Two-phase and double machine learning: with application and experiment

Li-Chun Zhang[1,2,3] **and** ***Susie Jentoft***[2]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***
[2] ***Statistisk sentralbyraa, Norway***
[3] ***Universitetet i Oslo***

Supervised machine learning (ML)

Sample of observations: $s = \{1, ..., n\}$

q-category outcome: dummy indexed $\boldsymbol{y}_k$, for $k \in s$

Feature: relevant p-covariates $\boldsymbol{x}_k$, for $k \in s$

Suppose training and validation sets: $s_1 \cup s_2 = s$

Supervised ML, e.g. random forrest (RF), based on s_1 yields a *classifier*

$$\hat{\boldsymbol{y}}_k = g(\boldsymbol{x}_k) \in \{0, 1\}^q \qquad \text{where} \quad \hat{\boldsymbol{y}}_k^\top \mathbf{1} = 1$$

and, at the same time, a *predictor*

$$\hat{\boldsymbol{\mu}}_k = h(\boldsymbol{x}_k) \in [0, 1]^q \qquad \text{where} \quad \hat{\boldsymbol{\mu}}_k^\top \mathbf{1} = 1$$

Accuracy of supervised ML

Confusion matrix based on validation set s_2

$q = 2$	Observed binary outcome y_k	
Classifier $\hat{y}_k$	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

Accuracy = (TP + TF)/n $\mapsto$ unconditional probability

$$\Pr(\delta_k = 1) \qquad \text{where} \quad \delta_k = \begin{cases} 1 & \text{if } y_k = \hat{y}_k \\ 0 & \text{if } y_k \neq \hat{y}_k \end{cases}$$

A more useful measure is the conditional probability

$$\eta_k = \Pr(\delta_k = 1 | \boldsymbol{x}_k, \boldsymbol{z}_k)$$

with possible additional features $\boldsymbol{z}_k$, for $i \in s$

Two-phase ML (2ML)

Based on different units in each phase:

- Phase 1: learn classifier $\hat{\boldsymbol{y}}_k$ based on $\{\boldsymbol{y}_k; k \in s_1\}$

  NB. can aim at predictor $\hat{\boldsymbol{\mu}}_k$ instead

- Phase 2: learn predictor $\hat{\eta}_k$ based on $\{\delta_k; k \in s_2\}$

Can target other accuracy measures, e.g. $y_k - \hat{y}_k$:

- unbiased $\hat{y}_k$ if FP = FN by confusion matrix, i.e.

$$\Pr(y_k = 0, \hat{y}_k = 1) = \Pr(y_k = 1, \hat{y}_k = 0)$$

- 2ML examines conditional unbiasedness, provided

$$\Pr(y_k = 0, \hat{y}_k = 1 | \boldsymbol{x}_k, \boldsymbol{z}_k) = \Pr(y_k = 1, \hat{y}_k = 0 | \boldsymbol{x}_k, \boldsymbol{z}_k)$$

An application of two-phase RF (2RF)

Observation of *partial absence* in Labour Force Survey (LFS)

- direct interview y_k for $k \in s^{direct}$ [NB. binary y_k]
- proxy interview $\tilde{y}_k$ for $k \in s^{proxy}$, where $\tilde{y}_k \neq y_k$ for some units

Q: is it better to impute by RF instead, for s^{proxy}?

- RF on $s_1^{direct} \cup s_2^{direct}$: unbiased predictor, biased classifier
- 2nd phase RF on s^{proxy}: negative bias of proxy observation $\tilde{y}_k$
- A test for H_0: unbiased $\tilde{y}_k$, provided MAR in $(s_1, ..., s_G)$:

$$t = \Delta^{proxy} - \Delta^{RF} \quad \text{where} \quad \begin{cases} \Delta^{proxy} = \sum_{g=1}^{G} w_g^{proxy} \tilde{\bar{y}}_g - \sum_{g=1}^{G} w_g^{direct} \bar{y}_g \\ \Delta^{RF} = \sum_{g=1}^{G} w_g^{direct} \hat{\bar{\mu}}_g - \sum_{g=1}^{G} w_g^{direct} \bar{y}_g \end{cases}$$

$$\Rightarrow \quad t = \sum_{g=1}^{G} w_g^{proxy} \tilde{\bar{y}}_g - \sum_{g=1}^{G} w_g^{direct} \hat{\bar{\mu}}_g = t_1(s^{proxy}) - t_2(s^{direct})$$

$$t = -0.0298, \ \mathrm{SE}(t) = 0.0048 \quad \Rightarrow \quad \text{Reject } H_0 \text{ at size } 0.05$$

An application of two-phase RF (2RF)

Group (g)	$\bar{y}_g^{dir}$	$\bar{y}_g^{proxy}$	$\bar{\mu}_g^{ML}$	n_g^{dir}	n_g^{proxy}	$\Delta_g(dir, proxy)$	$\Delta_g(dir, ML)$	Best
0_0_1_fast	0.13	0.09	0.11	7561	2204	-0.0346	-0.0161	ML
0_0_1_other	0.07	0.05	0.05	905	403	-0.0186	-0.0199	proxy
0_0_1_time	0.09	0.08	0.09	1319	597	-0.0170	-0.0035	ML
0_0_2_fast	0.10	0.09	0.13	968	470	-0.0077	0.0273	proxy
0_0_2_other	0.07	0.04	0.06	189	121	-0.0328	-0.0161	ML
0_0_2_time	0.05	0.03	0.06	428	263	-0.0195	0.0081	ML
0_0_3_fast	0.05	0.02	0.08	496	282	-0.0276	0.0257	ML
0_0_3_other	0.02	0.01	0.06	297	337	-0.0109	0.0436	proxy
0_0_3_time	0.03	0.01	0.04	569	709	-0.0197	0.0144	ML
0_1_1_fast	0.06	0.04	0.07	747	239	-0.0212	0.0128	ML
0_1_1_other	0.05	0.02	0.08	106	46	-0.0254	0.0370	proxy
0_1_1_time	0.05	0.02	0.07	135	95	-0.0308	0.0177	ML
0_1_2_fast	0.04	0.00	0.10	97	42	-0.0412	0.0596	proxy
0_1_2_other	0.00	0.00	0.04	16	14	0.0000	0.0435	proxy
0_1_2_time	0.02	0.03	0.07	43	40	0.0017	0.0489	proxy
0_1_3_fast	0.02	0.00	0.08	50	22	-0.0200	0.0573	proxy
0_1_3_other	0.00	0.00	0.06	22	19	0.0000	0.0626	proxy
0_1_3_time	0.08	0.02	0.06	40	58	-0.0578	-0.0160	ML
1_0_1_fast	0.50	0.47	0.52	1106	322	-0.0342	0.0166	ML
1_0_1_other	0.28	0.18	0.24	109	56	-0.1058	-0.0432	ML
1_0_1_time	0.36	0.46	0.43	181	90	0.0964	0.0660	ML
1_0_2_fast	0.26	0.36	0.37	155	83	0.0969	0.1074	proxy
1_0_2_other	0.27	0.40	0.27	22	10	0.1273	-0.0006	ML
1_0_2_time	0.14	0.19	0.21	50	32	0.0475	0.0664	proxy
1_0_3_fast	0.14	0.09	0.29	83	44	-0.0537	0.1500	proxy
1_0_3_other	0.09	0.04	0.22	34	45	-0.0438	0.1282	proxy
1_0_3_time	0.11	0.07	0.15	72	100	-0.0411	0.0366	ML

Double ML (DML)

2ML tells us how well ML does conditionally in a given situation, but not when an ML-technique can be expected to perform well.

[NB. e.g. linear regression can be expected to do well if so and so...]

[NB. e.g. a typical scenario: tried logistic regression, RF, and SVM, and RF looks the best — great, but what about next application?]

DML aims to learn the circumstances under which an ML-technique can be expected to do well, by incorporating *feedbacks* and *context*.

Feedback: about the outcome of classifier/predictor, e.g.

$$\text{pred\_good}_k = |2\hat{\mu}_k - 1| \quad \text{or} \quad \text{pred\_good}_k = |2\hat{\mu}_k - \bar{y}_s|$$

Context: characteristics at the dataset level, e.g. occurrence $\bar{y}_s$, pseudo R^2 from default (main-effects) logistic regression, etc.

Double ML (DML)

Stage-1: $s^{(1)}, ..., s^{(J)}$ applications of an ML-technique

[incl. when $y_k^{(1)}, ..., y_k^{(J)}$ are different outcomes in the same sample]

Stage-2: set $s = \bigcup_{j=1}^{J} s^{(j)}$; do ML of δ_k with features $\boldsymbol{x}_k$, feedbacks $\boldsymbol{z}_k$ and contextual variables $\boldsymbol{u}_k$, for $k \in s$

Alt. as an example of more pragmatic goal for DML, let

$$\delta_k = \begin{cases} 1 & \text{if } |\hat{\mu}_k^{ML} - y_k| \leq |\hat{\mu}_k^{Reg} - y_k| \\ 0 & \text{if } |\hat{\mu}_k^{ML} - y_k| > |\hat{\mu}_k^{Reg} - y_k| \end{cases}$$

Key challenge: which relevant feedbacks and context?

An experiment of DML with LFS data

Five variables: unemp., emp., partial absence, overtime, temp. job

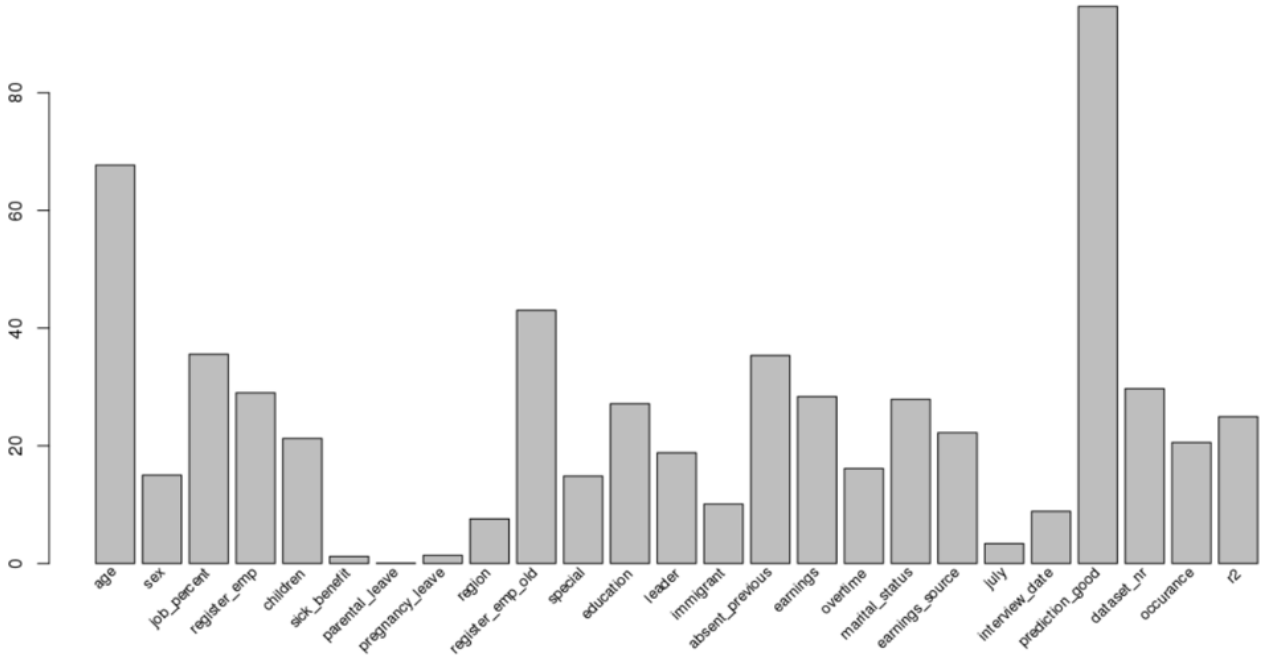
Stage-1: separate RF with (2/3, 1/3) training-validation partition

Stage-2: RF of $\delta(\hat{y}_k)$ with all stage-1 features $\boldsymbol{x}_k$, feedback pred_good$_k$ and contextual variables occurrence, pseudo R^2, dataset identifier j

Outcome	Occurrence	Pseudo R^2	Mean pred_good$_k$
Unemployment	0.03	0.2348	0.9645
Partial absence	0.10	0.1445	0.8313
Employment	0.73	0.8298	0.9325
Overtime	0.07	0.1032	0.8749
Temporary job	0.05	0.0972	0.8998

	Accuracy	$\frac{\text{FN}}{\text{FN+TP}}$	$\frac{\text{TN}}{\text{TN+FP}}$	$\hat{\bar{\delta}}$, Stage 2	$\bar{\delta}$, Stage 1
Unemployment	0.98	< 0.01	0.02	> 0.99	0.98
Partial absence	0.98	0.01	0.07	0.99	0.90
Employment	0.99	< 0.01	0.07	> 0.99	0.96
Overtime	0.99	< 0.01	0.02	> 0.99	0.93
Temporary job	0.99	< 0.01	0.03	> 0.99	0.95

An experiment of DML with LFS data



Contextual variables used: not very helpful

Most important covariate: pred_good$_k$
[potentially useful finding, despite it seems quite obvious]

Ongoing research! *Ideas?*