Academy of PhD Training in Statistics

Statistical Machine Learning

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

Error Estimation and Model Selection

This Section

- In-sample error estimation
  - Mallows' C_p
  - Akaike Information Criterion
  - Covariance penalities
  - Fixed -vs- random inputs
- Cross validation estimation
  - Train/Test/Validate
  - LOO
  - K-fold
  - Recent theoretical results

Why this matters

- Error estimation
  - training/apparent error is downward biased estimate
  - want to understand performance of model on future data

Why this matters

- Error estimation
  - training/apparent error is downward biased estimate
  - want to understand performance of model on future data
- Hyperparameter selection
  - how do we choose k and h from last section without relying on asymptotic theory?
  - may want hyperparameter tuned to non-standard loss for which no theory
  - there will by more hyperparameters to come ...

Why this matters

- Error estimation
  - training/apparent error is downward biased estimate
  - want to understand performance of model on future data
- Hyperparameter selection
  - how do we choose k and h from last section without relying on asymptotic theory?
  - may want hyperparameter tuned to non-standard loss for which no theory
  - there will by more hyperparameters to come ...
- Model selection
  - we may want to choose among multiple fitted models

What do we really need

Note:

- If we want to know our likely loss on future observations, we genuinely require a method that will produce and unbiased estimate of the generalisation error.

What do we really need

Note:

- If we want to know our likely loss on future observations, we genuinely require a method that will produce and unbiased estimate of the generalisation error.
- *But*, if we are just selecting between models/hyperparameters, we simply need a *consistent* estimator which correctly *orders* the performance.

In-sample estimation

May approximate the total expected in-sample loss:

$$\text{Err} := \sum_{i=1}^{n} \underbrace{\mathbb{E}_{Y \mid X=\mathbf{x}_i} \left[\mathcal{L}(Y, g_{\hat{f}}(\mathbf{x}_i)) \right]}_{\text{Err}_i}$$

with

$$\text{err} := \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, g_{\hat{f}}(\mathbf{x}_i))}_{\text{err}_i}$$

In-sample estimation

May approximate the total expected in-sample loss:

$$\text{Err} := \sum_{i=1}^{n} \underbrace{\mathbb{E}_{Y \mid X=\mathbf{x}_i} \left[\mathcal{L}(Y, g_{\hat{f}}(\mathbf{x}_i)) \right]}_{\text{Err}_i}$$

with

$$\text{err} := \sum_{i=1}^{n} \underbrace{\mathcal{L}(y_i, g_{\hat{f}}(\mathbf{x}_i))}_{\text{err}_i}$$

Problem: downward biased estimate.

Write: $\text{Err} = \text{err} + \omega$, where ω is *optimism* of apparent error.

In-sample estimation

- Can we estimate $\hat{\omega} \approx \omega$ only with training data?
- If so, then could use approximation

$$\widehat{\mathrm{Err}} = \mathrm{err} + \hat{\omega}$$

for error estimation, hyperparameter tuning and model selection.

Mallows' C_p (Mallows, 1973)

- Standard linear regression;
- homoskedastic Normal errors;
- $(Y \mid X = \mathbf{x}) \sim \mathrm{N}(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2)$;
- squared loss.

Then, Mallows' C_p is,

$$C_p := \frac{\sum (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2}{\sigma^2} - n + 2d$$

Mallows' C_p (Mallows, 1973)

- Standard linear regression;
- homoskedastic Normal errors;
- $(Y \mid X = \mathbf{x}) \sim \mathrm{N}(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2)$;
- squared loss.

Then, Mallows' C_p is,

$$C_p := \frac{\sum (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2}{\sigma^2} - n + 2d$$

Usual use: variable selection, with σ^2 estimated from data.

C_p and Err

$$\begin{aligned}
\text{Err}_i &= \mathbb{E}_{Y \mid X=\mathbf{x}_i}\left[\mathcal{L}(Y, g_{\hat{f}}(\mathbf{x}_i))\right] \\
&= \mathbb{E}_{Y \mid X=\mathbf{x}_i}\left[(Y - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2\right] \\
&= \left(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)^2 + \sigma^2
\end{aligned}$$

C_p and Err

$$\begin{aligned}\text{Err}_i &= \mathbb{E}_{Y \mid X=\mathbf{x}_i}\left[\mathcal{L}(Y, g_{\hat{f}}(\mathbf{x}_i))\right] \\ &= \mathbb{E}_{Y \mid X=\mathbf{x}_i}\left[(Y - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2\right] \\ &= \left(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)^2 + \sigma^2\end{aligned}$$

Mallows (1973) proved it is an unbiased estimator,

$$C_p \approx \sigma^{-2}\sum\left(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)^2$$

C_p and Err

$$\begin{aligned}\text{Err}_i &= \mathbb{E}_{Y \mid X=\mathbf{x}_i}\left[\mathcal{L}(Y, g_{\hat{f}}(\mathbf{x}_i))\right] \\ &= \mathbb{E}_{Y \mid X=\mathbf{x}_i}\left[(Y - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2\right] \\ &= \left(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)^2 + \sigma^2\end{aligned}$$

Mallows (1973) proved it is an unbiased estimator,

$$\begin{aligned}C_p &\approx \sigma^{-2}\sum\left(\mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}\right)^2 \\ \implies \sigma^2 C_p &\approx \text{Err} - n\sigma^2 \\ \implies \hat{\omega} &\approx 2d\sigma^2\end{aligned}$$

Alternative definition

Above relation leads to an alternative form:

$$\tilde{C}_p = \text{err} + 2d\sigma^2$$

Note, $\tilde{C}_p = \sigma^2(C_p + n)$, so minimised as by model, but $\tilde{C}_p$ is direct estimator of total loss.

General linear estimators

A *linear estimation rule* is any predictor of the form:

$$\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$$

where $\mathbf{M}$ does not depend on $\mathbf{y}$.

General linear estimators

A *linear estimation rule* is any predictor of the form:

$$\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$$

where $\mathbf{M}$ does not depend on $\mathbf{y}$.

$$\begin{aligned} \text{eg} \quad \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]\mathbf{y} \end{aligned}$$

so $\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ for standard linear regression.

Mallows' for linear estimators

Mallows' can be applied to any linear estimator (Efron, 2004),

$$\hat{\omega}_i = 2\sigma^2 M_{ii}$$

$$\implies \widehat{\text{Err}} = \text{err} + 2\sigma^2 \,\text{trace}(\mathbf{M})$$

Mallows' for linear estimators

Mallows' can be applied to any linear estimator (Efron, 2004),

$$\hat{\omega}_i = 2\sigma^2 M_{ii}$$

$$\implies \widehat{\text{Err}} = \text{err} + 2\sigma^2 \,\text{trace}(\mathbf{M})$$

$\text{trace}(\mathbf{M})$ is commonly referred to as the *degrees of freedom* (df) (Tibshirani, 2015).

Other linear estimators

KNN: in-sample nearest neighbour prediction of all responses is $\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$, where

$$M_{ij} = \begin{cases} \frac{1}{k} & \text{if } d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k,\mathbf{x}_i)}) \\ 0 & \text{otherwise} \end{cases}$$

Other linear estimators

KNN: in-sample nearest neighbour prediction of all responses is $\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$, where

$$M_{ij} = \begin{cases} \frac{1}{k} & \text{if } d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k,\mathbf{x}_i)}) \\ 0 & \text{otherwise} \end{cases}$$

Nadaraya-Watson: in-sample N-W prediction of all responses is $\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$, where

$$M_{ij} = \frac{K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h}\right)}{\sum_{\ell=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}_\ell}{h}\right)}$$

Other linear estimators

KNN: in-sample nearest neighbour prediction of all responses is $\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$, where

$$M_{ij} = \begin{cases} \frac{1}{k} & \text{if } d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_{(k,\mathbf{x}_i)}) \\ 0 & \text{otherwise} \end{cases}$$

Nadaraya-Watson: in-sample N-W prediction of all responses is $\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$, where

$$M_{ij} = \frac{K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h}\right)}{\sum_{\ell=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}_\ell}{h}\right)}$$

Others: ridge regression, lasso, group lasso, and spline smoothing (Arlot and Bach, 2009)

Akaike Information Criterion (AIC)

Akaike (1972) extended concept of maximum likelihood by examining maximum of expected log-likelihood. For each model, take MLE $\hat{\boldsymbol{\theta}}$ and compute:

$$\mathbb{E}_X[\log f_X(X \mid \hat{\boldsymbol{\theta}})] = \int_{\mathcal{X}} \log f_X(x \mid \hat{\boldsymbol{\theta}}) d\pi_X = -\mathcal{E}\left(\hat{f}(\cdot \mid \hat{\boldsymbol{\theta}})\right)$$

then choose model maximising this.

- equivalently minimising generalisation error for log-likelihood loss in full probabilistic model;
- equivalently minimising Err_i in fixed inputs case if model doesn't specify distribution on predictors.

AIC

Equivalently, maximise

$$\mathbb{E}_X\left[\log \frac{f_X(X \mid \hat{\boldsymbol{\theta}})}{\pi_X(X)}\right] = \int_{\mathcal{X}} \log \frac{f_X(x \mid \hat{\boldsymbol{\theta}})}{\pi_X(x)} d\pi_X$$

For us, this is $-\mathcal{E}\left(\hat{f}(\cdot \mid \hat{\boldsymbol{\theta}})\right) + \mathcal{E}\left(\pi_X\right)$ ... ie negative excess risk!

AIC

Equivalently, maximise

$$\mathbb{E}_X \left[\log \frac{f_X(X \mid \hat{\boldsymbol{\theta}})}{\pi_X(X)} \right] = \int_{\mathcal{X}} \log \frac{f_X(x \mid \hat{\boldsymbol{\theta}})}{\pi_X(x)} d\pi_X$$

For us, this is $-\mathcal{E}\left(\hat{f}(\cdot \mid \hat{\boldsymbol{\theta}})\right) + \mathcal{E}(\pi_X)$ ... ie negative excess risk!

Akaike (1972) derives the AIC:

$$\text{AIC} = -2 \left(\sum_{i=1}^{n} \log f_X(\mathbf{x}_i \mid \hat{\boldsymbol{\theta}}) \right) + 2d$$

where d is the dimension of $\boldsymbol{\theta}$

Relation to generalisation error

AIC is an estimator of the true generalisation error (Akaike, 1974):

$$\text{AIC} \approx 2n\mathbb{E}\left[\mathcal{E}\left(\hat{f}(\cdot \mid \hat{\boldsymbol{\theta}})\right)\right]$$

ie the optimism for a log-likelihood loss is approximately d.

Care in interpretation: full probabilistic model or fixed inputs?

Model selection -vs- prediction

- For model selection, compute AIC for candidates and selection one with smallest AIC.
  - BUT, it will *not* asymptotically choose ‘true’ model!

Model selection -vs- prediction

- For model selection, compute AIC for candidates and selection one with smallest AIC.
  - BUT, it will *not* asymptotically choose ‘true’ model!
- For prediction, *does* choose model which offers equivalent loss to the smallest available!
  - … what we want in machine learning!

Model selection -vs- prediction

- For model selection, compute AIC for candidates and selection one with smallest AIC.
  - BUT, it will *not* asymptotically choose ‘true’ model!
- For prediction, *does* choose model which offers equivalent loss to the smallest available!
  - … what we want in machine learning!
- If you are doing scientific modelling, investigate Bayesian Information Criterion (BIC)
  - In ML we would not often expect our model to represent the data generating process.

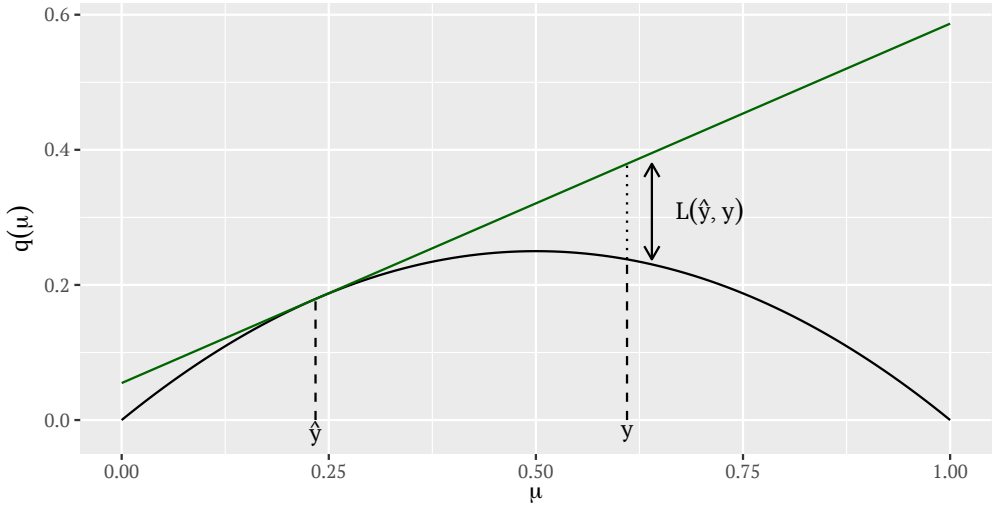
General covariance penalties

Efron (1986) generalised framework for analysing optimism in the apparent error by defining q-class of losses.

A loss is said to be a *q-class loss* if it can be written, for some concave function $q(\mu) : \mathcal{Y} \to \mathbb{R}$, as

$$\mathcal{L}(y, \hat{y}) = q(\hat{y}) + \left.\frac{dq}{d\mu}\right|_{\mu=\hat{y}} (y - \hat{y}) - q(y)$$

q-class loss (square loss)



Common q-losses

Many standard losses belong to the q-class:

- squared loss (shown in plot):

$$q(\mu) = \mu(1-\mu) \implies \mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

- 0-1 loss:

$$q(\mu) = \min\{\mu, 1-\mu\} \implies \mathcal{L}(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$$

- binary cross-entropy:

$$q(\mu) = -\mu \log \mu - (1-\mu)\log(1-\mu) \implies \mathcal{L}(y, \hat{\mathbf{p}}) = -\log \hat{p}_y$$

For full details see Efron (1986).

Optimisim Theorem (Efron, 2004)

Given a loss belonging to the q-class of losses, we have that,

$$\mathbb{E}[\text{Err}_i] = \mathbb{E}[\text{err}_i + \omega_i]$$

where

$$\omega_i = 2\text{Cov}(\hat{\lambda}_i, y_i)$$

with

$$\hat{\lambda}_i = -\frac{1}{2} \left. \frac{dq}{d\mu} \right|_{\mu=\hat{y}_i}$$

Example: square loss

$$q(\mu) = \mu(1-\mu) \implies \mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

$$\implies \hat{\lambda}_i = \hat{y}_i - \frac{1}{2}$$

$$\implies \omega_i = 2\text{Cov}(\hat{y}_i, y_i)$$

- if observation influences the value the model predicts for that observation, optimism higher
- observation is highly influential of its own prediction, then overfitting is likely
- calculation of ω_i can be very difficult or intractable (and we don't know true distribution of Y ... Bootstrap)

Data splitting methods

Many methods are based on splitting the available data up, using only part for model fitting.

New Notation

Full data,

$$\mathcal{D} = \mathcal{D}_n = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$$

Subset,

$$\mathcal{D}_{\mathcal{I}} := \{(\mathbf{x}_i, y_i) : i \in \mathcal{I}\} \quad \text{where} \quad \mathcal{I} \subset \{1, \ldots, n\}$$

Hold-out estimation (I)

Simplest approach and theoretical analysis easy!

Create partition,

- Training, $\mathcal{D}_{\mathcal{T}_r}$
- Testing, $\mathcal{D}_{\mathcal{T}_e}$

Where $\mathcal{T}_r \cap \mathcal{T}_e = \emptyset$ and $\mathcal{T}_r \cup \mathcal{T}_e = \{1, \ldots, n\}$.

Hold-out estimation (I)

Simplest approach and theoretical analysis easy!

Create partition,

- Training, $\mathcal{D}_{\mathcal{T}_r}$
- Testing, $\mathcal{D}_{\mathcal{T}_e}$

Where $\mathcal{T}_r \cap \mathcal{T}_e = \emptyset$ and $\mathcal{T}_r \cup \mathcal{T}_e = \{1, \ldots, n\}$.

$$\widehat{\mathrm{Err}}_{\mathrm{ho}} = \frac{1}{|\mathcal{T}_e|} \sum_{i \in \mathcal{T}_e} \mathcal{L}(y_i, \hat{f}(\mathbf{x}_i \,|\, \mathcal{D}_{\mathcal{T}_r}))$$

Hold-out estimation (II)

$\widehat{\text{Err}}_{\text{ho}}$ is an unbiased estimate for the generalisation error:

$$\mathcal{E}(\hat{f} \mid \mathcal{D}_{\mathcal{T}_r}) = \mathbb{E}_{XY}\left[\mathcal{L}(Y, \hat{f}(X \mid \mathcal{D}_{\mathcal{T}_r}))\right]$$

with standard error

$$\hat{s} := \sqrt{\frac{1}{|\mathcal{T}_e| - 1} \sum_{i \in \mathcal{T}_e} \left(\mathcal{L}(y_i, \hat{f}(\mathbf{x}_i \mid \mathcal{D}_{\mathcal{T}_r})) - \widehat{\text{Err}}_{\text{ho}}\right)^2}$$

Hold-out estimation (II)

$\widehat{\text{Err}}_{\text{ho}}$ is an unbiased estimate for the generalisation error:

$$\mathcal{E}(\hat{f} \mid \mathcal{D}_{\mathcal{T}_r}) = \mathbb{E}_{XY}\left[\mathcal{L}(Y, \hat{f}(X \mid \mathcal{D}_{\mathcal{T}_r}))\right]$$

with standard error

$$\hat{s} := \sqrt{\frac{1}{|\mathcal{T}_e| - 1} \sum_{i \in \mathcal{T}_e} \left(\mathcal{L}(y_i, \hat{f}(\mathbf{x}_i \mid \mathcal{D}_{\mathcal{T}_r})) - \widehat{\text{Err}}_{\text{ho}}\right)^2}$$

Catch: Model often refitted to whole data for production

Hold-out CI coverage

Is just a sample size adjustment needed? Sadly not (Bates et al., 2021):

$$\begin{aligned}\mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\widehat{\mathrm{Err}}_{\mathrm{ho}}\right] &= \mathbb{E}_{D_{|\mathcal{T}_r|}}\left[\frac{1}{|\mathcal{T}_e|}\sum_{i\in\mathcal{T}_e}\mathbb{E}_{XY}\left[\mathcal{L}(Y_i,\hat{f}(X_i\,|\,D_{|\mathcal{T}_r|}))\right]\right] \\ &= \mathbb{E}_{D_{|\mathcal{T}_r|}}\left[\mathbb{E}_{XY}\left[\mathcal{L}(Y_i,\hat{f}(X_i\,|\,D_{|\mathcal{T}_r|}))\right]\right] \\ &= \bar{\mathcal{E}}_{|\mathcal{T}_r|}\end{aligned}$$

Hold-out CI coverage

Is just a sample size adjustment needed? Sadly not (Bates et al., 2021):

$$\begin{aligned}
\mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\widehat{\mathrm{Err}}_{\mathrm{ho}}\right] &= \mathbb{E}_{D_{|\mathcal{T}_r|}}\left[\frac{1}{|\mathcal{T}_e|}\sum_{i\in\mathcal{T}_e}\mathbb{E}_{XY}\left[\mathcal{L}(Y_i,\hat{f}(X_i \mid D_{|\mathcal{T}_r|}))\right]\right] \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}}\left[\mathbb{E}_{XY}\left[\mathcal{L}(Y_i,\hat{f}(X_i \mid D_{|\mathcal{T}_r|}))\right]\right] \\
&= \bar{\mathcal{E}}_{|\mathcal{T}_r|}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\hat{s}^2\right] &= \mathbb{E}_{D_{|\mathcal{T}_r|}}\left[\mathrm{Var}\left(\widehat{\mathrm{Err}}_{\mathrm{ho}} \mid D_{|\mathcal{T}_r|}\right)\right] \\
&= \mathrm{Var}\left(\widehat{\mathrm{Err}}_{\mathrm{ho}}\right) - \mathrm{Var}\left(\mathbb{E}_{D_{|\mathcal{T}_r|}}\left[\widehat{\mathrm{Err}}_{\mathrm{ho}} \mid D_{|\mathcal{T}_r|}\right]\right) \quad \text{law of total variance} \\
&= \star
\end{aligned}$$

$$
\begin{aligned}
\operatorname{Var}\left(\widehat{\operatorname{Err}}_{\text{ho}}\right) &= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\operatorname{Err}}_{\text{ho}} - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2\right] \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\left(\widehat{\operatorname{Err}}_{\text{ho}} - \bar{\mathcal{E}}_n\right) + \left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)\right)^2\right] \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\operatorname{Err}}_{\text{ho}} - \bar{\mathcal{E}}_n\right)^2\right] + 2\left(\bar{\mathcal{E}}_{|\mathcal{T}_r|} - \bar{\mathcal{E}}_n\right)\left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right) + \left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2 \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\operatorname{Err}}_{\text{ho}} - \bar{\mathcal{E}}_n\right)^2\right] - \left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\left(\widehat{\mathrm{Err}}_{\mathrm{ho}}\right) &= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\mathrm{Err}}_{\mathrm{ho}} - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2\right] \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\left(\widehat{\mathrm{Err}}_{\mathrm{ho}} - \bar{\mathcal{E}}_n\right) + \left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)\right)^2\right] \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\mathrm{Err}}_{\mathrm{ho}} - \bar{\mathcal{E}}_n\right)^2\right] + 2\left(\bar{\mathcal{E}}_{|\mathcal{T}_r|} - \bar{\mathcal{E}}_n\right)\left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right) + \left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2 \\
&= \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\mathrm{Err}}_{\mathrm{ho}} - \bar{\mathcal{E}}_n\right)^2\right] - \left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2
\end{aligned}
$$

$$
\star = \mathbb{E}_{D_{|\mathcal{T}_r|}D_{|\mathcal{T}_e|}}\left[\left(\widehat{\mathrm{Err}}_{\mathrm{ho}} - \bar{\mathcal{E}}_n\right)^2\right] - \underbrace{\left(\bar{\mathcal{E}}_n - \bar{\mathcal{E}}_{|\mathcal{T}_r|}\right)^2}_{\text{sample size bias}} - \mathrm{Var}\left(\mathcal{E}(\hat{f} \mid \mathcal{D}_{\mathcal{T}_r})\right)
$$

Train/test/validate

Hold-out ok for error estimation ... model selection/hyperparameter tuning? Need a further split! Assume θ indexes models or is a hyperparameter.

Create partition,

- Training, $\mathcal{D}_{\mathcal{T}_r}$
- Validation, $\mathcal{D}_{\mathcal{V}}$
- Testing, $\mathcal{D}_{\mathcal{T}_e}$

Where $\mathcal{T}_r \cap \mathcal{T}_e = \emptyset, \mathcal{T}_r \cap \mathcal{V} = \emptyset, \mathcal{T}_e \cap \mathcal{V} = \emptyset$ and $\mathcal{T}_r \cup \mathcal{V} \cup \mathcal{T}_e = \{1, \ldots, n\}$.

Train/test/validate

Hold-out ok for error estimation ... model selection/hyperparameter tuning? Need a further split! Assume θ indexes models or is a hyperparameter.

Create partition,

- Training, $\mathcal{D}_{\mathcal{T}_r}$
- Validation, $\mathcal{D}_{\mathcal{V}}$
- Testing, $\mathcal{D}_{\mathcal{T}_e}$

Where $\mathcal{T}_r \cap \mathcal{T}_e = \emptyset, \mathcal{T}_r \cap \mathcal{V} = \emptyset, \mathcal{T}_e \cap \mathcal{V} = \emptyset$ and $\mathcal{T}_r \cup \mathcal{V} \cup \mathcal{T}_e = \{1, \ldots, n\}$.

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathcal{L}(y_i, \hat{f}_\theta(\mathbf{x}_i \,|\, \mathcal{D}_{\mathcal{T}_r}))$$

$$\widehat{\mathrm{Err}}_{\mathrm{te}} = \frac{1}{|\mathcal{T}_e|} \sum_{i \in \mathcal{T}_e} \mathcal{L}(y_i, \hat{f}_{\hat{\theta}}(\mathbf{x}_i \,|\, \mathcal{D}_{\mathcal{T}_r}))$$

Problems

$$\mathbf{X} = \begin{pmatrix} 0.391 & 0.153 & \\ 0.628 & 0.394 & \\ \vdots & \vdots & \dots \\ -1.556 & 2.421 & \end{pmatrix} \longrightarrow \begin{array}{lll} \mathbf{X}_{\text{tr}} = \begin{pmatrix} 0.391 & 0.153 & \\ \vdots & \vdots & \dots \\ 1.385 & 0.629 & \end{pmatrix} & \text{Fit models} \\ \mathbf{X}_{\text{te}} = \begin{pmatrix} -1.556 & 2.421 & \\ \vdots & \vdots & \dots \\ 0.996 & 0.056 & \end{pmatrix} & \text{Choose model/ pars} \\ \mathbf{X}_{\text{va}} = \begin{pmatrix} 0.628 & 0.394 & \\ \vdots & \vdots & \dots \\ -0.211 & -0.729 & \end{pmatrix} & \text{Final accuracy estimate} \end{array}$$

Match splits with response $\mathbf{y} = (\mathbf{y}_{\text{tr}}, \mathbf{y}_{\text{te}}, \mathbf{y}_{\text{va}})$.

Problems

$$\mathbf{X} = \begin{pmatrix} 0.391 & 0.153 & \\ 0.628 & 0.394 & \\ \vdots & \vdots & \cdots \\ -1.556 & 2.421 & \end{pmatrix} \longrightarrow \begin{array}{ll} \mathbf{X}_{\text{tr}} = \begin{pmatrix} 0.391 & 0.153 & \\ \vdots & \vdots & \cdots \\ 1.385 & 0.629 & \end{pmatrix} & \text{Fit models} \\ & 50\% \text{ of data} \\ \mathbf{X}_{\text{te}} = \begin{pmatrix} -1.556 & 2.421 & \\ \vdots & \vdots & \cdots \\ 0.996 & 0.056 & \end{pmatrix} & \text{Choose model/ pars} \\ & 25\% \text{ of data} \\ \mathbf{X}_{\text{va}} = \begin{pmatrix} 0.628 & 0.394 & \\ \vdots & \vdots & \cdots \\ -0.211 & -0.729 & \end{pmatrix} & \text{Final accuracy estimate} \\ & 25\% \text{ of data} \end{array}$$

Match splits with response $\mathbf{y} = (\mathbf{y}_{\text{tr}}, \mathbf{y}_{\text{te}}, \mathbf{y}_{\text{va}})$.

Train/test/validate problems

- Only some of the data is used in fitting, other parts never used during fit.
- Only some of data is used in evaluation (what if hard to predict observations are by chance allocated to train/test/...)
- Again, the final error estimate will usually be conservative, since once the best model is chosen we refit to the whole dataset and would expect slightly improved results.
- If little data, possibly hybrid: use in-sample estimators to choose hyperparameters and estimate error using hold-out

Cross-validation

The "original" cross-validation is called Leave One Out (LOO).

- Splits $\mathcal{I}_{-i} = \{1, \ldots, n\} \setminus \{i\}$ for $i \in \{1, \ldots, n\}$
- Total of n models fitted to each $\mathcal{I}_{-i}$
- Error for observation i assessed on model fitted to $\mathcal{I}_{-i}$

Overall,

$$\widehat{\mathrm{Err}}_{\mathrm{loo}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{f}(\mathbf{x}_i \,|\, \mathcal{D}_{\mathcal{I}_{-i}}))$$

Cross-validation

The "original" cross-validation is called Leave One Out (LOO).

- Splits $\mathcal{I}_{-i} = \{1, \ldots, n\} \setminus \{i\}$ for $i \in \{1, \ldots, n\}$
- Total of n models fitted to each $\mathcal{I}_{-i}$
- Error for observation i assessed on model fitted to $\mathcal{I}_{-i}$

Overall,

$$\widehat{\mathrm{Err}}_{\mathrm{loo}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{f}(\mathbf{x}_i \mid \mathcal{D}_{\mathcal{I}_{-i}}))$$

But: computationally expensive, asymptotically equivalent to AIC.

K-fold cross-validation

Arguably most popular form of cross validation.

Partition data into K equally sized disjoint parts, $\mathcal{I}_1, \ldots, \mathcal{I}_K$ with

$$\mathcal{I}_i \cap \mathcal{I}_j = \emptyset \; \forall \, i \neq j, \text{ and } \bigcup_{i=1}^{K} \mathcal{I}_i = \{1, \ldots, n\}$$

each fold having $|\mathcal{I}_j| = \frac{n}{K}$ observations.

(for simplicity, assume K divides n)

K-fold cross-validation setup

K-fold cross validation splits data into k equally sized groups, called 'folds':

$$\mathbf{X} = \begin{pmatrix} 0.391 & 0.153 & \\ 0.628 & 0.394 & \\ \vdots & \vdots & \dots \\ -1.556 & 2.421 & \end{pmatrix} \longrightarrow$$

$$\mathbf{X}_1 = \begin{pmatrix} 0.391 & 0.153 & \\ \vdots & \vdots & \dots \\ 1.385 & 0.629 & \end{pmatrix} \quad \text{fold 1}$$

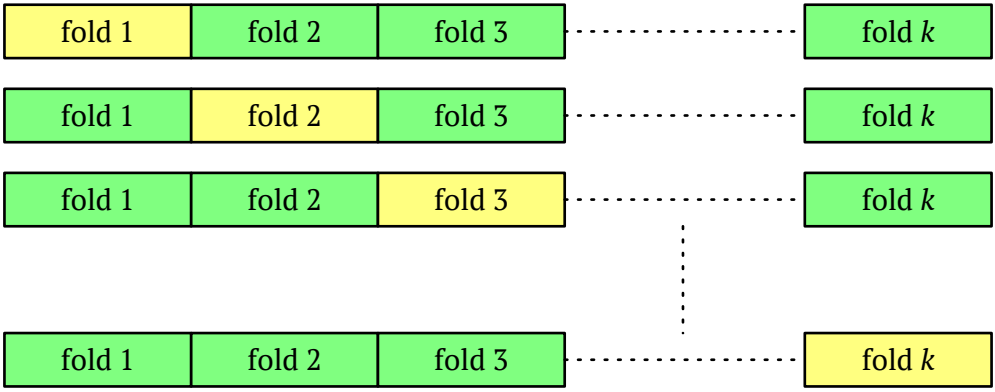
$$\mathbf{X}_2 = \begin{pmatrix} -1.556 & 2.421 & \\ \vdots & \vdots & \dots \\ 0.996 & 0.056 & \end{pmatrix} \quad \text{fold 2}$$

$$\vdots$$

$$\mathbf{X}_k = \begin{pmatrix} 0.628 & 0.394 & \\ \vdots & \vdots & \dots \\ -0.211 & -0.729 & \end{pmatrix} \quad \text{fold } k$$

Cross-validation folds

K times we: fit on green folds, evaluate error on the held out yellow fold.



Choice of K

- $K = n$ (ie LOO)
  - has the lowest bias, since each model is almost the same as the full data model!
  - *but* has very high variance since all models are so highly correlated with each other (mean of correlated variables has higher variance)
- $K = 2$
  - has high bias, for the same reason as train/test/validate
  - lower variance, as models have no data dependent correlation

$K = 5$ and $K = 10$ are common choices in the wild.

Choice of K

- $K = n$ (ie LOO)
  - has the lowest bias, since each model is almost the same as the full data model!
  - *but* has very high variance since all models are so highly correlated with each other (mean of correlated variables has higher variance)
- $K = 2$
  - has high bias, for the same reason as train/test/validate
  - lower variance, as models have no data dependent correlation

$K = 5$ and $K = 10$ are common choices in the wild.

CAUTION! Time series, hidden ordering and concept drift all require careful attention.

Theory

- Wager (2020): Cross-validation is,
  - asymptotically consistent in identifying the better performing of two models,

Theory

- Wager (2020): Cross-validation is,
  - asymptotically consistent in identifying the better performing of two models,
  - biased estimate of generalisation error.

Theory

- Wager (2020): Cross-validation is,
  - asymptotically consistent in identifying the better performing of two models,
  - biased estimate of generalisation error.
- Bates et al. (2021): Cross-validation is,
  - more closely estimating expected (prediction) error, $\bar{\mathcal{E}}_n(\cdot)$

Theory

- Wager (2020): Cross-validation is,
  - asymptotically consistent in identifying the better performing of two models,
  - biased estimate of generalisation error.
- Bates et al. (2021): Cross-validation is,
  - more closely estimating expected (prediction) error, $\bar{\mathcal{E}}_n(\cdot)$
  - but, confidence intervals calibrated to include $\mathcal{E}(\hat{f} \mid \mathcal{D})$, can be constructed using *nested* cross validation

Bootstrap

You have covered in previous APTS, so we just provide definition (and a little more in notes):

Consider data set $\mathcal{D} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ and a statistic $S(\cdot)$ one wishes to estimate.

To construct a bootstrap estimate of the confidence interval of the statistic $S(\cdot)$, draw B new samples of size n *with replacement* from $\mathcal{D}$, $\mathcal{D}^{\star 1}, \ldots, \mathcal{D}^{\star B}$ and compute:

$$\widehat{\text{Var}}(S(\mathcal{D})) = \frac{1}{B-1} \sum_{b=1}^{B} \left(S(\mathcal{D}^{\star b}) - \bar{S}^{\star} \right)^2$$

where $\bar{S}^{\star} = \frac{1}{B} \sum_{b=1}^{B} S(\mathcal{D}^{\star b})$.

References I

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723. DOI: 10.1109/TAC.1974.1100705

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle, in: Proceedings of the 2nd International Symposium on Information Theory. pp. 267–281.

Arlot, S., Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. *arXiv* (0909.1884). URL `https://arxiv.org/abs/0909.1884`

Bates, S., Hastie, T., Tibshirani, R. (2021). Cross-validation: What does it estimate and how well does it do it? *arXiv* (2104.00673). URL `https://arxiv.org/abs/2104.00673`

References II

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association* **99**(467), 619–632. DOI: 10.1198/016214504000000692

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**(394), 461–470. DOI: 10.2307/2289236

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**(4), 661–675. DOI: 10.1080/00401706.1973.10489103

Tibshirani, R.J. (2015). Degrees of freedom and model search. *Statistica Sinica* **25**(3), 1265–1296. DOI: 10.5705/ss.2014.147

Wager, S. (2020). Cross-validation, risk estimation, and model selection: Comment on a paper by Rosset and Tibshirani. *Journal of the American Statistical Association* **115**(529), 157–160. DOI: 10.1080/01621459.2020.1727235