Introduction Of Machine Learning

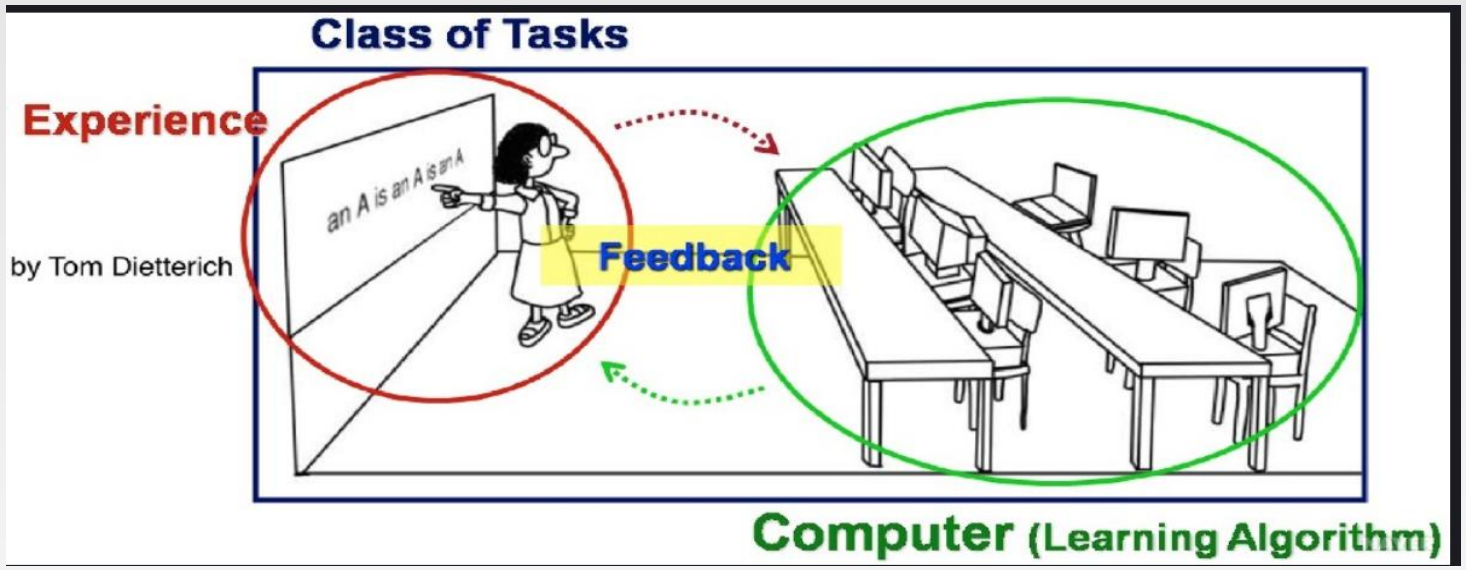
민형규

목차

1. 머신러닝이란?

- “컴퓨터에 명시적인 프로그램 없이 배울 수 있는 능력을 부여하는 연구 분야” – 아서 사무엘,1959

- 컴퓨터로 하여금 과거의 데이터를 이용해 새로운 함수를 만들게 하는 것

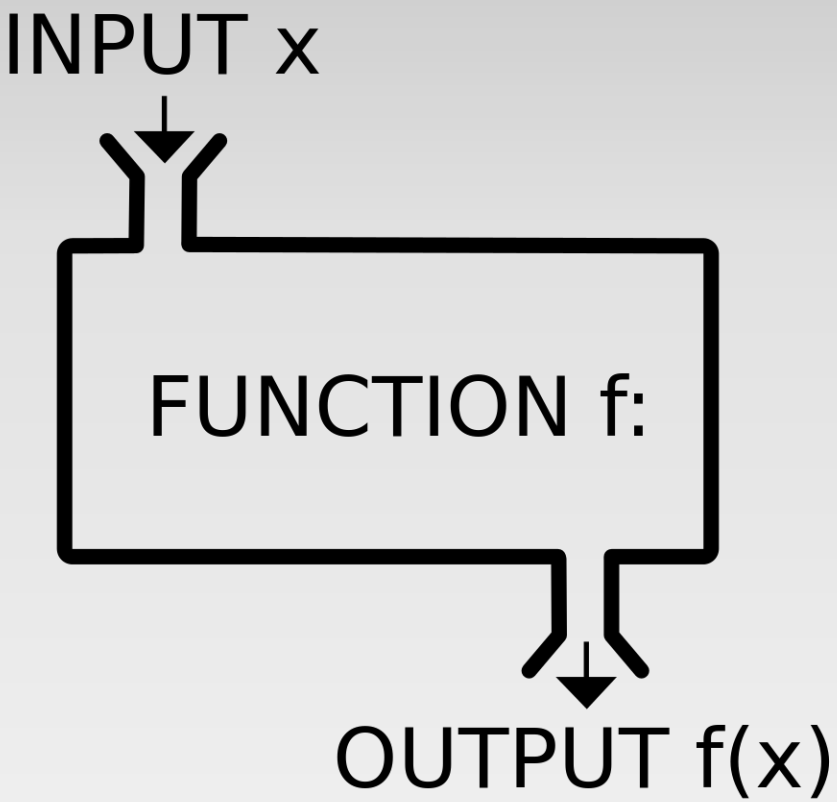


1. 머신러닝이란? - 머신러닝의 종류

- 지도 학습 (Supervised Learning)
  - 정답지(y)를 이용하는 학습
  - Y ~ X
- 비지도 학습 (Un-supervised Learning)
  - 정답지 없이 데이터의 패턴만 이용하는 학습
  - X
- 강화 학습 (Reinforcement Learning)
  - 가상 환경을 조성하여 특정 환경과 특정 행위의 보상점수를 이용하는 학습
  - A~R,S

1. 머신러닝이란? – 모델

- 모델
  - 이상이 되는 규범
  - 머신 러닝의 목표
  - 함수 y=f(x)의 역할과 유사



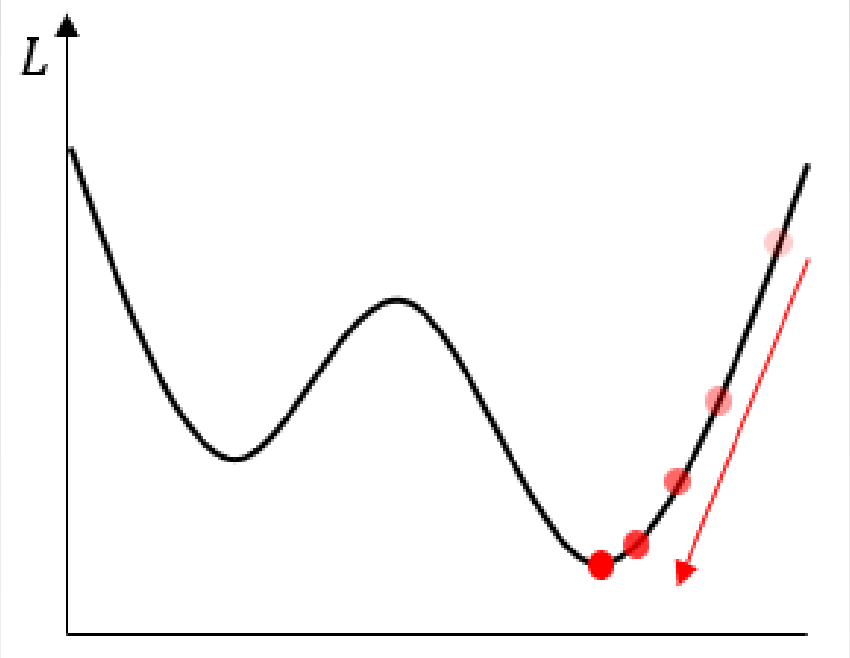
1. 머신러닝이란? – 손실 함수와 비용 함수

- Loss Fucntion (≈Cost function)
  - 컴퓨터에게 전달하는 피드백
  - 모델을 이용한 예상치와 실제 값의 차이(오차)에 관한 함수
  - 여러 파라미터를 반복해서 실험해 봄으로써 오차를 최소화 시키는 최적의 파라미터를 찾음

1. 머신러닝이란? – 수치해석

- MSE(Mean Square Error) = $\frac{1}{n}\sum(y-\hat{y})^2$
  - 함수의 형태 -> 수치해석을 통한 근 찾기 가능
  - $\hat{\theta} = argmin_{\theta} f(\theta)$

- 경사 하강법 (Gradient Descent)
  - $\theta_{t+1} = \theta_t - \gamma * \frac{\delta}{\delta\theta} loss(\theta)|_{\theta=\theta_t}$



2. 모델 만들기 -모델링의 과정

A. 데이터 정제 (Data Cleansing)

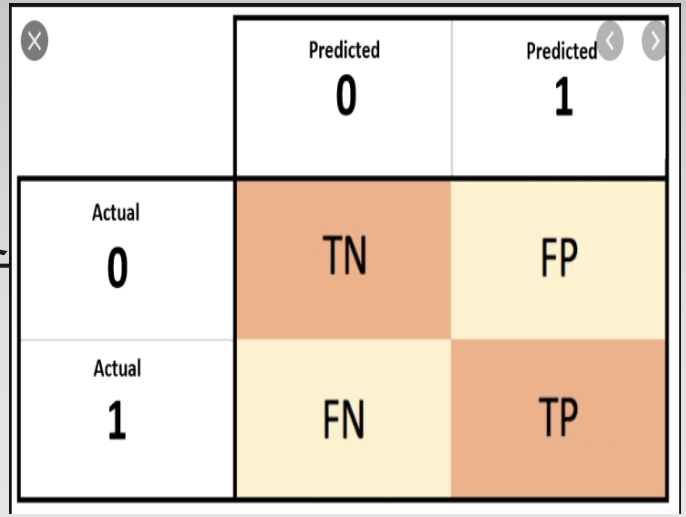
B. 모델 적합 (Model Fitting)

C. 모델 평가 (Model Evaluation)

-실습 노트 참고

2. 모델 만들기 – 모델 평가 점수

Classification
-정확도(accuracy) – 전체 중 맞춘 것의 개수
-정밀도(precision) – 1로 예측한 것 중 실제 1의 개수
-재현율(recall) – 실제 1 중 1로 예측해낸 것의 개수
-F1-score – 정밀도와 재현율의 조화 평균



		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

2. 모델 만들기 – 모델 평가 점수

Regression

-MSE(평균 제곱 오차)
$= \frac{1}{n}\sum(y - \hat{y})^2$

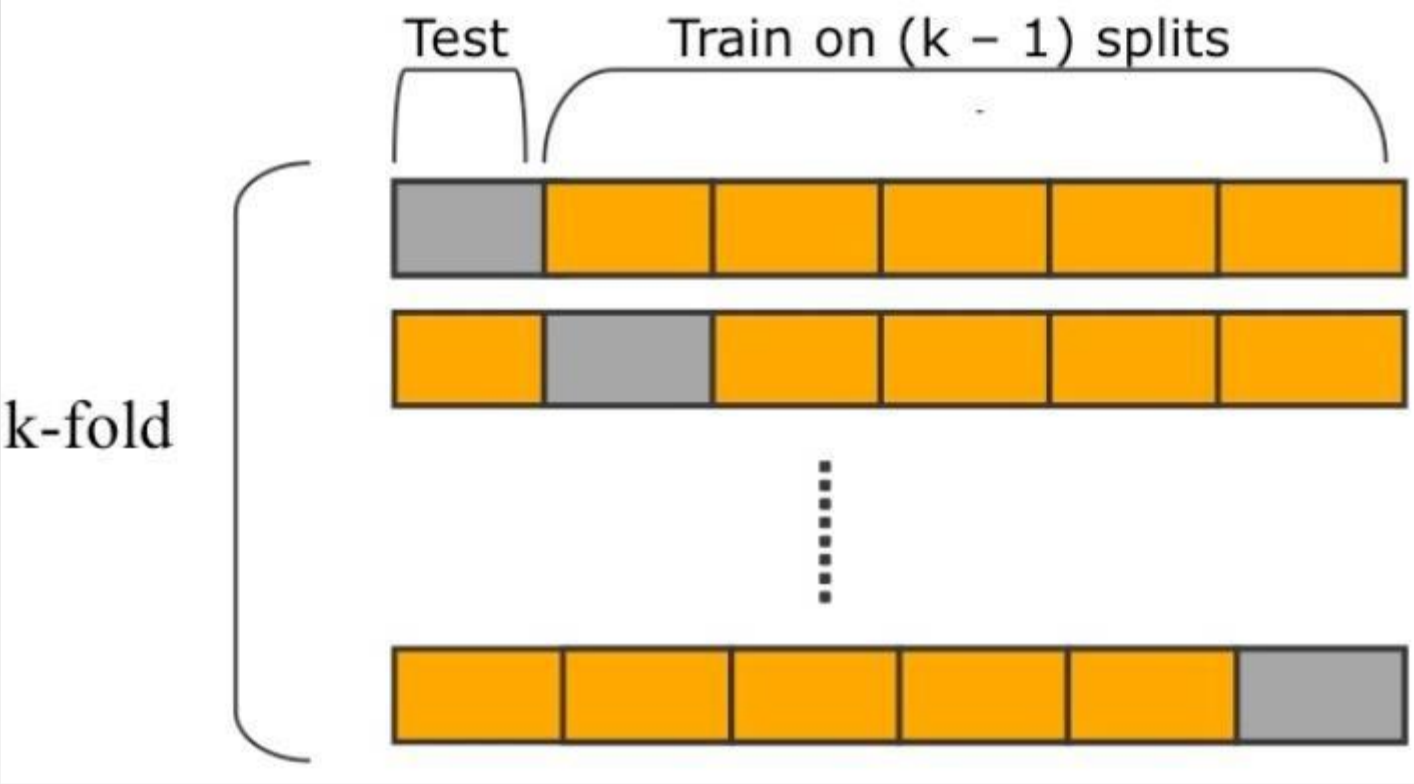
-R2(결정계수) = SSR/SSTO
$= \sum(\hat{y} - \bar{y})^2 / \sum(y - \bar{y})^2$

-기타
AIC,BIC

2. 모델 만들기 – Cross Validation

- 훈련하는 데이터와 검사하는 데이터를 분리 시킴으로써 오버피팅을 확인 할 수 있음

- Cross Validation은 트레인 셋을 테스트로, 테스트 셋을 트레인으로 사용함으로써 교차검증 하는 방식



3. 좋은 모델 만들기 – 오버피팅

- Train 데이터에 과하게 적합되어 새로운 데이터에 적합이 덜 되는 현상

- 피쳐(설명변수)가 많아지면 많아질수록 데이터는 과적합 된다.

- $MSE(\hat{\theta}) = E(y - \hat{y})^2 = E(y - E(\hat{y}) + E(\hat{y}) - \hat{y})^2$

$$= E\left(y - E(\hat{y})\right)^2 + E\left(\hat{y} - E(\hat{y})\right)^2$$

$$= Bias(\hat{y})^2 + Var(\hat{y})$$

- 실습 노트 참고

3. 좋은 모델 만들기 – 자료의 양상 파악

유의해야할 데이터 양상들

-Type

-Outlier

-Sparce Data

-Imblanced Data

3. 좋은 모델 만들기 – 자료의 양상 파악

Type

-데이터의 타입에 따라 처리나 적용이론이 달라짐

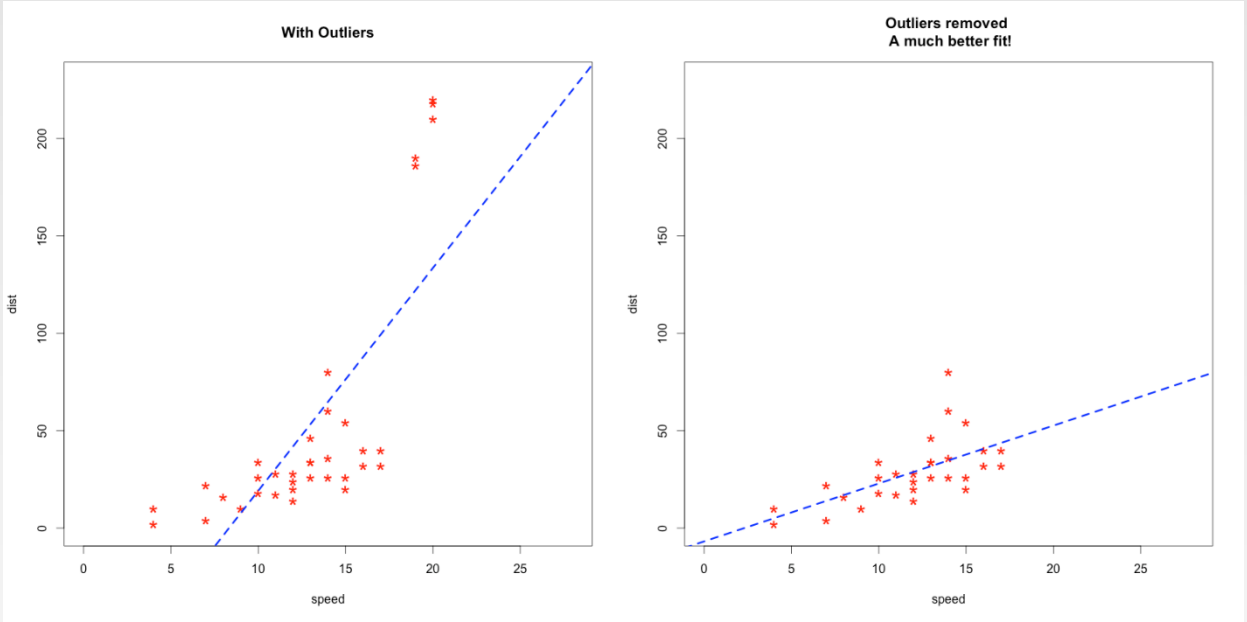
- 데이터가 수치형 인지 명목형 인지
- 데이터가 연속형 인지 이산형 인지
- 데이터가 string인지 integer인지 float인지

3. 좋은 모델 만들기 – 자료의 양상 파악

Outlier

- 아웃 라이어는 알고리즘에 따라 모델에 크게 영향을 줌

- ➢ 아웃라이어를 제거하거나
- ➢ 아웃라이어에 강건한 모델 사용



3. 좋은 모델 만들기 – 자료의 양상 파악

Sparse Data

- 데이터 프레임에 빈 공간이 많이 생기는 현상

- 더미 변수를 너무 많이 만들 시 정보를 거의 담고 있지 않은 변수들이 만들어지며 이것이 결과를 왜곡시킴

➢ 적합한 알고리즘을 사용해 주거나 차원 축소를 해줘야 함

3. 좋은 모델 만들기 – 자료의 양상 파악

Imbalanced Data

-데이터의 불균형은 추정에 있어서 영향이 큼

1. 타깃(y)가 불균형 할 시 (ex)발암 확률 예측)

➢ Threshold 조절이 필요해짐

2. 설명변수(x)가 불균형 할 시 (ex)킹카운티의 waterfront 변수)

➢ 알고리즘에 따라 샘플을 축소 시켜서 보는 것과 마찬가지임

➢ 업 샘플링/다운 샘플링을 통해 변수 비율을 맞춰줄 수 있음

Ex) GAN, Smote

3. 좋은 모델 만들기 –기타 유의점

- 분류 모델의 경우, 문제에 따라 재현율과 정밀도의 가치가 달라진다

Ex) 암 검사의 경우, 보험 사기 포착의 경우

- 모델이 실제로 사용 가능한지를 항상 염두에 둬야 한다.

Ex)
인과의 역전
(기후 예측 모델에 당해 농산물 출하량을 사용하는 경우)
포착이 힘든 설명변수 사용
(마트 고객의 사용금액 예측 모델에 고객의 자산을 사용하는 경우)