Model Evaluation

CS 418. Introduction to Data Science

Model Evaluation
Introduction

- The data used for **training** and **selecting** a model must be kept separate from the data used for **evaluating** the model.
- Otherwise, the error rate computed cannot be considered a representative of the performance of the model on **unseen** observations (**generalization error**), since it can be inaccurately low due to **overfitting**.
- The correct approach for **model evaluation** would be to partition the dataset into a **training set**, which is used for training and selecting the model, and a **test set**, which is used for computing the error rate.
- There are two main approaches for partitioning the dataset:
  - **Holdout method**.
  - **Cross-validation method**.

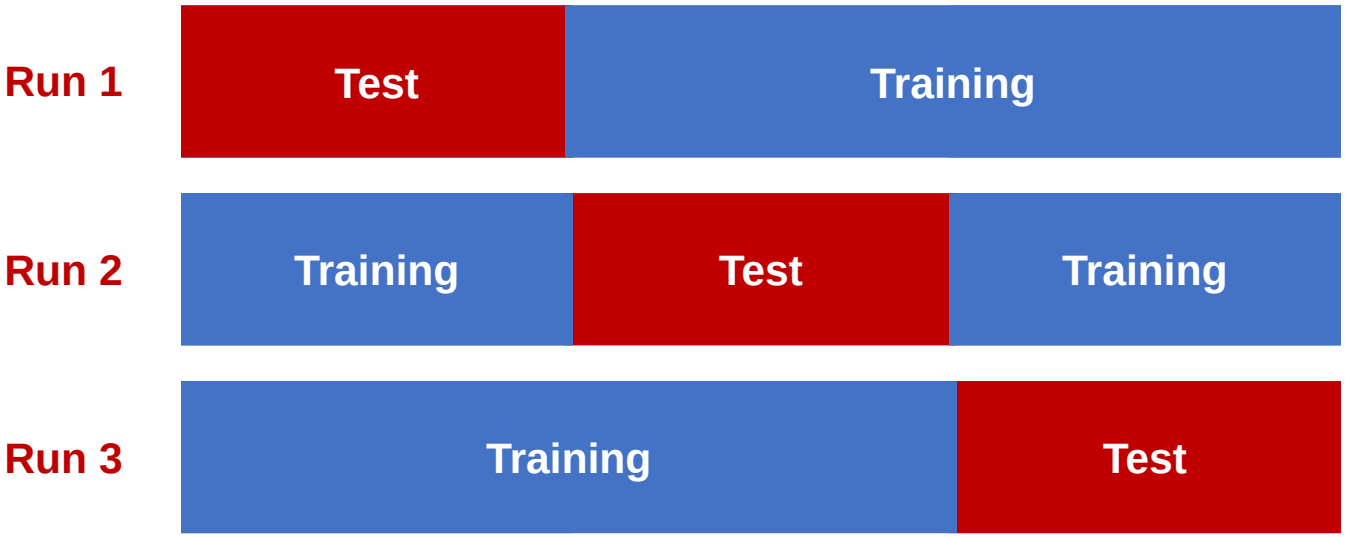
Model Evaluation
Holdout Method

- The most basic approach for partitioning a dataset is the **holdout method**, where the dataset is **randomly** partitioned into two **disjoint** sets called the **training set** and the **test set**.

- The error rate on the **test set** is used as an estimate of the **generalization error** rate of the model.
- **How to choose the proportion of data for the training and the test set?**
  - If the **training set** is **too small**, the model may not be built properly.
  - If the **test set** is **too small**, the error rate on the test set may be less reliable and can have a high variance.

Model Evaluation
Cross-Validation Method (I)

- Another widely used approach for partitioning a dataset is the ***k*-fold cross-validation method**, where the dataset is **randomly** partitioned into ***k* equal-sized** sets (or **folds**). During the ith run, the ith partition is chosen as the **test set** while the rest of the partitions are used as the **training set**.



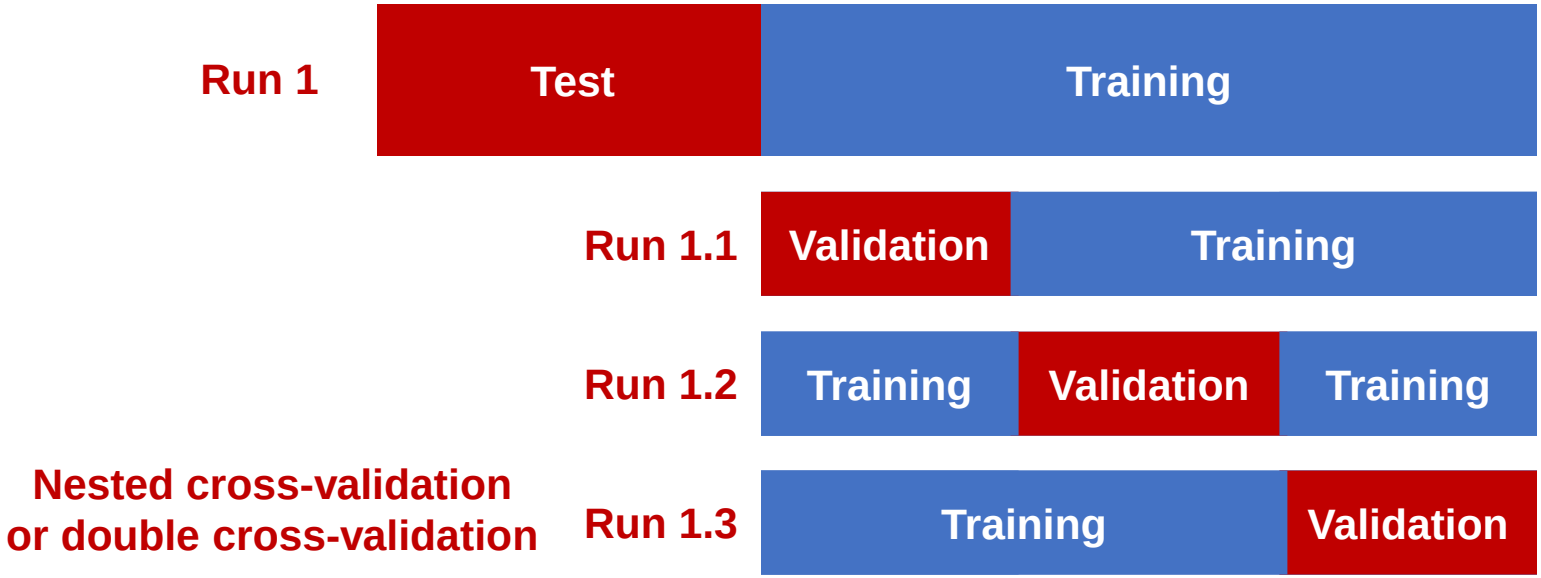
- The **aggregated** (e.g., average) error rate on the **test sets** is used as an estimate of the **generalization error** rate of the model.

Model Evaluation
Cross-Validation Method (II)

- **How to choose the value of k?**
  - A small value of k will result in a smaller **training set** at every run, while a large value of k will result in a smaller **test set**.
  - For most practical applications, a value of k **between 5 and 10** provides a reasonable approach for estimating the **generalization error** rate because each fold is able to use 80% to 90% of the data for training.
- **What are the advantages of the cross-validation method?**
  - We make effective use of all the data for both training and testing.
- **What are the disadvantages of the cross-validation method?**
  - **More computation time** is required.
  - The **aggregated** error rate does not reflect the **generalization error** rate of any of the k models.

Model Evaluation
Hyper-Parameter Selection

- **Hyper-parameters** are parameters that need to be determined before building the model.
- We can determine the values of the **hyper-parameters** by further partitioning the dataset into a **training set** and a **validation set** and selecting the **hyper-parameter** values that provide the lowest **error rate** on the **validation set**.



Model Evaluation
References

- Daniel Chen. *Pandas for Everyone* (2018).
- Joel Grus. *Data Science from Scratch* (2015).
- Cathy O'Neil and Rachel Schutt. *Doing Data Science* (2013).
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. *Introduction to Data Mining* (2019).