Hyperparameter tuning

Sang Yup Lee

Machine Learning

- 파라미터의 종류
  - 학습을 통해서 그 값이 결정되는 파라미터
  - 사용자가 그 값을 결정하는 파라미터 => 이러한 파라미터를 hyperparameter라고 함
  - 어떠한 hyperparameter를 갖는지는 모형 마다 달라짐
    - 또는 어떠한 방법을 사용해서 학습을 하느냐에 따라 달라짐
  - Hyperparameter가 취할 수 있는 값은 여러가지
    - 이중 모형의 성능을 좋게하는 값을 선택하는 것이 필요

Logistic Regression

- LogisticRegression class에서의 Hyperparameter 의 예
  - C, penalty, solver 등
  - 그렇다면 어떠한 값으로 설정을 하는게 좋은가?
- Hyperparameter의 값 결정
  - 사람이 결정하기 때문에 자동적으로 최적의 값을 결정하기 어려워
  - 사전 지식을 가지고 몇가지 값을 시도
  - 모형의 성능이 더 좋은 것을 선택

Logistic Regression

- Model tuning
  - hyperparameter의 값을 변경하는 것 (혹은 그러한 과정을 거쳐서 성능이 더 좋은 모형을 찾는 것)
- Example
  - Penalty 종류에 따른 모형의 성능 비교
- 주의!
  - Hyperparameter tuning의 결과를 파악하기 위해서 평가 데이터를 사용하면 안됨
  - Test dataset은 모형의 최종 평가 목적으로 사용되므로 학습이나 튜닝에 사용되면 안됨 => 즉, 모형의 성능을 개선하는 목적으로 사용하면 안되고, 오직 (최종) 모형의 성능을 평가하는 목적으로 사용

Logistic Regression

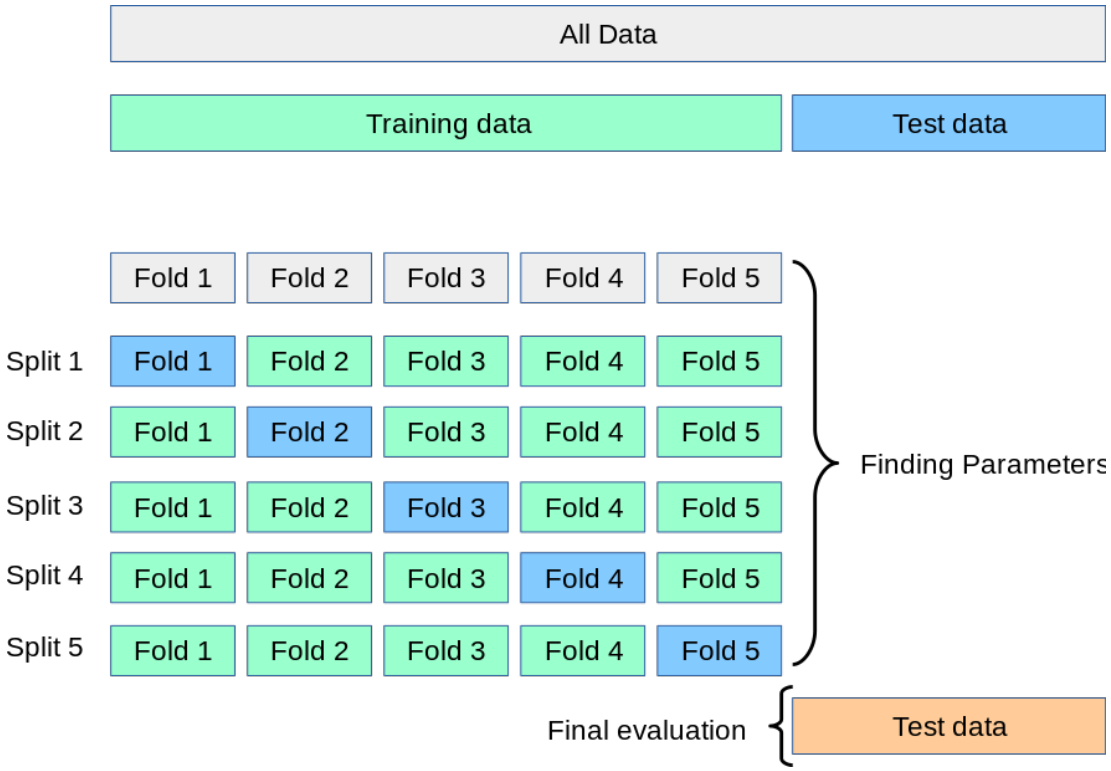
- Validation dataset
  - Model tuning의 목적으로 사용
  - 학습데이터의 일부를 validation dataset으로 사용
- Penalty 유형의 예
  - 서로 다른 규제화 방법에 대한 모형들 중에 어떠한 모형의 성능이 좋은가를 평가하기 위해서는 validation dataset을 사용 => 그 후, 더 성능이 좋은 모형을 이용하여 test data를 사용하여 평가

Logistic Regression

- Validation
  - Validation dataset을 이용해서 모형의 성능을 파악해 보는 것 (최종 평가가 아님)
- K-fold cross validation
  - The training set is split into k smaller sets, called “folds”. The following procedure is followed for each of the k “folds”:
    - A model is trained using k-1 of the folds as training data;
    - The resulting model is validated (i.e., tested) on the remaining part of the data
  - This process is repeated k times
  - Validation을 여러번 수행하는 이유 => 모형의 일반화 정도를 높이기 위해서

Logistic Regression

- Example: 5-fold cross validation

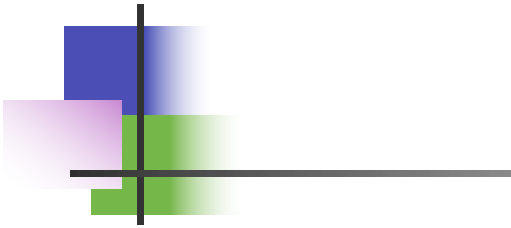


Logistic Regression

- K-fold cross validation
  - from sklearn.model_selection import cross_val_score
  - scores = cross_val_score(model, X_train, y_train, cv=5)
- Refer to “k_fold_validation_grid_search.ipynb”

Grid search

- How to find the optimal values of hyperparameters?
- What is it?
  - Grid-search is used to find the optimal *hyperparameters* of a model which results in the most 'accurate' predictions.
- When to use?
  - When a hyperparameter can take numerous values
  - For a set of values that the user sets, the grid search method automatically finds the best hyperparameter values that leads to the best model (i.e, the model with best performance)
- Refer to "k_fold_validation_grid_search.ipynb"



Q & A