Academy of PhD Training in Statistics

Statistical Machine Learning

Louis J.M. Aslett (`louis.aslett@durham.ac.uk`)

Model Assessment and Combination

Durham University

This Section

- Baseline models
- Cut-off selection
- Binary metrics
- Calibration
- Learning curves
- Additional topics
- Super learners

This Section

- Baseline models
- Cut-off selection
- Binary metrics
- Calibration
- Learning curves
- Additional topics
- Super learners

Coming discussion rooted in binary classification, but much applies generally.

Evaluating applied models

- Working in applied problems requires deep thought about the subject matter problem

Evaluating applied models

- Working in applied problems requires deep thought about the subject matter problem
- Are you optimising for the right quantity?

Evaluating applied models

- Working in applied problems requires deep thought about the subject matter problem
- Are you optimising for the right quantity?
- Don’t blindly examine accuracy!

Evaluating applied models

- Working in applied problems requires deep thought about the subject matter problem
- Are you optimising for the right quantity?
- Don't blindly examine accuracy!
  - Especially accuracy with a default 0.5 cutoff!

Evaluating applied models

- Working in applied problems requires deep thought about the subject matter problem
- Are you optimising for the right quantity?
- Don't blindly examine accuracy!
  - Especially accuracy with a default 0.5 cutoff!
  - Beware advice to over/undersampe to rescue this often inappropriate metric!

Evaluating applied models

- Working in applied problems requires deep thought about the subject matter problem
- Are you optimising for the right quantity?
- Don't blindly examine accuracy!
  - Especially accuracy with a default 0.5 cutoff!
  - Beware advice to over/undersampe to rescue this often inappropriate metric!
- Do worry about calibration and where possible make probabilistic forecasts, rather than raw labels alone

Baseline models

Always include a very simple model in your analysis to act as a baseline.

eg. "featureless" model:

$$\hat{f}_j(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i \in \mathcal{T}_r} \mathbb{1}\{y_i = j\} > \sum_{i \in \mathcal{T}_r} \mathbb{1}\{y_i = k\} \; \forall \, k \\ 0 & \text{otherwise} \end{cases}$$

Baseline models

Always include a very simple model in your analysis to act as a baseline.

eg. "featureless" model:

$$\hat{f}_j(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i \in \mathcal{T}_r} \mathbb{1}\{y_i = j\} > \sum_{i \in \mathcal{T}_r} \mathbb{1}\{y_i = k\} \; \forall \, k \\ 0 & \text{otherwise} \end{cases}$$

Or at least a standard logistic regression. You'll be amazed how often it's good enough! (Christodoulou et al., 2019)

Baseline models

Always include a very simple model in your analysis to act as a baseline.

eg. "featureless" model:

$$\hat{f}_j(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i \in \mathcal{T}_r} \mathbb{1}\{y_i = j\} > \sum_{i \in \mathcal{T}_r} \mathbb{1}\{y_i = k\} \ \forall \, k \\ 0 & \text{otherwise} \end{cases}$$

Or at least a standard logistic regression. You'll be amazed how often it's good enough! (Christodoulou et al., 2019)

Baseline model can really save you from unwarrented excitement (eg accuracy)

Cut-off selection

Under 0-1 loss, Bayes optimally:

$$g^\star(\mathbf{x}) = \arg \max_{j \in \{0,1\}} f_j(\mathbf{x}) = \begin{cases} 0 & \text{if } f_1(\mathbf{x}) < 0.5 \\ 1 & \text{if } f_1(\mathbf{x}) \geq 0.5 \end{cases}$$

Cut-off selection

Under 0-1 loss, Bayes optimally:

$$g^\star(\mathbf{x}) = \arg\max_{j \in \{0,1\}} f_j(\mathbf{x}) = \begin{cases} 0 & \text{if } f_1(\mathbf{x}) < 0.5 \\ 1 & \text{if } f_1(\mathbf{x}) \geq 0.5 \end{cases}$$

But, in general for other losses:

$$g_{\hat{f}}(\mathbf{x}) = \begin{cases} 0 & \text{if } \hat{f}_1(\mathbf{x}) < \alpha \\ 1 & \text{if } \hat{f}_1(\mathbf{x}) \geq \alpha \end{cases}$$

(and often practicioners prefer to think in terms of related performance measures rather than trying to specify a loss directly)

Standard binary performance measures

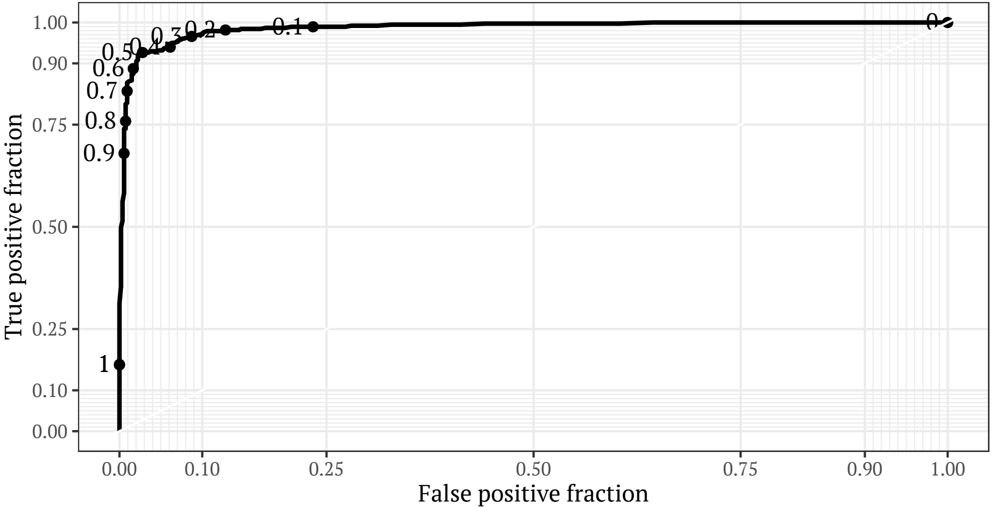
- True positive (TP) rate (aka sensitivity or recall): is the conditional probability of correctly predicting 1 given true label 1.
- True negative (TN) rate (aka specificity): is the conditional probability of correctly predicting a 0 given true label 0.
- False positive (FP) rate: probability of Type I error (ie 1-specificity).
- False negative (FN) rate: probability of Type II error (ie 1-sensitivity).
- Positive predictive value (PPV) (aka precision): is the conditional probability of the true label being 1 given a prediction of 1.
- Negative predictive value (NPV): is the conditional probability of the true label being 0 given a prediction of 0.

Standard binary performance measures

<table>
<tr><td colspan="2"></td><td colspan="2">True condition</td><td colspan="3"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR–}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR–) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Source: Wikipedia

ROC/AUC


True positive fraction
False positive fraction
1.00
0.90
0.75
0.50
0.25
0.10
0.00
0.00
0.10
0.25
0.50
0.75
0.90
1.00
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Calibration

Informally: we want the probabilistic forecasts from our model to have good frequency properties …

Calibration

Informally: we want the probabilistic forecasts from our model to have good frequency properties …

… if we predict '1' with probability 0.6 at some $\mathbf{x}$, does this happen 60% of the time?

Calibration

Informally: we want the probabilistic forecasts from our model to have good frequency properties …

… if we predict '1' with probability 0.6 at some x, does this happen 60% of the time?

- Calibration-in-the-large
- Weak calibration
- Moderate calibration
- Strong calibration

Cox (1958)

David Cox (1958) proposed logistic regression to assess agreement between observed binary events and probabilities.

Cox (1958)

David Cox (1958) proposed logistic regression to assess agreement between observed binary events and probabilities.

If model predicts '1' with probability $\hat{f}(\mathbf{x})$ and that probability is accurate, then a logistic regression:

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)}\right) = \beta_0 + \beta_1 \log\left(\frac{\hat{f}(\mathbf{x}_i)}{1 - \hat{f}(\mathbf{x}_i)}\right)$$

should have $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$.

Cox (1958)

David Cox (1958) proposed logistic regression to assess agreement between observed binary events and probabilities.

If model predicts '1' with probability $\hat{f}(\mathbf{x})$ and that probability is accurate, then a logistic regression:

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)}\right) = \beta_0 + \beta_1 \log\left(\frac{\hat{f}(\mathbf{x}_i)}{1 - \hat{f}(\mathbf{x}_i)}\right)$$

should have $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$.

In other words, produce out-of-sample predictions from the model and perform a logistic regression of the true response y_i against a univariate derived feature $\log\left(\frac{\hat{f}(\mathbf{x}_i)}{1-\hat{f}(\mathbf{x}_i)}\right)$.

Cox (1958)

David Cox (1958) proposed logistic regression to assess agreement between observed binary events and probabilities.

If model predicts '1' with probability $\hat{f}(\mathbf{x})$ and that probability is accurate, then a logistic regression:

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = 0)}\right) = \beta_0 + \beta_1 \log\left(\frac{\hat{f}(\mathbf{x}_i)}{1 - \hat{f}(\mathbf{x}_i)}\right)$$

should have $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$.

In other words, produce out-of-sample predictions from the model and perform a logistic regression of the true response y_i against a univariate derived feature $\log\left(\frac{\hat{f}(\mathbf{x}_i)}{1-\hat{f}(\mathbf{x}_i)}\right)$.

Could perform hypothesis tests for $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 1$.

Calibration-in-the-large

Simplest and easiest form of calibration for a model to satisfy.

Require,

$$\frac{1}{n}\sum_{i=1}^{n} y_i \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\mathbf{x}_i)$$

Calibration-in-the-large

Simplest and easiest form of calibration for a model to satisfy.

Require,

$$\frac{1}{n}\sum_{i=1}^{n} y_i \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\mathbf{x}_i)$$

Either, compute informally, or use the method of Cox (1958) with β_1 fixed at 1 (ie only estimate $\hat{\beta}_0$)

Weak calibration

Is the name given to the procedure of Cox (1958), with both slope and intercept fitted and tested.

Weak calibration

Is the name given to the procedure of Cox (1958), with both slope and intercept fitted and tested.

- If the slope condition is satisfied, but β_0 is not plausibly zero and: - is negative, this corresponds to general overestimation of probabilities for the event '1'.
    - is positive, this corresponds to general underestimation of probabilities for the event '1'.

Weak calibration

Is the name given to the procedure of Cox (1958), with both slope and intercept fitted and tested.

- If the slope condition is satisfied, but β_0 is not plausibly zero and: - is negative, this corresponds to general overestimation of probabilities for the event '1'.
    - is positive, this corresponds to general underestimation of probabilities for the event '1'.
- If the intercept condition is satisfied, but β_1 is not plausibly 1 and:
    - is smaller than 1, this corresponds to probabilities being pushed out to extremities (ie probabilities for the event '1' are too large, and probabilities for the event '0' are too small).
    - is larger than 1, this corresponds to probabilities being too under dispersed (ie probabilities for the event '1' are too small, and probabilities for the event '0' are too large).

Moderate calibration (I)

Perhaps closest to our intuition: requires observations with the same predicted probability of '1' have an observed rate of '1' with the same probability.

Moderate calibration (I)

Perhaps closest to our intuition: requires observations with the same predicted probability of '1' have an observed rate of '1' with the same probability.

Continuum of probabilities means usually need to bin predicted probabilities,

$$[0.0, 0.1), [0.1, 0.2), \ldots, [0.9, 1.0]$$

and observe empirical frequency of responses y_i for corresponding observations in those bins.

Moderate calibration (I)

Perhaps closest to our intuition: requires observations with the same predicted probability of '1' have an observed rate of '1' with the same probability.

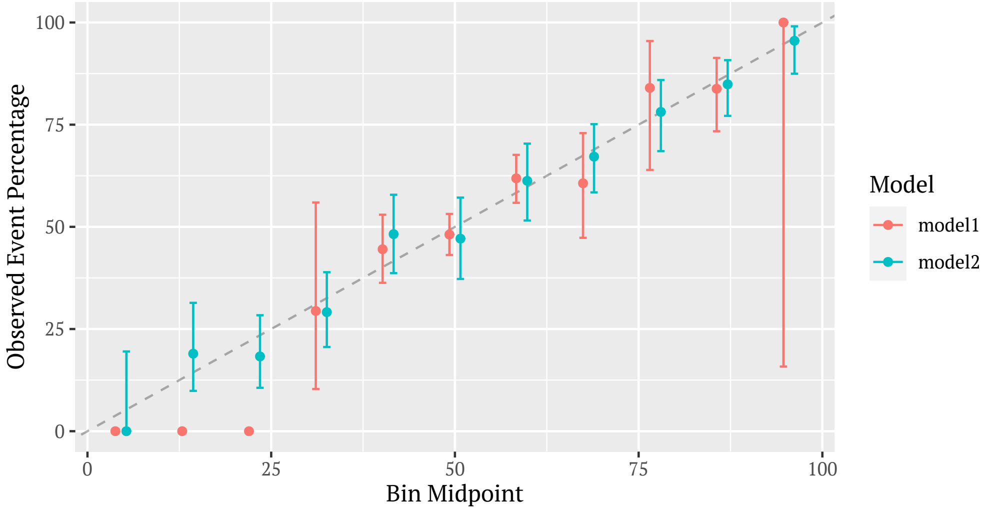
Continuum of probabilities means usually need to bin predicted probabilities,

$$[0.0, 0.1), [0.1, 0.2), \dots, [0.9, 1.0]$$

and observe empirical frequency of responses y_i for corresponding observations in those bins.

Can construct Binomial test confidence intervals for each bin, looking to cover $y = x$ line.

Moderate calibration (II)



Strong calibration

- Requires moderate calibration *for all* fixed feature combinations.

Strong calibration

- Requires moderate calibration *for all* fixed feature combinations.
- Impractical for continuous features

Strong calibration

- Requires moderate calibration *for all* fixed feature combinations.
- Impractical for continuous features
- Grouping on categorical variables may make impractically small validation sets

Strong calibration

- Requires moderate calibration *for all* fixed feature combinations.
- Impractical for continuous features
- Grouping on categorical variables may make impractically small validation sets

Therefore rarely is strong calibration practical to examine except in the largest data problems.

Calibration corrections

What to do if calibration is not satisfied?

- **Platt scaling:** fit logistic regression of the form:

$$p_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 \hat{f}(\mathbf{x}_i))}$$

  - sometimes use pseudo-responses to regularise (see notes)

Calibration corrections

What to do if calibration is not satisfied?

- **Platt scaling:** fit logistic regression of the form:

$$p_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 \hat{f}(\mathbf{x}_i))}$$

  - sometimes use pseudo-responses to regularise (see notes)

- **Isotonic regression:** fits a piecewise constant monotonically increasing function — principled method of correcting the probabilities within certain bins, allowing for bin location and size to be learned.

Learning curves

Compares performance on training and test data over model learned on increasing proportion of training data

- can model learn relationship in the data fast?

Learning curves

Compares performance on training and test data over model learned on increasing proportion of training data

- can model learn relationship in the data fast?

Performance should improve as training set size increases

Learning curves

Compares performance on training and test data over model learned on increasing proportion of training data

- can model learn relationship in the data fast?

Performance should improve as training set size increases

Should be fast initially and then slower

- at saturation (zero gradient), learned as much as possible for this model
- gap between train and test lines remains constant

Learning curves

Compares performance on training and test data over model learned on increasing proportion of training data

- can model learn relationship in the data fast?

Performance should improve as training set size increases

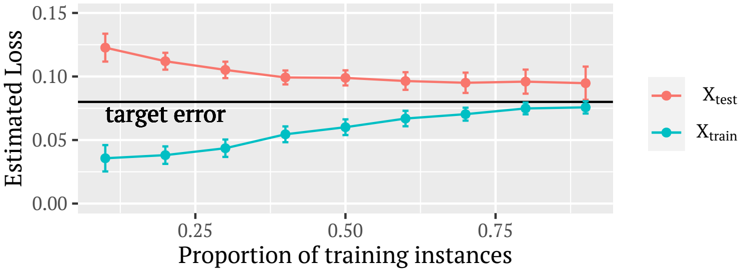
Should be fast initially and then slower

- at saturation (zero gradient), learned as much as possible for this model
- gap between train and test lines remains constant

Aim to find 'sweet spot' minimising the bias and variance at the right level of model complexity

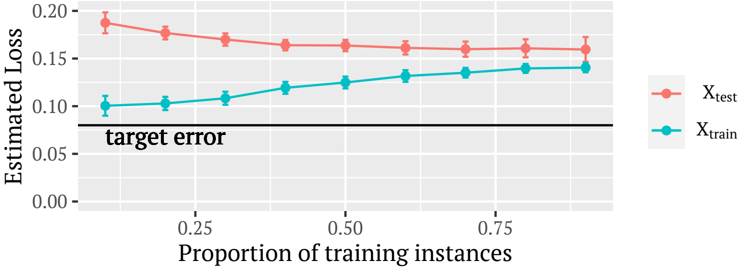
Learning curves — perfect learning curve

A ‘perfect’ learning curve would be like this:



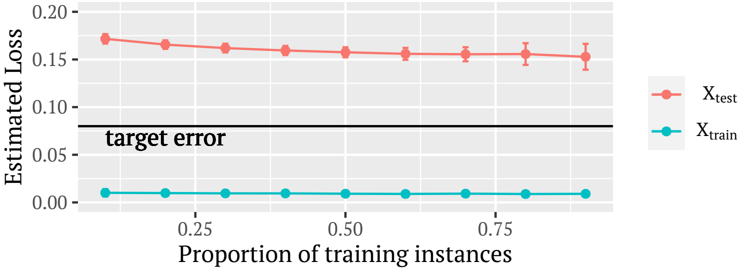
Learning curves — high bias

- Train and test errors converge but remain very high
  - more data not enough — the model is just insufficient to represent the true $f(\cdot)$
- Poor fit
- Poor generalisation to new data



Learning curves — high variance

- Large gap between train and test error
- Clear evidence more data needed
- Need to simplify model with fewer and/or less complex features



Further topics

- Ethics & Fairness

Further topics

- Ethics & Fairness
- Reproducibility

Further topics

- Ethics & Fairness
- Reproducibility
- Reporting frameworks

Further topics

- Ethics & Fairness
- Reproducibility
- Reporting frameworks
- Interpretable machine learning

Further topics

- Ethics & Fairness
- Reproducibility
- Reporting frameworks
- Interpretable machine learning
- Feature engineering

Further topics

- Ethics & Fairness
- Reproducibility
- Reporting frameworks
- Interpretable machine learning
- Feature engineering
- Privacy and confidentiality

Further topics

- Ethics & Fairness
- Reproducibility
- Reporting frameworks
- Interpretable machine learning
- Feature engineering
- Privacy and confidentiality
- Model updating

Further topics

- Ethics & Fairness
- Reproducibility
- Reporting frameworks
- Interpretable machine learning
- Feature engineering
- Privacy and confidentiality
- Model updating
- Missing data

Super learners (I)

Models: $f^{(1)}, \ldots, f^{(s)}$

Super learners (I)

Models: $f^{(1)}, \dots, f^{(s)}$

K-fold cross validation: $\mathcal{D}_{\mathcal{I}_1}, \dots, \mathcal{D}_{\mathcal{I}_K}$

Super learners (I)

Models: $f^{(1)}, \dots, f^{(s)}$

K-fold cross validation: $\mathcal{D}_{\mathcal{I}_1}, \dots, \mathcal{D}_{\mathcal{I}_K}$

Fit models:

$$\begin{gathered}
\hat{f}^{(1)}(\cdot \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_1}), \dots, \hat{f}^{(1)}(\cdot \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_K}) \\
\hat{f}^{(2)}(\cdot \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_1}), \dots, \hat{f}^{(2)}(\cdot \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_K}) \\
\vdots \\
\hat{f}^{(s)}(\cdot \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_1}), \dots, \hat{f}^{(s)}(\cdot \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_K})
\end{gathered}$$

Super learners (II)

For each observation i, in fold $\mathcal{I}_j$, construct predictions:

$$\hat{y}_i^{(1)} = \hat{f}^{(1)}(\mathbf{x}_i \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_j}), \dots, \hat{y}_i^{(s)} = \hat{f}^{(s)}(\mathbf{x}_i \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_j})$$

Super learners (II)

For each observation i, in fold $\mathcal{I}_j$, construct predictions:

$$\hat{y}_i^{(1)} = \hat{f}^{(1)}(\mathbf{x}_i \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_j}), \dots, \hat{y}_i^{(s)} = \hat{f}^{(s)}(\mathbf{x}_i \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_j})$$

Construct design matrix with original responses:

$$\mathbf{X} := \begin{pmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \dots & \hat{y}_1^{(s)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \dots & \hat{y}_2^{(s)} \\ & & \vdots & \\ \hat{y}_n^{(1)} & \hat{y}_n^{(2)} & \dots & \hat{y}_n^{(s)} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Super learners (II)

For each observation i, in fold $\mathcal{I}_j$, construct predictions:

$$\hat{y}_i^{(1)} = \hat{f}^{(1)}(\mathbf{x}_i \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_j}), \dots, \hat{y}_i^{(s)} = \hat{f}^{(s)}(\mathbf{x}_i \mid \mathcal{D} \setminus \mathcal{D}_{\mathcal{I}_j})$$

Construct design matrix with original responses:

$$\mathbf{X} := \begin{pmatrix} \hat{y}_1^{(1)} & \hat{y}_1^{(2)} & \dots & \hat{y}_1^{(s)} \\ \hat{y}_2^{(1)} & \hat{y}_2^{(2)} & \dots & \hat{y}_2^{(s)} \\ & & \vdots & \\ \hat{y}_n^{(1)} & \hat{y}_n^{(2)} & \dots & \hat{y}_n^{(s)} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Finally, build super learner model (eg logistic regression) to predict based on above.

References I

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* **110**, 12–22. DOI: 10.1016/j.jclinepi.2019.02.004

Cox, D.R. (1958). Two further applications of a model for binary regression. *Biometrika* **45**(3-4), 562–565. DOI: 10.1093/biomet/45.3-4.562