Data Science

Lab 4

Learning Models and Evaluation

Ok-Ran Jeong

Contents

- 1. Linear Regression
  - Holdout method
- 2. Decision Tree
  - Holdout method
- 3. k-Nearest Neighbors
  - K-fold cross validation

Problem 1: Linear Regression

- Using a linear regression model, predict and evaluate the results of the median_house_value.
- Dataset: California Housing Prices

https://www.kaggle.com/camnugent/california-housing-prices

- Training and Testing
  - Split the dataset into 4/5 (then 3/5) for training and 1/5 (2/5) for testing.
  - Evaluate the model using the holdout method, with the shuffle and stratify options in dataset split.
- Show all the results.

Problem 2: Decision Tree

- Using a decision tree model, predict, and evaluate the results of wine quality.
- Dataset: Wine Quality

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

- Training and Testing
    - Each sample data contains all feature values separated by semi-colons. You need to split them. Use only the red wine dataset.
    - Split the dataset into 9/10 (then 8/10, 7/10) for training and 1/10 (then 2/10, 3/10) for testing.
    - Evaluate the model using the holdout method, with the shuffle option in dataset split
- Show all the results.

Problem 3: K-NN

- Using the k-nearest neighbors model, predict, and evaluate the results of digit recognition.
- Dataset: MNIST (Use only the train dataset.)

https://www.kaggle.com/oddrationale/mnist-in-csv

- Split the dataset into 5 subsets of equal size
  - Use 5-fold cross validation method for evaluation
  - After the initial model testing with k=3 and k=5, do hyperparameter tuning by using GridSearch and Randomized GridSearch.
- Show and compare the results for the base model and hypertuned models.

End of lab