NOVA IMS Information Management School

Neural and Evolutionary Learning

Class 1 - Machine Learning foundations

Prof.: Karina Brotto Rebuli

krebuli@novaims.unl.pt

2025

Algorithms comparision

1. Bias and Variance Balance (Dataset split)

2. Metrics

3. Statistical tests

4. Plots

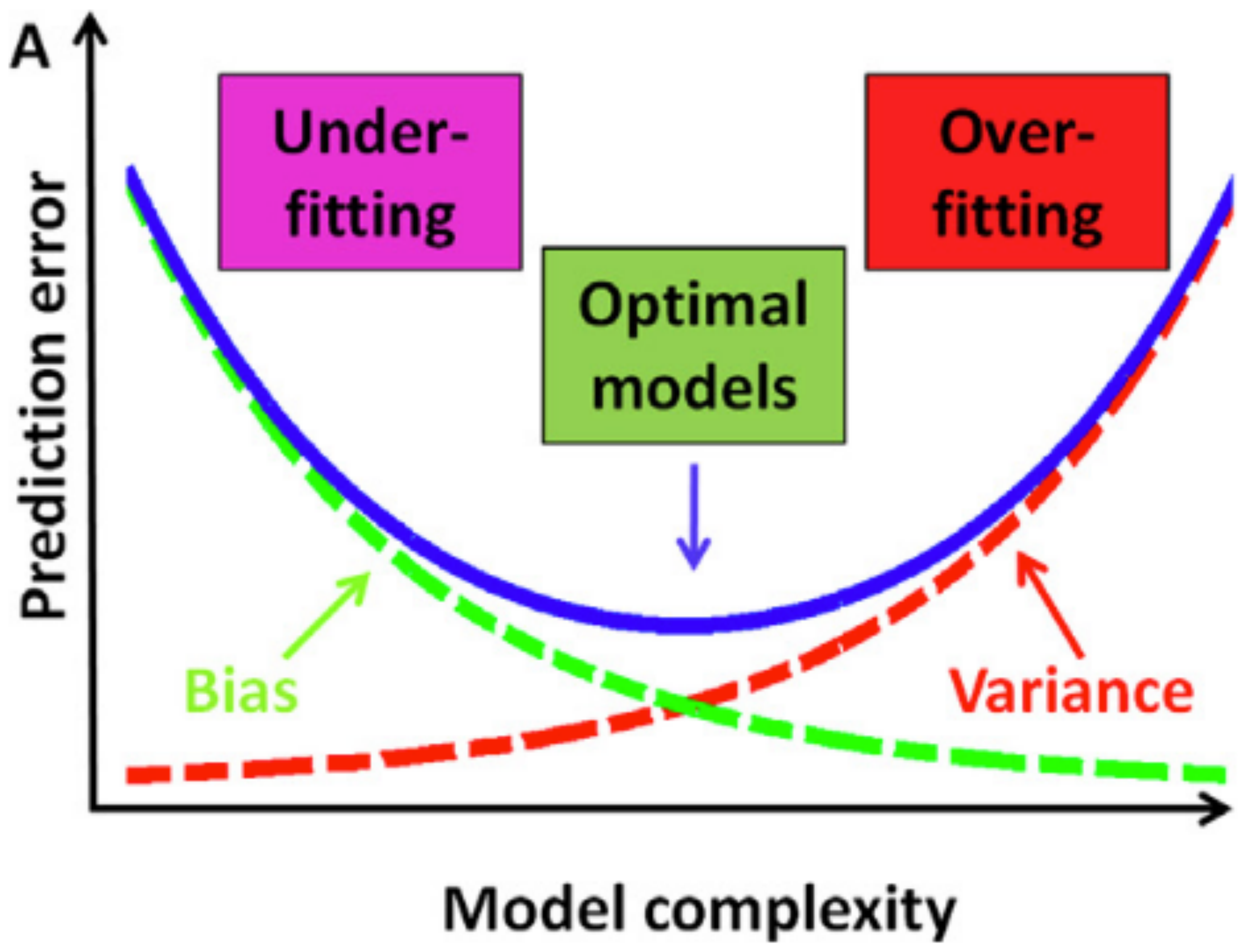
Algorithms comparision

1. Bias and Variance Balance

G. C. Cawley and N. L. C. Talbot. (2010) *On over-fitting in model selection and subsequent selection bias in performance evaluation*, Journal of Machine Learning Research, vol. 11, pp. 2079–2107. https://dl.acm.org/doi/10.5555/1756006.1859921

Algorithms comparision

1. Bias and Variance tradeoff



Source: Deng et al. (2015). *A new strategy to prevent over-fitting in partial least squares models based on model population analysis*, Analytica Chimica Acta, 880, 32-41.

Algorithms comparision

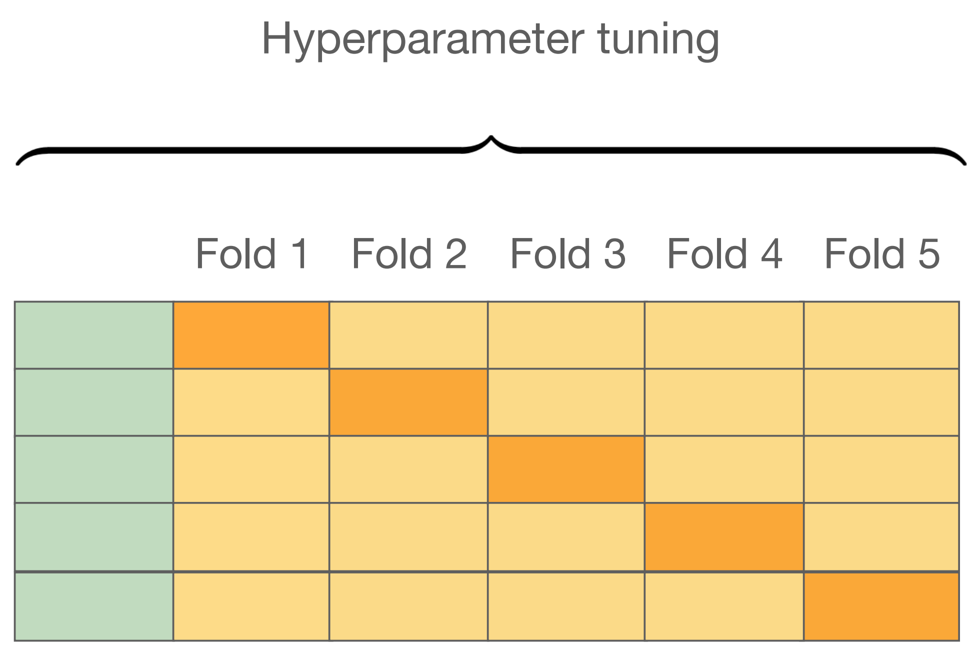
It is always necessary to evaluate the performance of the algorithms on **unseen** data. Thus, the data should be split into the following partitions:

- **Train**: used for the algorithm training (learning) phase;
- **Validation**: used for hyperparameters tuning;
- **Test**: **unseen** data, is used to assess the generalisation ability of the algorithm; therefore, it only should be used after the training and tuning phases, when model architecture and hyperparameters are defined.

Using the test set before evaluating algorithm performance is cheating!!!
And you know… **cheating is strictly prohibited under all circumstances** 🚫

Algorithms comparision

1. Cross-validation (*k*-fold cross-validation)



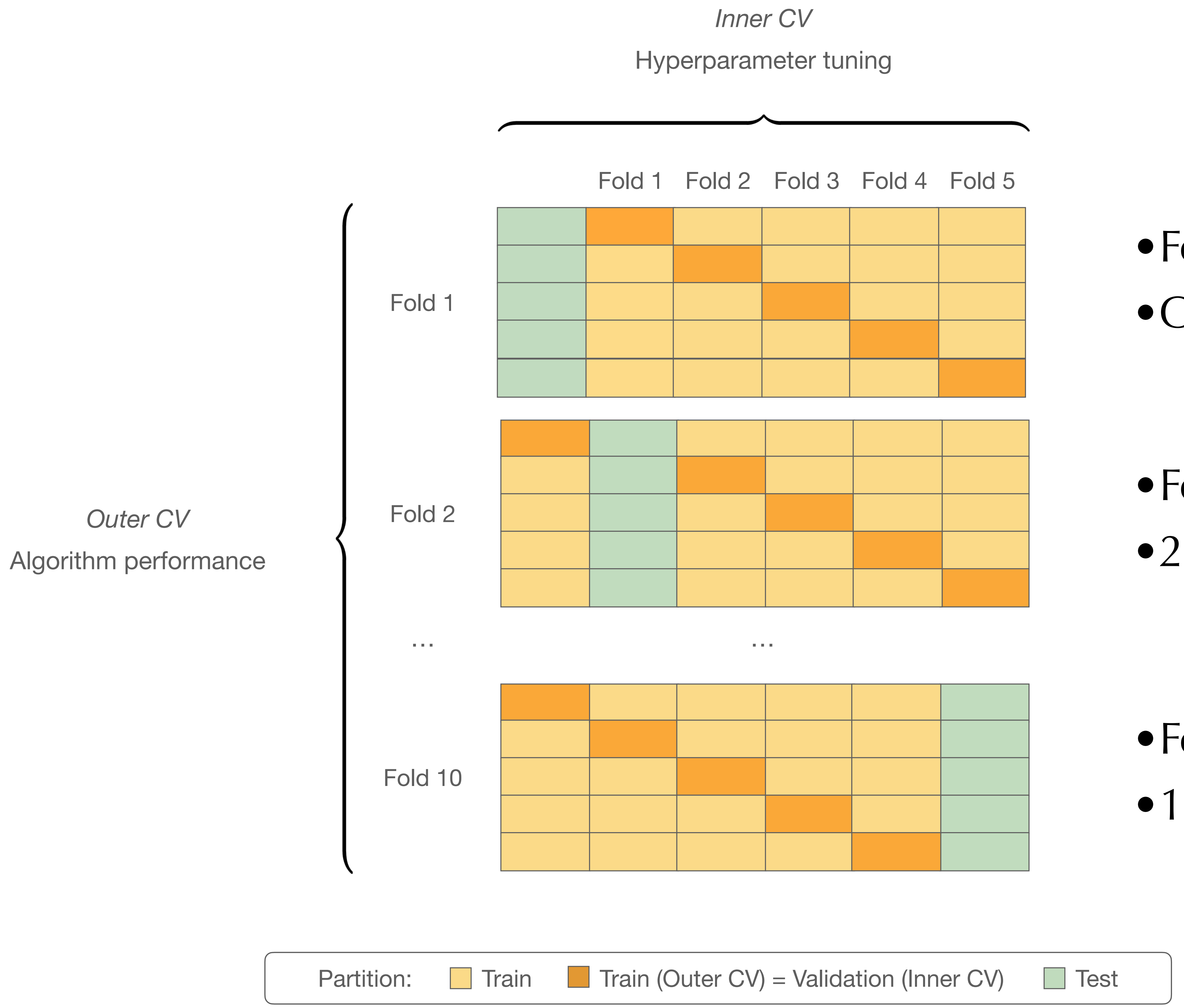
- Folds 1 to 5: best hyperparameters.
- One single Model evaluation on test set.

What conclusions can be taken about the algorithm's performance?

Scientifically speaking, none.

Algorithms comparision

1. Nested Cross-validation (*k*-fold cross-validation)



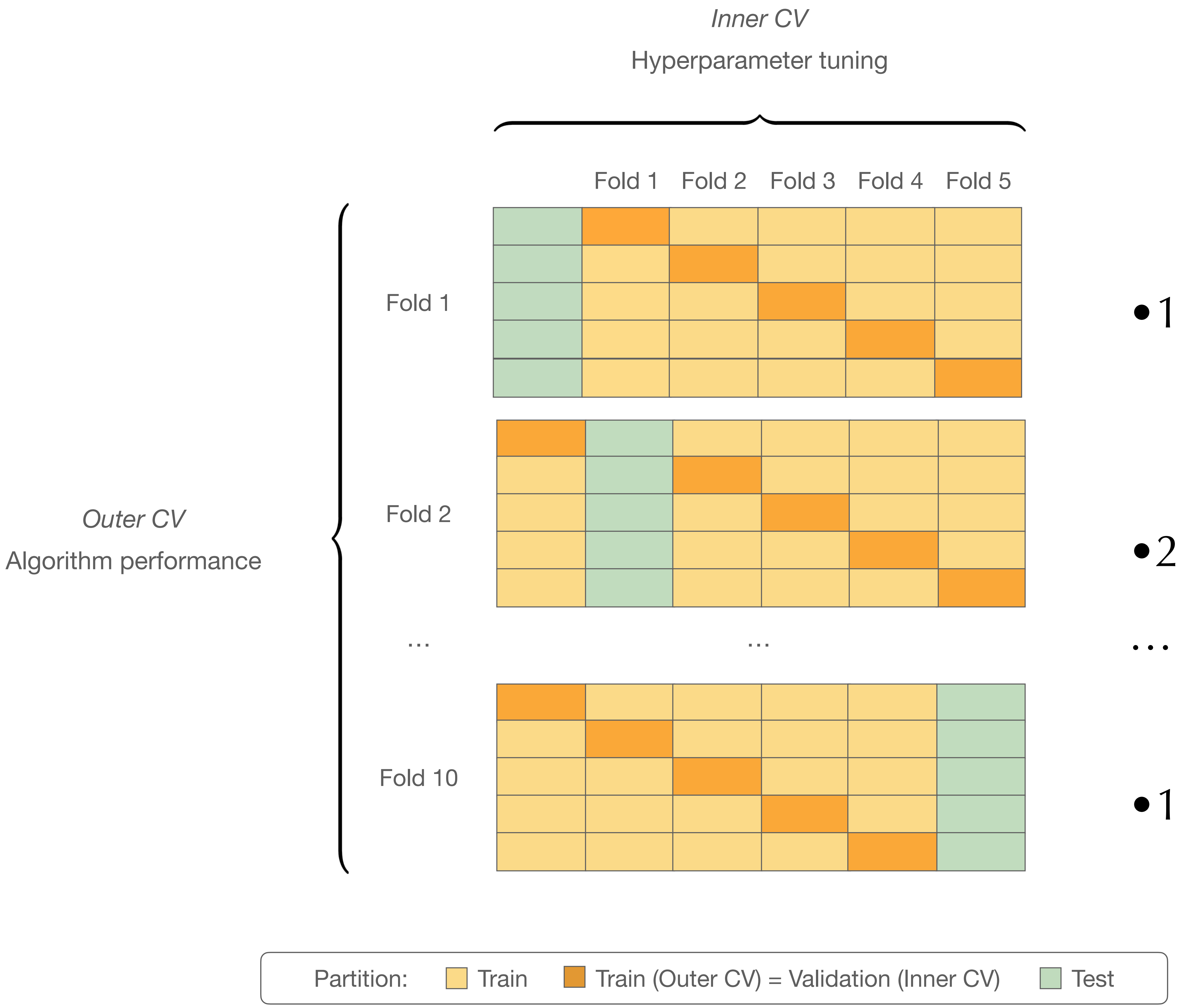
- Folds 1 to 5: best hyperparameters.
- One single Model evaluation on test set.

- Folds 1 to 5: best hyperparameters.
- 2 Model evaluations on test set.

- Folds 1 to 5: best hyperparameters.
- 10 Model evaluations on test set.

Algorithms comparision

1. Nested Cross-validation (*k*-fold cross-validation)



- 1 Model evaluation on test set.
- 2 Model evaluations on test set.

…

- 10 Model evaluations on test set.

Algorithms comparision

1. Nested Cross-validation (*k*-fold cross-validation)

- 1 Model evaluation on test set.
- 2 Model evaluations on test set.

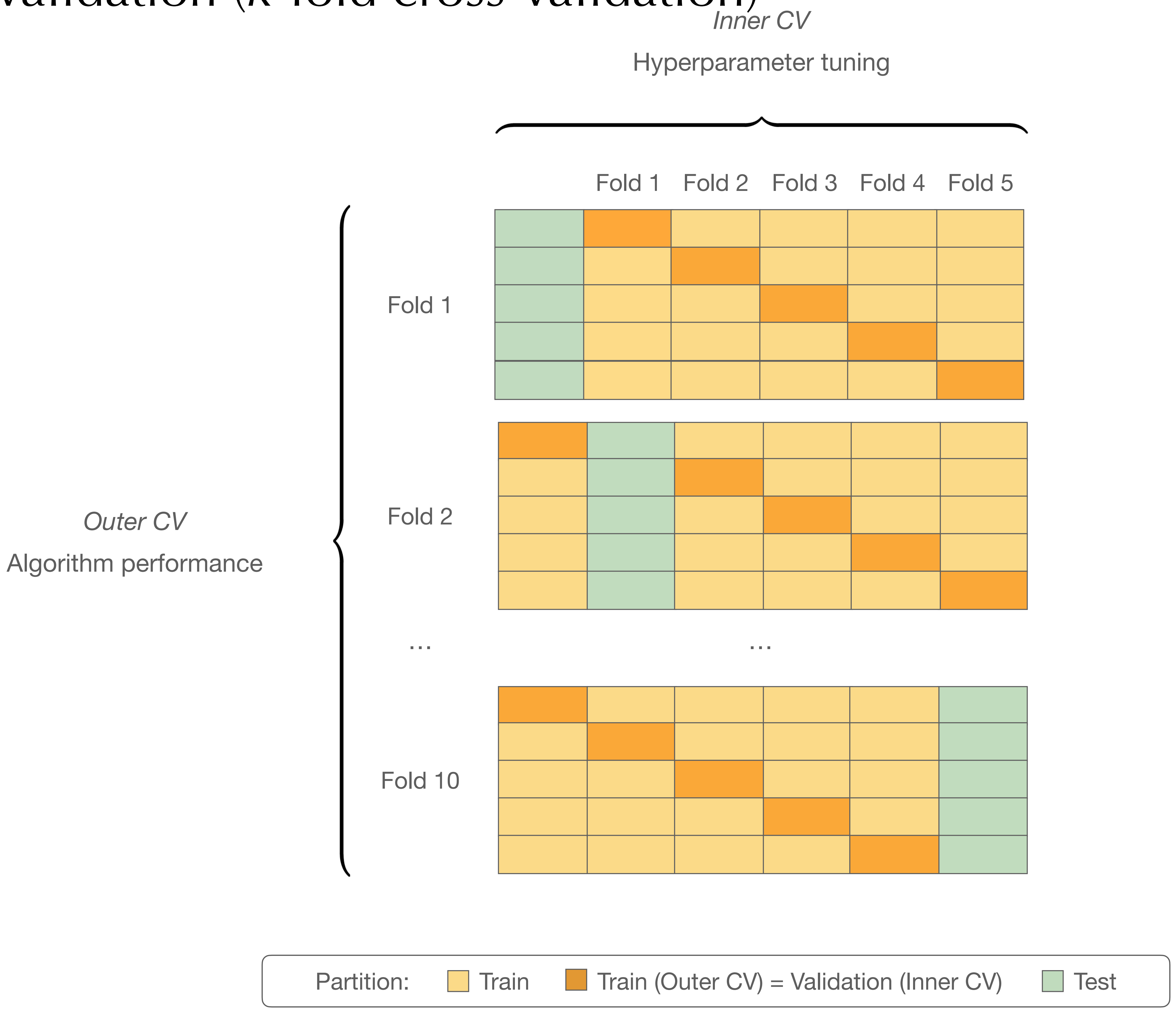
…

- 10 Model evaluations on test set.

Samples for statistical tests.
Scientific conclusions can be discussed.

Algorithms comparision

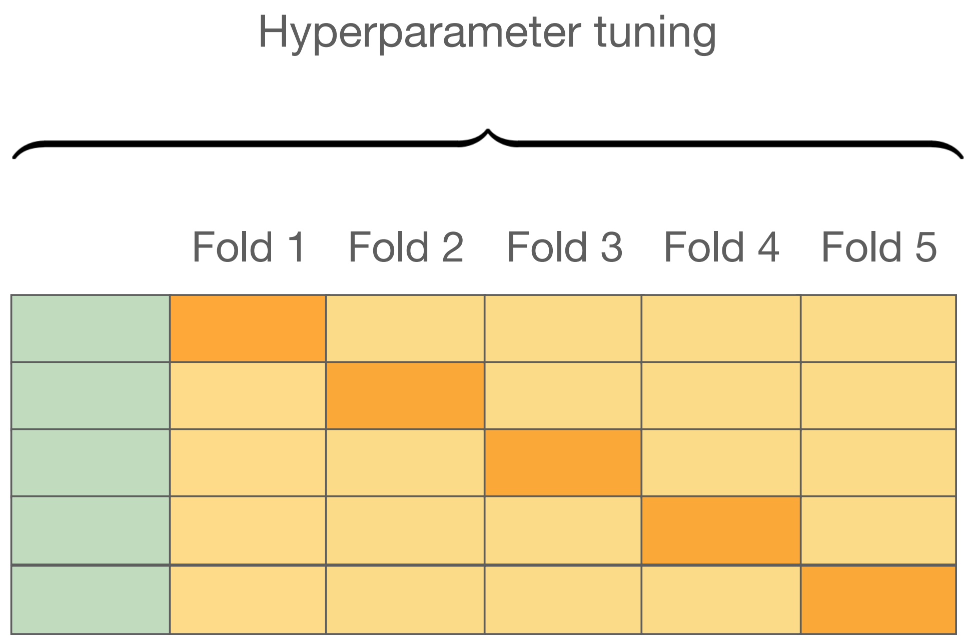
1. Nested Cross-validation (*k*-fold cross-validation)



Drawback: computational time.

Algorithms comparision

1. Cross-validation (*k*-fold cross-validation)



Therefore, if using single-level cross-validation:

- *Justify it;*
- *Discuss the results accordingly;*

Algorithms comparision

2. Metrics for regression problems

- **Error**: MAE (mean absolute error), MSE (mean squared error), RMSE (root mean squared error);
- **Good of fitness**: Pearson correlation coefficient, Spearman correlation coefficient, R^2.

Algorithms comparision

3. Statistical tests

- **Experiment 1**: What would you do to determine whether the age of students in the Information Management course is the same as that of students in the Data Science course in NOVA IMS?

Census

Sampling and Statistical analysis

Algorithms comparision

3. Statistical tests

Sampling and Statistical analysis

1. Define the Hypotheses

Null hypothesis (H_0): The average age of students is the same in both courses.

Alternative hypothesis (H_1): The average age of students is different.

2. Collect Samples

Take a random sample of students from each course (e.g., 30–50 students).

3. Choose a Statistical Test

Set the required significance (p-value) and define the statistical test to be used.

4. Run the test and report results.

Algorithms comparision

3. Statistical tests

2. Collect Samples

Take a random sample of students from each course (e.g., **30–50 students**).

We apply the exact same reasoning for answering the question "Does algorthm A outperforms algorithm B?"

Algorithms comparision

3. Statistical tests

Sampling and Statistical analysis

1. Define the Hypotheses

Null hypothesis (H_0): The average **age** of **students** is the same in both courses.

Alternative hypothesis (H_1): The average **age** of **students** is different.

Algorithms comparision

3. Statistical tests

Sampling and Statistical analysis

1. Define the Hypotheses

Null hypothesis (H_0): The average **RMSE** is the same in both **algorithms**.

Alternative hypothesis (H_1): The average **RMSE** of the **algorithms** is different.

2. Collect Samples

Take a random sample of **students** from each **course** (e.g., 30–50 students).

Algorithms comparision

3. Statistical tests

Sampling and Statistical analysis

1. Define the Hypotheses

Null hypothesis (H_0): The average **RMSE** is the same in both **algorithms**.

Alternative hypothesis (H_1): The average **RMSE** of the **algorithms** is different.

2. Collect Samples

Different **runs** of both **algorithms** (e.g., 30–50 runs).

3. Choose a Statistical Test

Set the required significance (p-value) and define the statistical test to be used.

4. Run the test and report results.

Algorithms comparision

3. Statistical tests

- **Frequentist Tests**:

  The Null Hypothesis is that there is no *real* difference among models' performance, and the Alternative Hypothesis is that there is a *real* difference among models' performances;

  The *p*-value is the probability of having, by chance, a value of the statistic of the test that is equal to or more extreme than the observed value; thus, small p-values mean a low probability of rejecting H_0 when H_0 is true.

  The same training and test sets should be used for all the models to be compared;

Algorithms comparision

3. Non-parametric statistical tests

Two models

- Unpaired data: Mann-Whitney U Test
- **Paired data: Wilcoxon Test**

More than two models

- Unpaired data: Kruskall-Wallis Test, followed by Dunn Test if significative.
- **Paired data: Friedman Test follwoed by Nemenyi Test if significative.**

Algorithms comparision

3. Statistical tests

Rainio, O., Teuho, J. & Klén, R. *Evaluation metrics and statistical tests for machine learning*. Sci Rep 14, 6086 (2024). https://doi.org/10.1038/s41598-024-56706-x

Algorithms comparision

4. Plots

- The cleaner, the better;
- The more standardized, the better;
- Each characteristic of the plot should be used to give the reader a new information;
- Take care with the scale of the axis;
- As results refer to experiments, there is variability in the data. Include this in the plots.
- Example: https://cavalab.org/srbench/results/

Algorithms comparision

Coding time

Questions?

https://forms.gle/EV9VkExNtfNckMSM8
Register your feedback