Course 2

Cross Validation

Developer Student Clubs

Al-Azhar University

AGENDA
CV

01

INTRODUCTION

-What Why Cross validation-

02

CV

- Cross Validation Strategies.

03

IMPLEMENTATION

- Direct Coding

04

Conclusions

- Limitations
- Recommendations

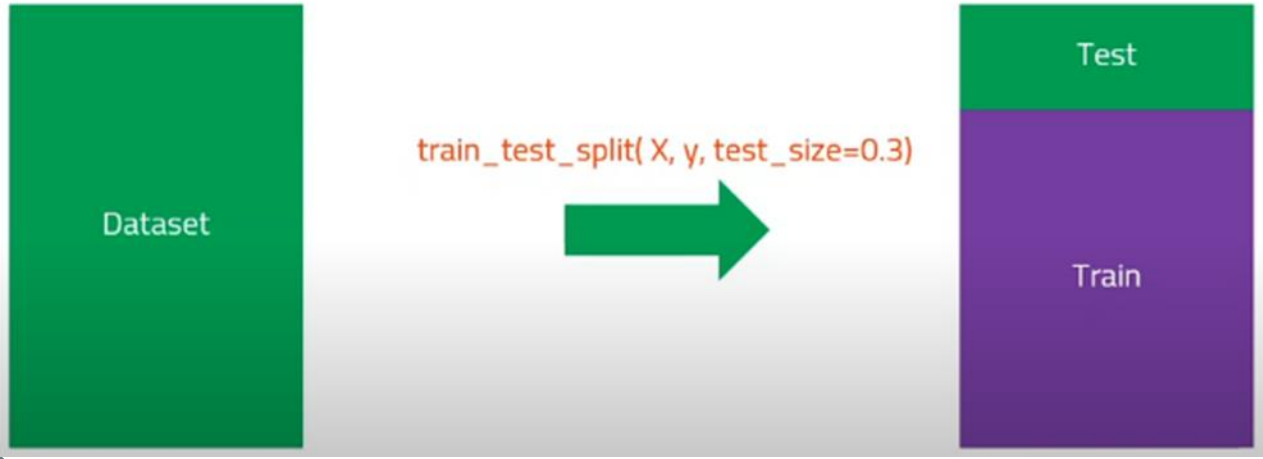
01

INTRODUCTION

What Why

Cross validation

Train Test Split Method



bigger training set

↓

better **learning**

bigger test set

↓

better **testing**

Cross Validation

Cross Validation is a technique which involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it.

Why cross-validate?

bigger training set → better **learning**

bigger test set → better **testing**

Key: Train & test sets must be **disjoint.**
And the dataset or sample size is fixed.
They grow at the expense of each other!

⇨ **cross**-validate to maximize both

02

Cross Validation Strategies.

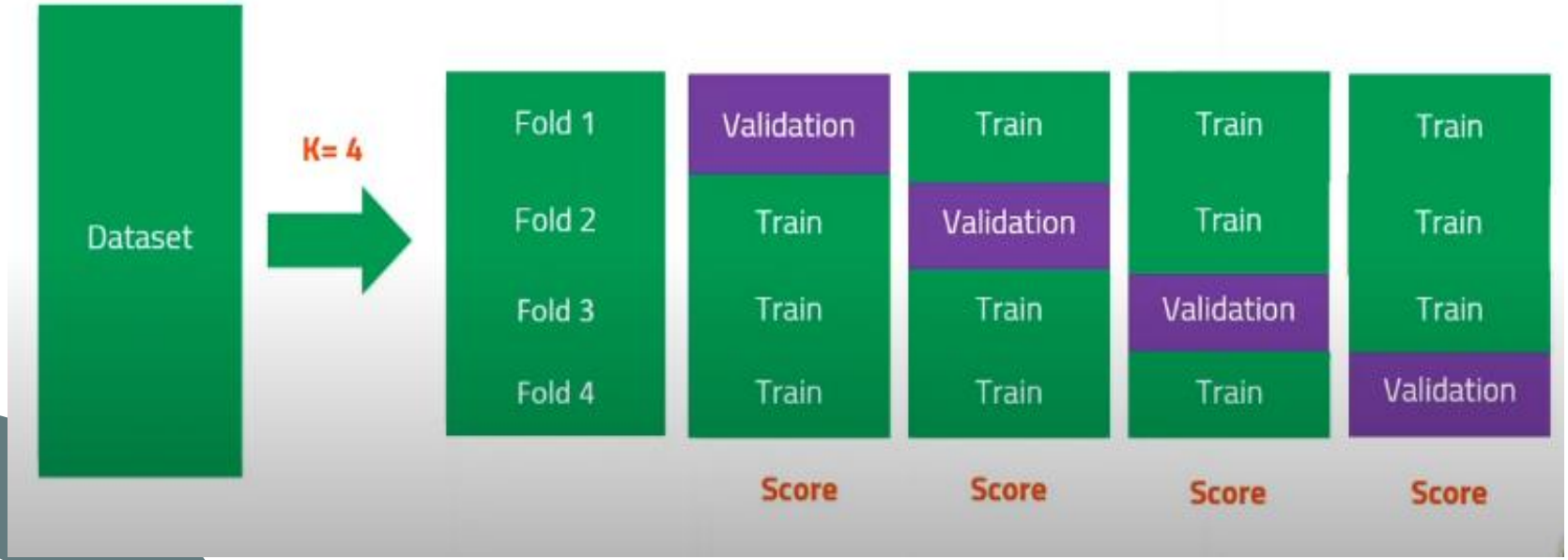
Cross Validation Strategies.

K-fold

Shuffle & Split

Leave-One-Out - LOO

K-Fold Cross Validation

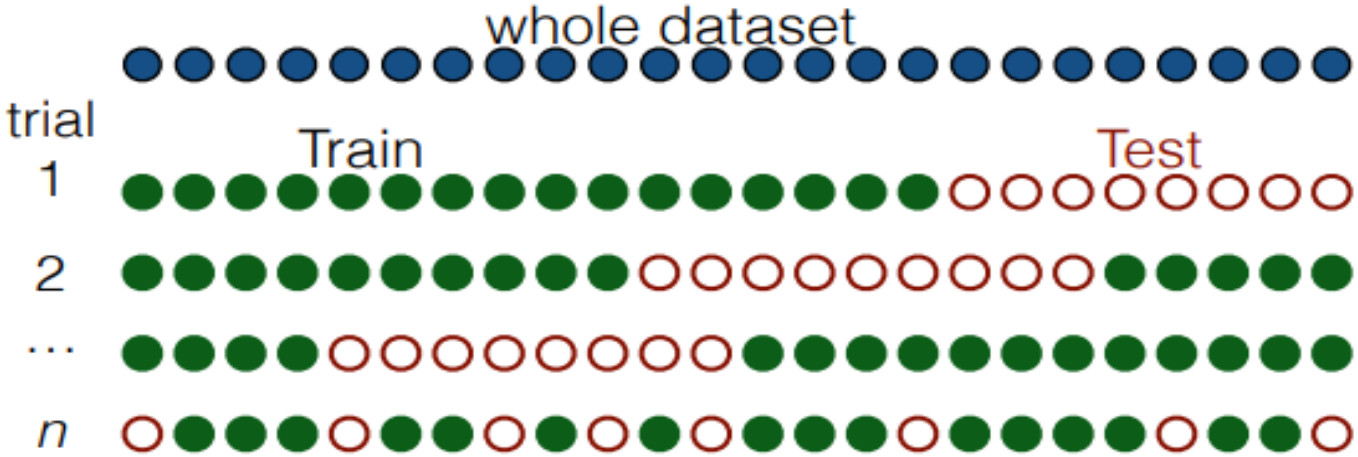


K-Fold Cross Validation

$$\frac{\text{Score} + \text{Score} + \text{Score} + \text{Score}}{K} = \text{Average Score}$$

Repeated Holdout CV

Set aside an independent subsample (e.g. 30%) for testing



Leave one out cross validation (LOOCV)

- In this approach, we reserve only one data point from the available dataset, and train the model on the rest of the data. This process iterates for each data point. This also has its own advantages and disadvantages.

We make use of all data points, hence the bias will be low

We repeat the cross validation process n times (where n is number of data points) which results in a higher execution time

This approach leads to higher variation in testing model effectiveness because we test against one data point. So, our estimation gets highly influenced by the data point. If the data point turns out to be an outlier, it can lead to a higher variation

03

IMPLEMENTATION

04

Conclusions

Limitations of CV

- Number of CV repetitions increases with
  - sample size: • large sample —> large number of repetitions
  - esp. if the model training is computationally expensive.
- number of model parameters, exponentially
  - to choose the best combination!

Recommendations

- Ensure the test set is truly independent of the training set!
- easy to commit mistakes in complicated analyses!
- Use repeated-holdout (10-50% for testing)
  - respecting sample/dependency structure
- ensuring independence between train & test sets
  - Use biggest test set, and large # repetitions when possible
  - Not possible with leave-one-sample-out.

Conclusions

- Results could vary considerably with a different CV scheme
- CV results can have variance (>10%)
- Document CV scheme in detail:
    - type of split
    - number of repetitions
    - Full distribution of estimates
- Proper splitting is not enough, proper pooling is needed too.

PARCTICE

. Your Turn Now

THANKS

Any questions?

Developer Student Clubs

Al-Azhar University