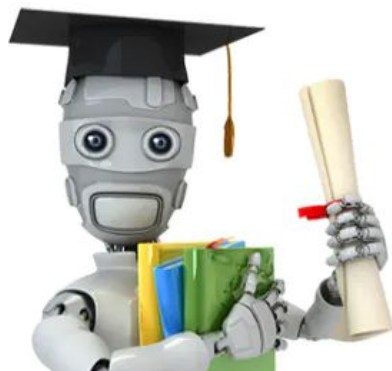


Review of ML Process & Classification



2022.01.11

Dongmin Kim (tommy.dm.kim@kaist.ac.kr)

TODO

- 머신러닝의 과정 리뷰
- 분류 (Classification) 리뷰
- Kaggle Titanic EDA

I. Machine Learning Process

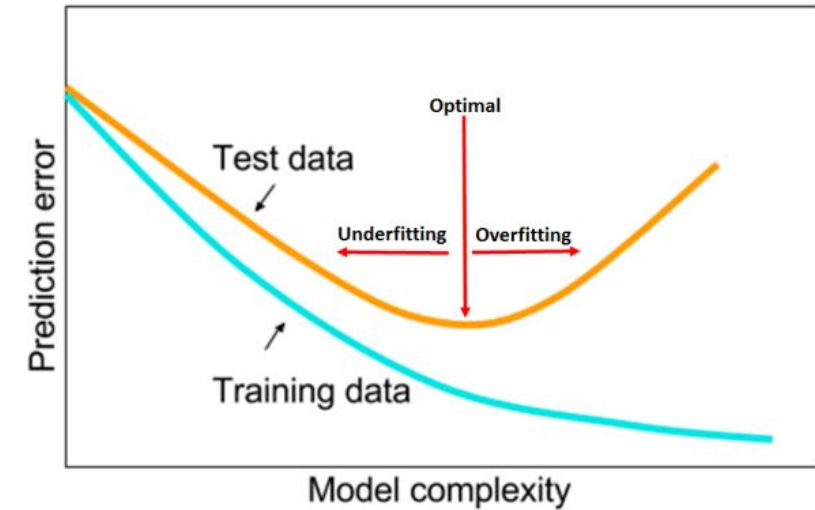
머신러닝의 과정 리뷰



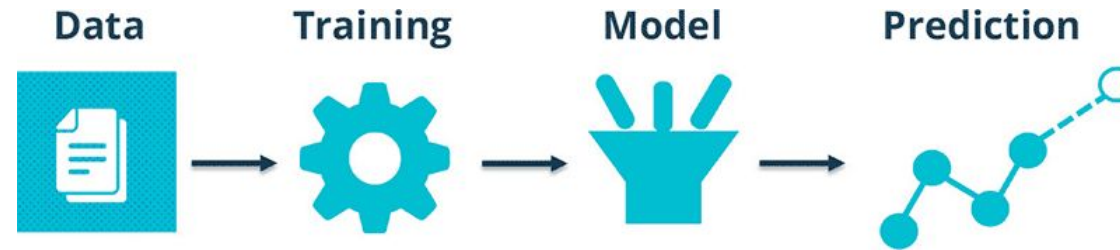
머신러닝의 과정 리뷰



An example of overfitting, underfitting and a model that's "just right!"



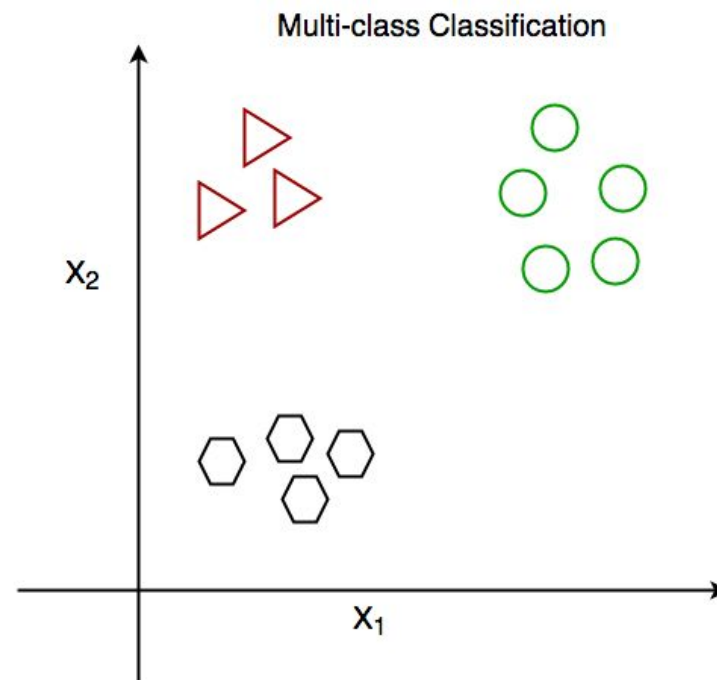
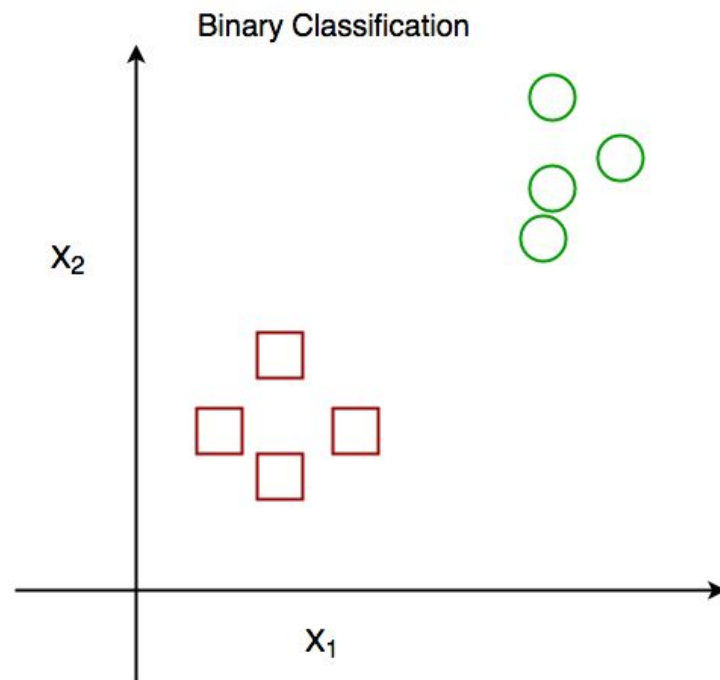
머신러닝의 과정 리뷰



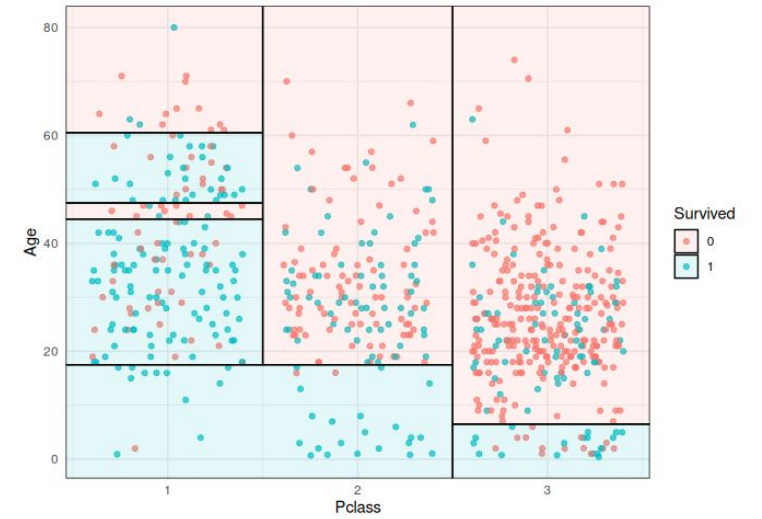
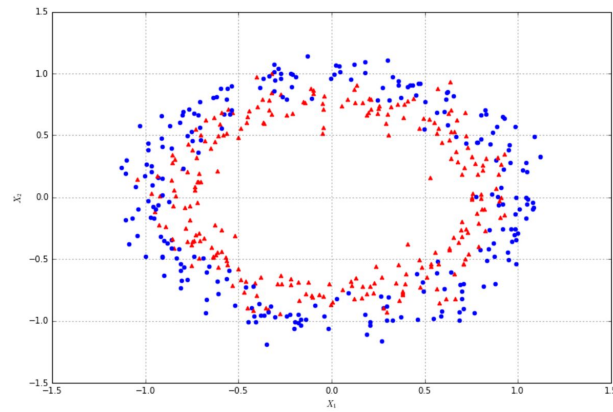
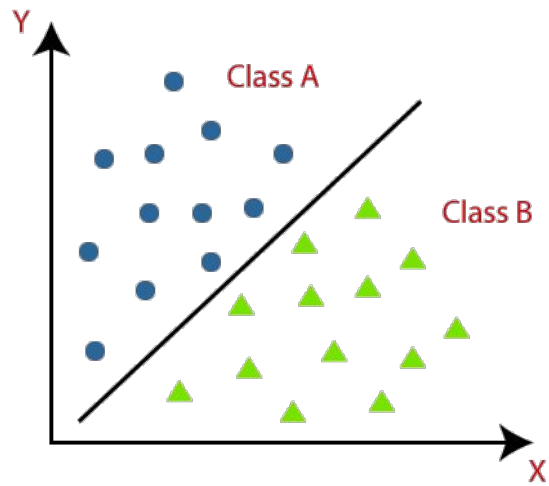
- Data
 - EDA (Exploratory Data Analysis)
 - 전처리 (Preprocessing)
- 학습 (Training)
 - 모델 선정
 - Loss Function + Regularization
 - hyperparameter tuning
- 평가 (Evaluation)
 - 정확도 (accuracy), MSE 등의 metric으로 test set에 대해 평가
 - 결과의 시각화

2. Classification

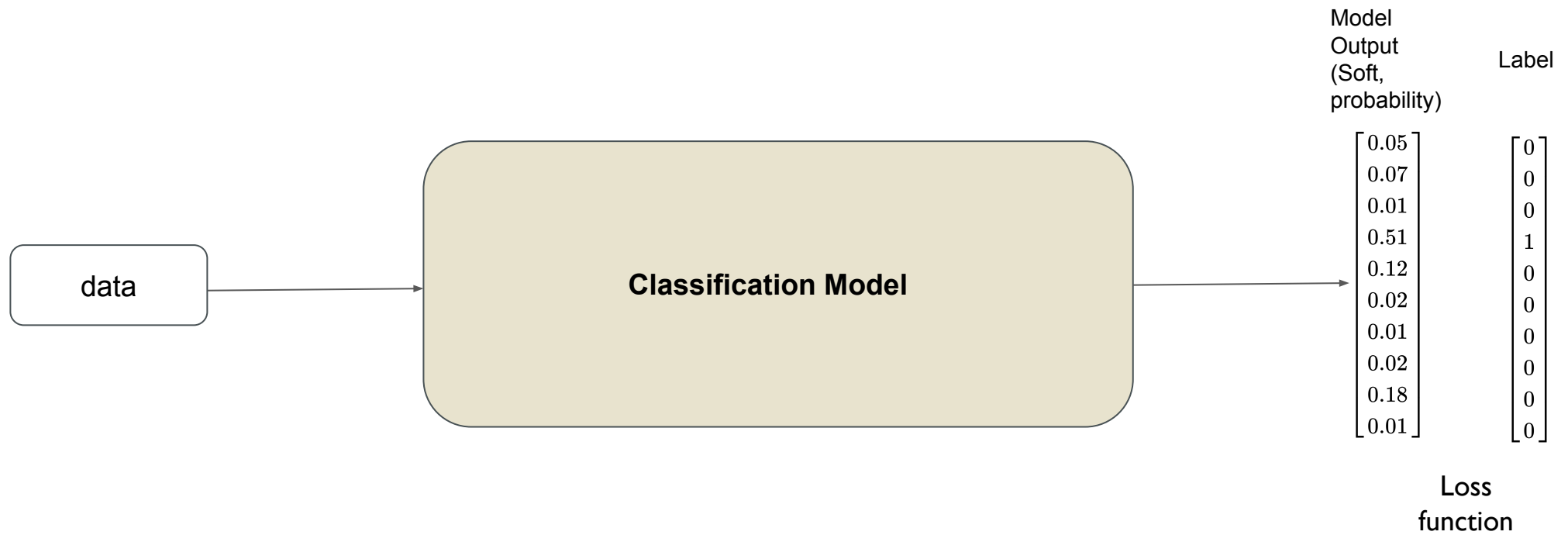
Data



Classification == 선긋기



Classification의 입출력

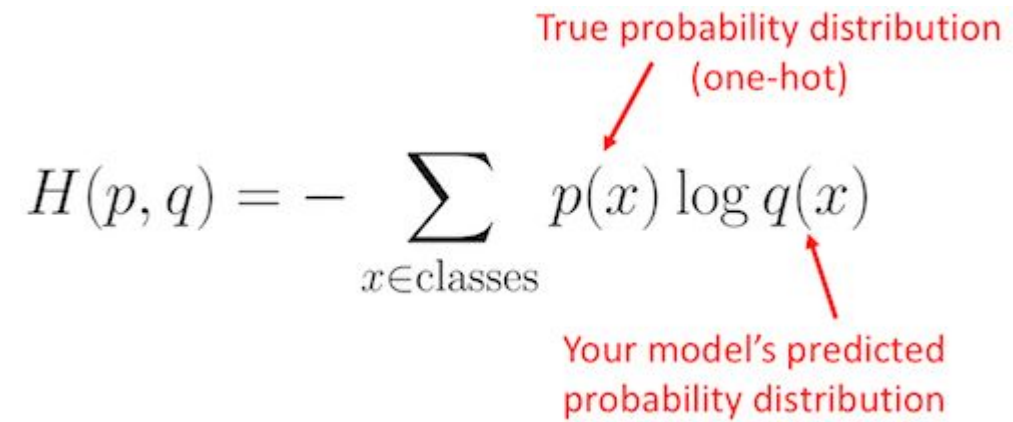


Cross Entropy

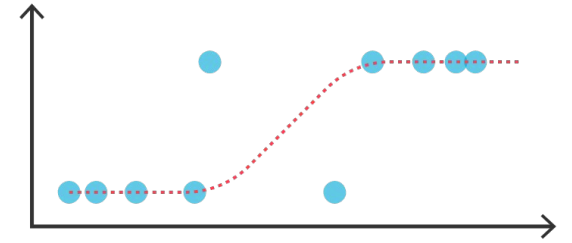
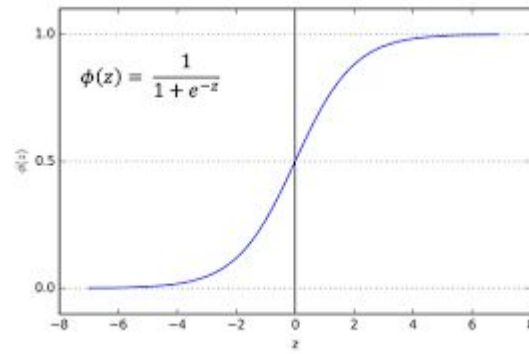
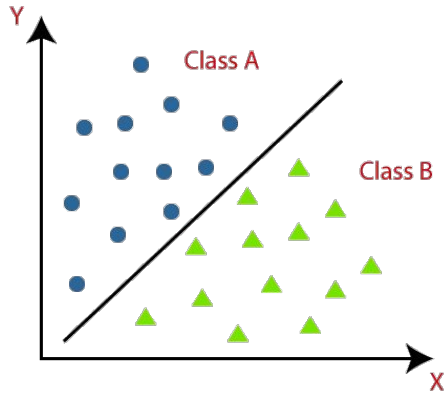
$$H(p, q) = - \sum_{x \in \text{classes}} p(x) \log q(x)$$

True probability distribution
(one-hot)

Your model's predicted
probability distribution



Logistic Regression의 출력



Metrics 1: Accuracy

$\text{Accuracy (정확도)} = \text{맞은 개수} / \text{총 데이터 수}$

Metrics 2: Precision, Recall and F1-score

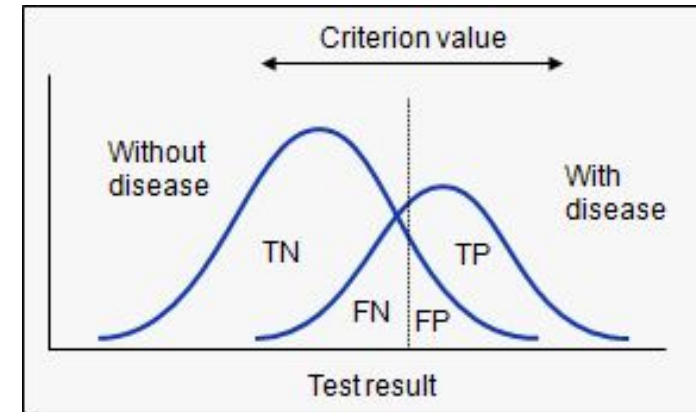
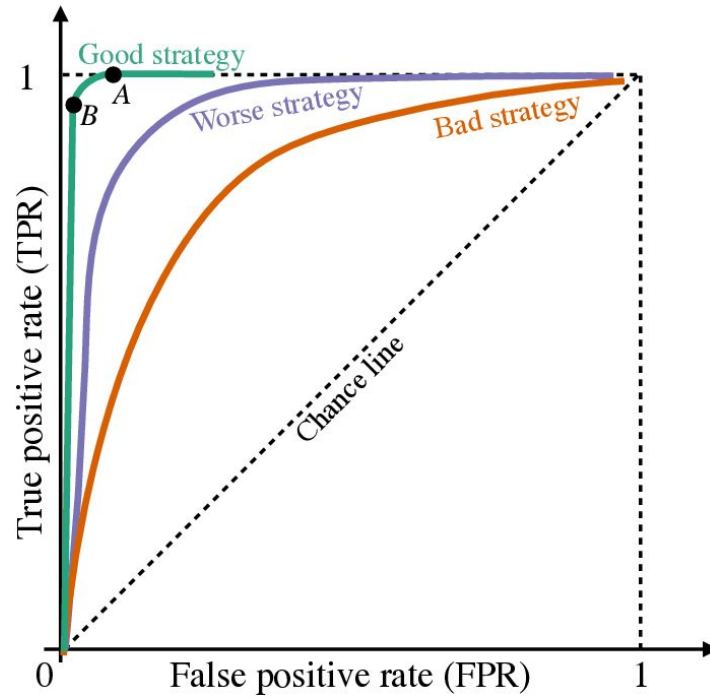
TP,TN, FP, FN

True	Positive
True	Negative
False	Positive
False	Negative

Precision, Recall, F1 score

Accuracy		
Precision		
Recall		
F1 Score		

Metrics 3: ROC, AUC



Summary

- 머신러닝의 과정 (supervised setting) 리뷰
 - Model input / output
 - Cost function, Regularization
 - 데이터 전처리
 - Data split (Train / Validation / Test) and hyperparameter tuning
- 분류 (Classification) 과정 리뷰
 - Classification objective
 - Cost function of classification task
 - Accuracy and other metrics

3. Kaggle

Kaggle Titanic EDA

0. Kaggle 가입이 안되어 있다면, 가입하기
1. 일단 제출해보기 [Submission 1]
 - 1.1. Pandas를 활용하여 train.csv, test.csv, gender_submission.csv 확인해보기
 - 1.2. Survived 값을 모두 1 혹은 모두 0으로 하여, 제출해보기
2. EDA
 - 2.1. train data의 “Survived” column의 분포 보기. 0 (dead), 1 (Survived) 의 개수 출력
 - 2.2. Pclass, 성별에 따른 생존자 분포를 확인해보세요 (hint: sns.barplot).

다 하신 분들은, 추가적으로 다른 feature들도 살펴보고 plot 해보며 insight를 얻어 봅시다.
 - 2.3. 뭔가 쓰기 힘들 것 같은 feature가 있나요? Nan값이 있는지 살펴보고, 숫자로 처리하기 쉬운만한 데이터만 처리하려 합니다.

리스트 input_features, output_features를 선언하고, feature 이름들을 넣어주세요. 그리고 변수 X에 input, y에 output을 할당해주세요.
3. 데이터 가공하기
 - 3.1. 결측치를 처리합시다. na (혹은 null) 값을 모두 -1로 채워주세요.
 - 3.2. train.csv를 읽은 pandas dataframe을 train/test 80%, 20% 비율로 분리해주세요.