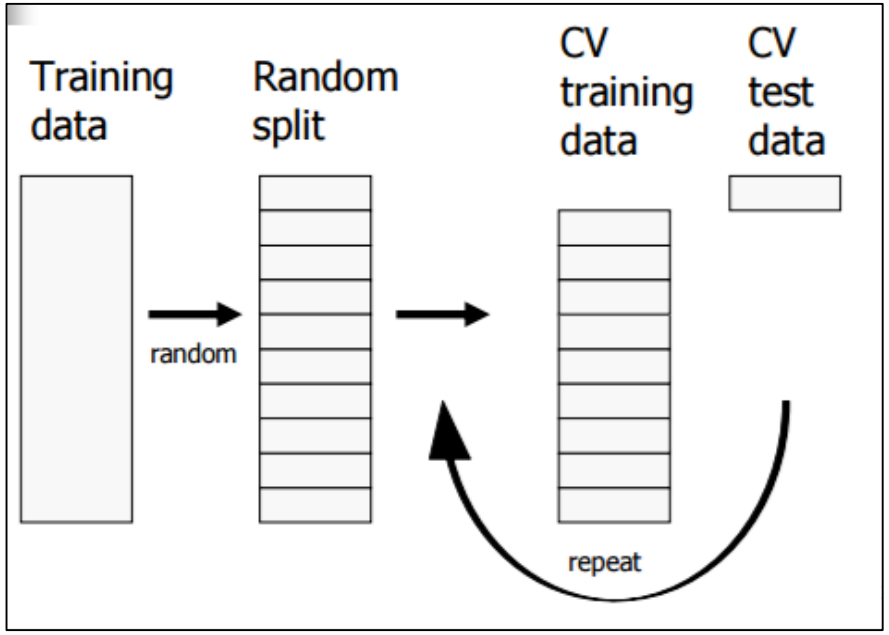
- **모델 검증 및 평가를 위한 데이터의 구분 : Hold-out 방식**
  - 주어진 자료를 다음의 세 그룹으로 랜덤하게 분할한 뒤, 주어진 목적에 따라 각각 모델의 훈련, 검증, 평가에 활용함.

- **훈련 데이터 (Training data)**
  - 모델의 학습을 위해 사용되는 자료.
- **검증 데이터 (Validation data)**
  - 훈련 자료로 적합되는 모델을 최적의 성능으로 튜닝하기 위해 사용되는 자료.
  - 훈련에 필요한 하이퍼파라미터(hyperparameter)를 조정하거나, 변수선택(model selecting) 등에 이용.
- **평가 데이터 (Test data)**
  - 훈련 및 검증 자료로 적합된 최종 모형이 미래에 주어질 새로운 자료에 대하여 얼마나 좋은 성과를 갖는지를 평가하는데 사용되는 자료.

머신러닝 모델 분석방법

⊙머신러닝 모델의 검증 및 평가

- **모델 검증 및 평가를 위한 데이터의 구분 : K-fold 교차검증(CV, Cross-Validation) 방식**
  - 자료의 수가 충분하지 않은 경우에는 훈련 데이터에서 너무 많은 양의 데이터를 검증 또는 평가 데이터에 뺏기지 않도록 교차 검정(CV) 기법을 사용.

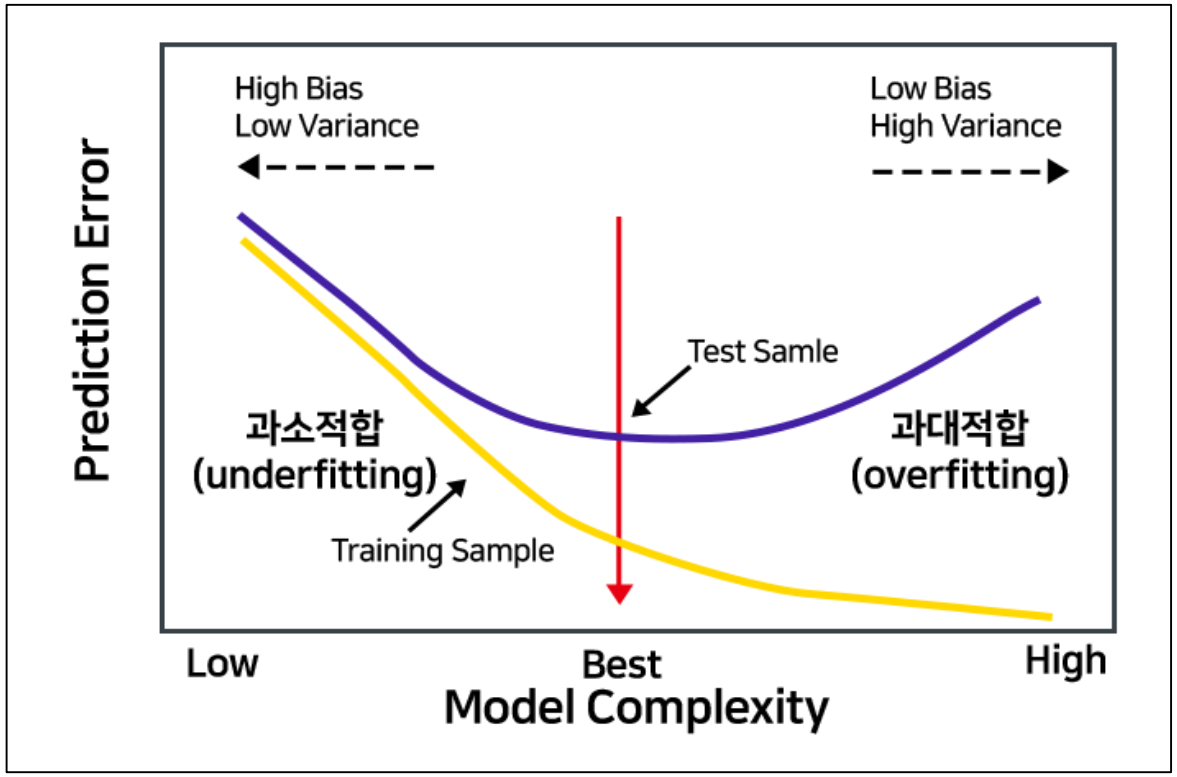


- **K-fold 교차검증(CV) 절차**
  - 자료를 균등하게 k개의 그룹으로 분할한 뒤
  - 각 j에 대하여, j번째 그룹을 제외한 나머지 $k-1$개 그룹의 자료를 이용하여 모델을 적합
  - j번째 그룹의 자료에 적합된 모델을 적용한 뒤 예측 오차를 구함.
  - $j = 1, \ldots, k$ 에 대하여 위의 과정을 반복한 뒤, k 개의 예측 오차의 평균을 구함.
  - 예측 오차의 평균값을 기준으로, 모델의 검증 또는 평가를 수행

머신러닝 모델 분석방법

⊙일반화 오차(Generalized Error)및 편향-분산 트레이드 오프(Bias-Variance Trade off)

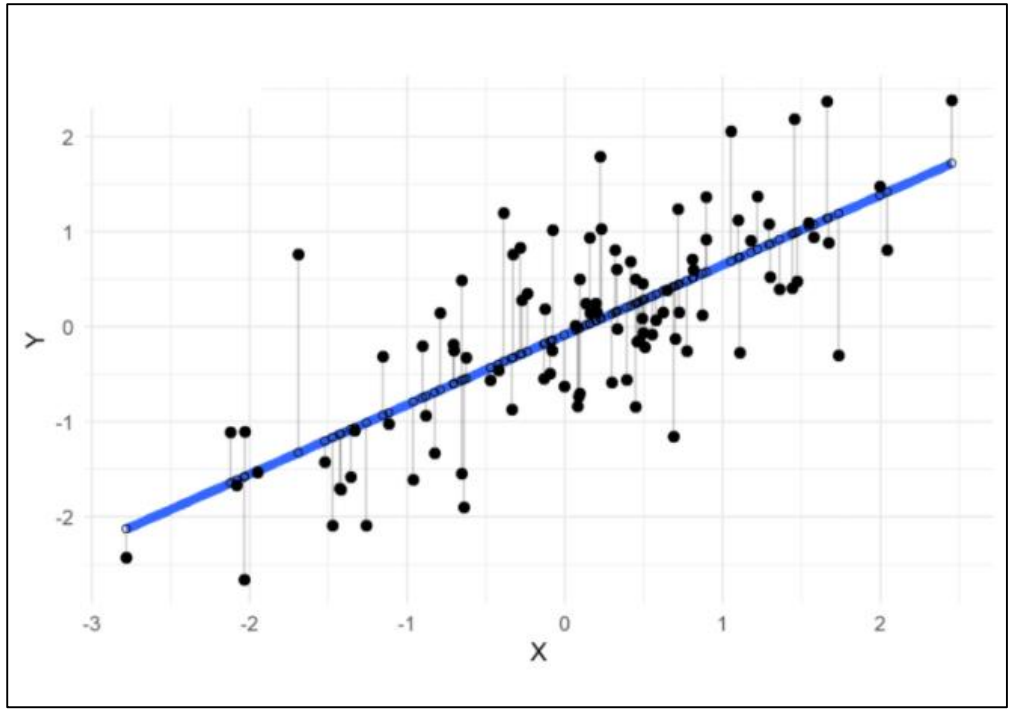
- **모델 복잡도에 따른 예측 오차**
  - 모델의 복잡한 정도에 따라 훈련 데이터와 평가 데이터의 예측 오차는 일반적으로 다음과 같은 패턴을 보이게 됨.



머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- 회귀(Regression) 모델의 평가지표



▫ **RMSE (Root mean square error)**

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

▫ **R square (Coefficient of determination,결정계수)**

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

머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- **분류(Classification) 모델의 평가지표**

ID	X1	…	Xk	Y	P(Y=1) 예측값	Y 예측값
1	0.5736		0.5	1	0.9960	1
2	0.9876		0.2	1	0.9875	1
3	0.4366		0.7	1	0.9845	1
4	0.8791		0.3	1	0.8893	1
5	0.8462		0.0	0	0.7628	1
6	0.2198		0.4	1	0.7070	1
7	0.2911		0.2	0	0.6808	1
…	…	…	…	…	…	…
89	0.1512		0.4	0	0.0480	0
90	0.9824		0.1	0	0.0383	0
91	0.6375		0.7	1	0.0249	0
92	0.4177		0.7	1	0.0218	0
93	0.0116		0.0	0	0.0161	0
94	0.5114		0.4	0	0.0036	0

분류기준값 : 0.5		예측범주	
		0	1
실제범주	0	40	12
	1	7	35

머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- **분류(Classification) 모델의 평가지표**
  - 정오분류표

정오분류표		모형에 의한 예측	
		Negative	positive
실제 자료	Negative	A (**TN**, true negative)	B (**FP**, false positive)
	Positive	C (**FN**, false negative)	D (**TP**, true positive)

머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- **분류(Classification) 모델의 평가지표**
  - 정확도, 정분류율 (Accuracy) :
    - 전체 관찰치 중 정분류된 관찰치의 비중

$$\frac{A+D}{A+B+C+D}=\frac{TN+TP}{TN+FP+FN+TP}$$

정오분류표		모형에 의한 예측	
		Negative	positive
실제자료	Negative	A (**TN**)	B (**FP**)
	Positive	C (**FN**)	D (**TP**)

  - 정밀도 (Precision)
    - Positive 로 예측한 것 중에서 실제 범주도 Positive인 데이터의 비율

$$\frac{D}{B+D}=\frac{TP}{FP+TP}$$

  - 재현율 (Recall)
    - 실제 범주가 Positive인 것 중에서 Positive 로 예측된 데이터의 비율

$$\frac{D}{C+D}=\frac{TP}{FN+TP}$$

  - F1 Score
    - 정밀도와 재현율의 조화평균.

$$2\times\frac{Precision\times Recall}{Precision+Recall}$$

머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- **분류(Classification) 모델의 평가지표**
  - ROC(Receiver operating characteristic) 도표
    - 분류의 결정임계값(threshold)에 따라 달라지는 **TPR**(민감도, sensitivity)과 **FPR**(1-특이도, 1-specificity)의 조합을 도표로 나타냄.
      - **TPR** : True Positive Rate (=sensitivity(민감도))
        1인 케이스에 대해 1로 잘 예측한 비율.
      - **FPR** : False Positive Rate (=1-specificity(특이도))
        0인 케이스에 대해 1로 잘못 예측한 비율.

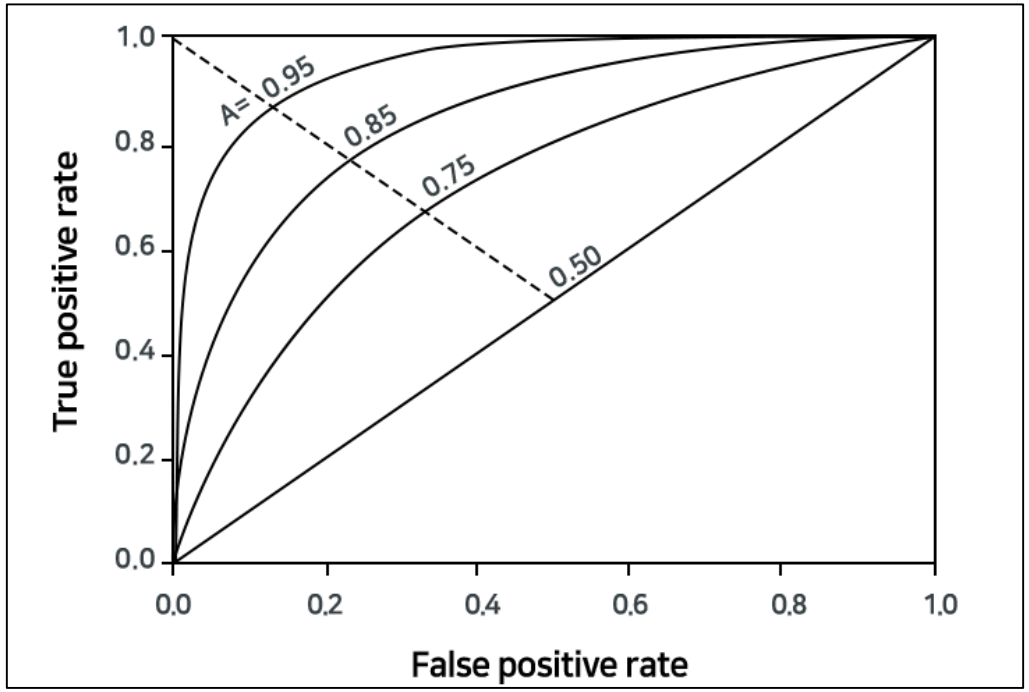
정오분류표		모형에 의한 예측	
		Negative	positive
실제자료	Negative	A (**TN**)	B (**FP**)
	Positive	C (**FN**)	D (**TP**)

머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- **분류(Classification) 모델의 평가지표**
  - ROC(Receiver operating characteristic) 도표
    - 임계값이 1이면 FPR=0, TPR=0
    - 임계값을 1에서 0으로 낮춰감에 따라 FPR과 TPR은 동시에 증가함.
    - FPR이 증가하는 정도보다 TPR이 빠르게 증가하면 이상적

      ⇒ **왼쪽 위 꼭지점에 가까울수록 좋음**



머신러닝 모델의 평가

⊙ 지도학습 모델의 평가 지표

- **분류(Classification) 모델의 평가지표**
  - AUC (Area Under the Curve)
    - ROC 곡선 아래의 면적.
    - 가운데 대각선의 직선은 랜덤한 수준의 이진 분류에 대응되며, 이 경우 AUC는 0.5임.
    - 1에 가까울수록 좋은 수치. FPR이 작을 때 얼마나 큰 TPR을 얻는지에 따라 결정됨.