Lab Session 2

Seok-Ju Hahn

Master course student @ UNIST Data Mining Lab.
(Prof. Junghye Lee)

Contents

- Revisit: Data Mining Workflow
- Example: Heart Disease Prediction
- Good References
- Kaggle: data science playground

Revisit: Data Mining Workflow

- Observation & Hypothesis (1: **Data Preparation**)
    - Collecting & Storing data (← nowadays a main role of a *data engineer*)
    - Exploratory data analysis (EDA)
    - Cleansing data (munging, imputation, removing duplicated values, feature scaling…)
- Experiment (2: **Modeling**)
    - Selection of appropriate machine learning algorithms
    - Training & evaluating & tuning models
- Analysis (3: **Evaluation & Feedback**)
    - Testing models
    - Delivering the result to stakeholders
    - Deploying & monitoring services/systems or releasing products (← nowadays a main role of a *data engineer*)

Example: Heart Disease Prediction

- Observation & Hypothesis
  - Heart disease is problematic; once it occurs, it causes serious results on patients.
  - As a clinician, I want to prognose the potential cause of heart disease in advance.
  - From what I learned in medical school and based on previous studies in this field, I filtered out 12 variables to predict heart disease.
    - Those variables represent my 'hypothesis'.
- Experiment & Analyze
  - I will train simple classification model; logistic regression model
  - If my hypothesis works well, I expect the classification performance to be reached to some level.
    - e.g. ROC-AUC score over 0.7, F1-score over 0.4, etc.
- Let's code!

Good References

- MatPlotLib Tutorial by Aurelien Geron (author of Hands on Machine Learning <- Highly recommended book!!!)
  - https://colab.research.google.com/github/ageron/handson-ml2/blob/master/tools_matplotlib.ipynb?fbclid=IwAR0CvkDat3uEWwDhU--UPkGUzO6oCMAIKbQy4wn9wQrpJweqOSQisZvWF34
  - https://datascienceschool.net/view-notebook/39569f0132044097a15943bd8f440ca5/ (KOR)
- Pandas Tutorial
  - https://pandas.pydata.org/pandas-docs/stable/getting_started/tutorials.html
- Scikit-learn Tutorial
  - https://scikit-learn.org/stable/tutorial/index.html

Kaggle: data science playground

- Many smart experts are there!
- They share their codes, skills, insights, datasets, etc.
- It is the best place to learn and experience data science!

Questions?

Thank You