Data preprocessing

Marcin Kuta

Data imputation

- removal of columns or rows with missing values
- imputation of missing values with mean, median or mode
- imputation of missing values with the most fequent values, zero value or random value
- imputation of missing values with k-NN method

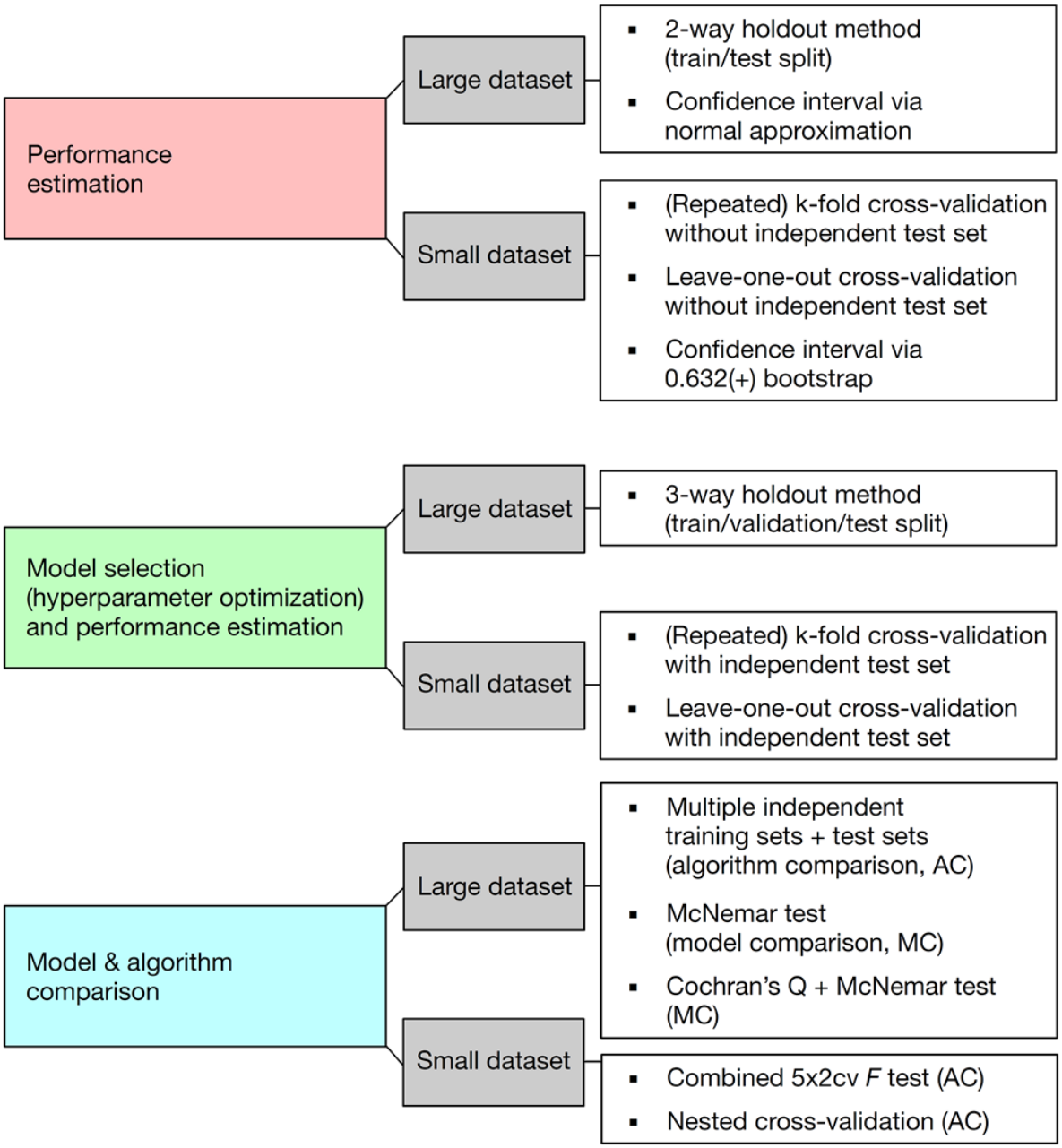
Feature encoding

- one-hot encoding

Feature scaling

- normalization (max-min scaling)
  $x \leftarrow \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$
- standarization
  $x \leftarrow \frac{x - \mu}{\sigma}$
- soft-max scaling
  $y \leftarrow \frac{x - \mu}{r\sigma}$
  $x \leftarrow \frac{1}{1 + \exp(-y)}$

Evaluation


Performance estimation
Large dataset
2-way holdout method (train/test split)
Confidence interval via normal approximation
Small dataset
(Repeated) k-fold cross-validation without independent test set
Leave-one-out cross-validation without independent test set
Confidence interval via 0.632(+) bootstrap
Model selection (hyperparameter optimization) and performance estimation
Large dataset
3-way holdout method (train/validation/test split)
Small dataset
(Repeated) k-fold cross-validation with independent test set
Leave-one-out cross-validation with independent test set
Model & algorithm comparison
Large dataset
Multiple independent training sets + test sets (algorithm comparison, AC)
McNemar test (model comparison, MC)
Cochran's Q + McNemar test (MC)
Small dataset
Combined 5x2cv F test (AC)
Nested cross-validation (AC)

References I

[1] https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.04-Feature-Engineering.ipynb

[2] https://github.com/rasbt/machine-learning-book/tree/main/ch04

[3] Sebastian Raschka,
Model Evaluation, Model Selection and Algorithm Selection in Machine Learning, 2018.