Cite as: Ioannis Tsamardinos, Vincenzo Lagani, Automated Machine Learning and Knowledge Discovery, ECCB 2018 Tutorial;

Slides are a modified excerpt from the ECCB 2018 tutorial

Automated Machine Learning and Knowledge Discovery

IOANNIS TSAMARDINOS
PROFESSOR, CSD, UNIVERSITY OF CRETE
GNOSIS DATA ANALYSIS, CO-FOUNDER

VINCENZO LAGANI
ILIA STATE UNIVERSITY
GNOSIS DATA ANALYSIS, CO-FOUNDER

A fictional execution trace of Nested Cross Validation (NCV)

- Choose among Pipelines (Learners) *a, b*
- Split data to 3 Folds named 1, 2, 3
- All accuracies reported as fictional estimates

- Inner Cross-Validation loop: cross-validation of a single pipeline to determine the winning pipeline
- Outer Cross-Validation loop: cross-validation of the selecting-the-winner procedure to determine the predictive performance of the winning pipelines

NCV Trace: Model Production

- Choose among Pipelines (learners) *a, b*
- Split data to Folds 1, 2, 3

Model Performances on held out fold

Train On	With Pipeline	Produce	Apply on	Accuracy
1, 2	a	M_1	3	0.7
1, 3	a	M_2	2	0.8
2, 3	a	M_3	1	0.6
				Mean_a = 0.7
1, 2	b	M_4	3	0.6
1, 3	b	M_5	2	0.7
2, 3	b	M_6	1	0.5
				Mean_b = 0.6
Select **a** as winning				
1, 2, 3	a	M_7	**N/A**	

No direct estimation of the performance of the returned model.
No loss of data to estimation

Final Model to return using Cross Validation with Tuning (CVT): M_7

NCV Trace: Performance Estimation (1of 3)

- Outer loop iteration 1
- Fold 3 is held-out as an Estimation set ; the other folds serve as Tune sets in the inner CV loop.

Train On	With Pipel.	Produce	Apply on	Accuracy
1	a	M_8	2	0.7
2	a	M_9	1	0.8
				Mean$_a$ = 0.75
1	b	M_{10}	2	0.6
2	b	M_{11}	1	0.7
				Mean$_a$ = 0.65
Select a as winning				
1, 2	a	M_{12}	**3**	**0.9**

Folds 1 and 2 (inner loop folds) serve multiple times as Tune sets

Fold 3 (outer loop folds) serves a single time as an Estimation set.

NCV Trace: Performance Estimation (2 of 3)

- Outer loop iteration 2
- Fold 2 is held-out as an Estimation set ; the other folds serve as Tune sets in the inner CV loop.

Train On	With Pipel.	Produce	Apply on	Accuracy
1	a	M_{13}	3	0.6
3	a	M_{14}	1	0.7
				Mean$_a$ = 0.65
1	b	M_{15}	3	0.7
3	b	M_{16}	1	0.8
				Mean$_a$ = 0.75
Select b				
1, 3	b	M_{17}	**2**	**0.7**

NCV Trace: Performance Estimation (3 of 3)

- Outer loop iteration 3
- Fold 1 is held-out as an Estimation set ; the other folds serve as Tune sets in the inner CV loop.

Train On	With Pipel.	Produce	Apply on	Accuracy
2	a	M_{18}	3	0.8
3	a	M_{19}	2	0.6
				Mean$_a$ = 0.7
2	b	M_{20}	3	0.6
3	b	M_{21}	2	0.6
				Mean$_a$ = 0.6
Select a				
2, 3	a	M_{22}	**1**	**0.8**

Final Estimate to return is the Learner performance : mean accurate on Estimation folds over three iterations 0.9 + 0.7 + 0.8 = **0.8**

How many models trained?

C: number of pipelines

K: number of folds

- To produce the final model the inner CV loop is called with K folds
  - C pipelines × K folds for estimating the winning pipeline
  - +1 times to train on the full dataset
  - $= C \times K + 1$
- To estimate the performance of the returned model
  - Run the inner CV with K-1 folds, K times
  - $= (C \times (K-1) + 1) \times K$
- **Total number of models trained for model production and estimation**
- $\mathbf{= C \times K^2 + K + 1}$