Cross validation, Bootstrap, and Tests

Model assessment

Cross validation:

- Data splitting revisited
- Uncertainty for the expected test score
- Multiple splitting and cross validation

Bootstrap

- Data resampling
- Bias and variance estimation for estimators
- Parametric bootstrap

Exploiting exchangability

- Residual resampling, basic exchangeability
- Randomisation/permutation tests

Proper scores and data splits

- Recall the expectation of a score for a forecast F under a true distribution G,

$$S(F, G) = \mathsf{E}_{y \sim G}[S(F, y)].$$

- A (negatively oriented) score is *proper* if $S(F, G) \geq S(G, G)$ for all predictions F.
- In simulation studies, we know the true G, so if we can simulate from it or do direct calculations, we can estimate or compute $S(F, G)$ for a given forecast F.
- In real applications, G is unknown.

Before, we split the data in subsets for *observation*/*estimation*/*training*, $\mathcal{Y}^{\text{train}}$, and *test*, $\mathcal{Y}^{\text{test}}$.

Basic estimation and testing with data splitting

- Estimate the model using $\mathcal{Y}^{\text{train}}$
- Construct forecasts F_i^{test} for y_i in $\mathcal{Y}^{\text{test}}$.
- Estimate $\overline{S}(\{F_i^{\text{test}}\}, \{G_i^{\text{test}}\})$ with $\overline{S}(\{F_i^{\text{test}}\}, \{y_i^{\text{test}}\}) = \frac{1}{|\mathcal{Y}^{\text{test}}|} \sum_{i=1}^{|\mathcal{Y}^{\text{test}}|} S(F_i^{\text{test}}, y_i^{\text{test}})$

Generalised splitting

- Observation/Estimation/Training — Data used to estimate a model
- Validation — Data for assessing estimates and taking modelling decisions
- Test — Data used in a final step to assess the resulting model

- Decisions based on scores evaluated on Training or Validation data might lead to overestimation of the predictive ability.
- *Holding out* a separate Test set provides a safer way of assessing predictive ability.

Basic score uncertainty

- We're interested in the expected average prediction test score $\overline{S}(F^{\text{test}}, G^{\text{test}})$
- Before looking at the Test data, we only have access to an *estimate* of $\overline{S}(F^{\text{test}}, G^{\text{test}})$, based on a training/validation data split:

$$\widehat{S}^{\text{valid}} = \frac{1}{N} \sum_{i=1}^{N} S(F_i^{\text{valid}}, y_i^{\text{valid}})$$

- Note: To investigate the *difference* in expected score between two models or methods, just replace $S(F_i, y_i)$ by the pairwise differences $S_i^{\Delta} = S^{\Delta}(F_i, F_i', y_i) = S(F_i, y_i) - S(F_i', y_i)$ everywhere.
- The empirical variance estimate for $\widehat{S}^{\text{valid}}$ is

$$\widehat{\text{Var}}[\widehat{S}^{\text{valid}}] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left[S(F_i^{\text{valid}}, y_i^{\text{valid}}) - \widehat{S}^{\text{valid}} \right]^2$$

- The variance estimate may be biased due to dependence between the scores.

Cross validation

- We're interested in the expected average prediction test score $\overline{S}(F^{\text{test}}, G^{\text{test}})$ when using all the Training and Validation data to estimate the parameters of the final model.
- The Training set is a subset; may lead to overestimation of the expected score
- The Validation set is a small subset; high variability in the score estimator
- Different splits might give different score estimates and hence different modelling decisions
- Partial solution: Do multiple splits

K-fold Cross Validation: CV(K)

- Split the N data points $\mathcal{D}$ into K subsets $\mathcal{D}_k^{(\text{K})}$, each of size N/K.
- Iterate over the K subsets, treating each as a Validation set, $\mathcal{D}_k^{\text{valid}} = \mathcal{D}_k^{(\text{K})}$, and the remaining $K-1$ subsets as Training data $\mathcal{D}_k^{\text{train}} = \cup_{j \neq k} \mathcal{D}_j^{(\text{K})}$.
- Average over the resulting K score estimates.

Cross-validation scores

- For each of the K *folds*, the estimator of the expected score is

$$\widehat{S}_k^{(\mathsf{K})} = \frac{K}{N} \sum_{i=1}^{N/K} S(F_{ki}^{\mathsf{valid}}, y_{ki}^{\mathsf{valid}})$$

- The combined cross-validation score is

$$\widehat{S}^{\mathsf{CV}(\mathsf{K})} = \frac{1}{K} \sum_{k=1}^{K} \widehat{S}_k^{(\mathsf{K})}$$

- There are many options for estimating the variance of the combined CV score. Simple:

$$\widehat{\mathsf{Var}}[\widehat{S}^{\mathsf{CV}(\mathsf{K})}] = \frac{1}{K(K-1)} \sum_{k=1}^{K} [\widehat{S}_k^{(\mathsf{K})} - \widehat{S}^{\mathsf{CV}(\mathsf{K})}]^2$$

- No universal rule for what K and splitting choices will minimise the bias and variance of the estimators.
  Common choice is $K = 10$ and random splitting.

Common problem-dependent splitting options

- Leave-one-out CV; LOOCV=CV(N)
  In general very expensive, but for some model classes fast approximations are possible; Notably in Gaussian time series and spatial models
- Structured, only partially random, cross-validation examples:
  - Leave-station-out (to assess spatial predictive ability)
  - Leave-country-out (to assess macro scale generalisability, including potentially different measurement systems)
  - Leave-timepoint-out (to assess temporal interpolation ability)
  - Leave-individual-out (to assess generalisability of medical treatment between patients)
  - Related non-cross-validation example: Leave-future-out (to assess forecasting ability)
- Alternative: Instead of complete splitting, do multiple Validation subset selections as random subsamples with replacement (related to *Bootstrap*)

Cross validation and Bootstrap

- Cross validation splits the data in K parts and performs model estimation on $(K-1)$ parts and validation assessment on the Kth part, all for each of the parts.
- Bootstrap resamples *with replacement* to obtain a random sample of the same size as the original sample.

Basic Bootstrap resampling

Let $Y = \{(y_i, x_i), i = 1, \ldots, N\}$ be a data collection with response values y_i and predictors/covariates x_i.

- Define a *Bootstrap sample* $Y^{(j)}$ by drawing N pairs (y_i, x_i) from Y with equal probability, and with replacement.
- Repeat this procedure for $j = 1, \ldots, J$, with $J \gg 1$.

The resampling procedure draws a random sample from the *empirical distribution* for the data collection.

The Bootstrap principle

- Each boostrap sample $Y^{(j)}$ can be used to apply some model estimation procedure, each generating a parameter estimate $\widehat{\theta}^{(j)}$.
- We want to use these bootstrapped estimates to say something about the properties of the estimator $\widehat{\theta}$ which is based on the original data Y.
- Idea: The parameter estimate is a deterministic function of the data, the *empirical parameter value* for the observed sample Y: $\widehat{\theta} = \theta(Y)$ and $\widehat{\theta}^{(j)} = \theta(Y^{(j)})$.

The Bootstrap principle

According to the *Bootstrap principle*, the errors of the bootstrapped estimates have the same distribution as the error of $\widehat{\theta}$. In particular, if the true parameter is θ_{true}, then

$$\mathrm{E}(\widehat{\theta} - \theta_{\text{true}}) = \mathrm{E}(\widehat{\theta}^{(j)} - \widehat{\theta}),$$
$$\mathrm{Var}(\widehat{\theta} - \theta_{\text{true}}) = \mathrm{Var}(\widehat{\theta}^{(j)} - \widehat{\theta}).$$

Bootstrap estimation

- The usual expectation and variance estimators can be used:

$$\widehat{\mathsf{E}}(\widehat{\theta} - \theta_{\mathsf{true}}) = \frac{1}{J}\sum_{j=1}^{J}(\widehat{\theta}^{(j)} - \widehat{\theta}) = \overline{\widehat{\theta}^{(\cdot)}} - \widehat{\theta}, \quad \widehat{\mathsf{Var}}(\widehat{\theta} - \theta_{\mathsf{true}}) = \frac{1}{J-1}\sum_{j=1}^{J}\left(\widehat{\theta}^{(j)} - \overline{\widehat{\theta}^{(\cdot)}}\right)^2.$$

- Bias adjusted estimator $\widehat{\theta} - \left(\overline{\widehat{\theta}^{(\cdot)}} - \widehat{\theta}\right)$. Properties? Need *double bootstrap*!
- Confidence intervals for θ_{true}: Consider the quantiles of the error distribution:
  Find a and b such that $\mathsf{P}(a < \widehat{\theta}^{(j)} - \widehat{\theta} < b) = 1 - \alpha$ from the empirical quantiles of the Bootstrap sample residuals $\widehat{\theta}^{(j)} - \widehat{\theta}$.
  According to the Bootstrap principle,

$$\mathsf{P}(a < \widehat{\theta}^{(j)} - \widehat{\theta} < b) = \mathsf{P}(a < \widehat{\theta} - \theta_{\mathsf{true}} < b),$$

  so that

$$\mathsf{P}(\widehat{\theta} - b < \theta_{\mathsf{true}} < \widehat{\theta} - a) = 1 - \alpha,$$

  and a confidence interval is given by

$$\mathsf{CI}_{\mathsf{boot}}(\theta) = \left(\widehat{\theta} - b, \widehat{\theta} - a\right)$$

Alternative confidence interval derivation

Instead of using the empirical quantiles of the boostrap *residuals*, we can use the samples themselves:
Find A and B such that $\mathsf{P}(A < \widehat{\theta}^{(j)} < B) = 1 - \alpha$ from the empirical quantiles of the Bootstrap samples $\widehat{\theta}^{(j)}$.
According to the Bootstrap principle,

$$\mathsf{P}(A < \widehat{\theta}^{(j)} < B) = \mathsf{P}(A - \widehat{\theta} < \widehat{\theta}^{(j)} - \widehat{\theta} < B - \widehat{\theta}) = \mathsf{P}(A - \widehat{\theta} < \widehat{\theta} - \theta_{\mathsf{true}} < B - \widehat{\theta}),$$

so that

$$\mathsf{P}(2\widehat{\theta} - B < \theta_{\mathsf{true}} < 2\widehat{\theta} - A) = 1 - \alpha,$$

and a confidence interval is given by

$$\mathsf{CI}_{\mathsf{boot}}(\theta) = \left(2\widehat{\theta} - B, 2\widehat{\theta} - A\right).$$

Bootstrap is not posterior sampling

- In Bayesian statistics, *Credible Intervals* can be obtained as lower and upper empirical quantiles of samples from the *posterior distribution*.
- It could therefore be tempting to just use the empirical Bootstrap sample quantiles (A, B) as the interval, but this would not work in the presence of Bias or skewness of the Bootstrap distributions.
- Instead, the *Bootstrap confidence interval* construction involves the *upper* quantile of the Bootstrap sample in the *lower* endpoint, and vice versa.

Parametric bootstrap

- If the data size is small, basic Boostrap that resamples with replacement from the raw data has very little information.
- An alternative is to sample entirely new data from the model that was estimated on the whole data set.

Parametric Bootstrap sampling

Let $Y = \{(y_i, x_i), i = 1, \ldots, N\}$ be a data collection with response values y_i and predictors/covariates x_i, such that the conditional data density function is $p(y_i|x_i, \theta)$. Let $\widehat{\theta}$ be the maximum likelihood estimate of θ.

- Define the *Bootstrap sample* $Y^{(j)}$ by drawing N pairs $(y_i^{(j)}, x_i)$, where for each x_i, $y_i^{(j)}$ is drawn from $p(y_i|x_i, \widehat{\theta})$.
- Repeat this procedure for $j = 1, \ldots, J$, with $J \gg 1$.

Depending on the problem, the x_i values can either be kept fixed to their values in the original data set, resampled with replacement, or simulated from some generative model.

Exchangeability

- We can relax model assumptions, e.g. if we don't trust a Gaussian assumption for regression residuals.
- Two model components can be said to be *exchangeable* is swapping their values gives the same joint distribution.
- Instead of assuming a specific residual distribution (usually Gaussian), we can relax the assumption to saying only that the residuals all have the same, but unknown, distribution; the individual residuals are (probabilistically) indistinguishable.
- Instead of resampling the raw data, we resample the model *residuals*.

Residual resampling

Residual resampling in regression models

Let $Y = \{(y_i, x_i), i = 1, \ldots, N\}$ be a data collection with response values y_i and predictors/covariates x_i, such that $\mu_i = \mathsf{E}(y_i|x_i, \theta)$ is the conditional expectation in a regression model. Let $\widehat{\theta}$ be the maximum likelihood estimate of θ, and $\widehat{\mu}_i = \mathsf{E}(y_i|x_i, \widehat{\theta})$.

- Define the residuals $r_i = y_i - \widehat{\mu}_i$, and construct a *residual Bootstrap sample* $Y^{(j)}$ by drawing N pairs $(y_i^{(j)}, x_i)$, where for each x_i, $y_i^{(j)} = \widehat{\mu}_i + r_i^{(j)}$, where $r_i^{(j)}$ is drawn with replacement from the empirical distributon of $\{r_1, r_2, \ldots, r_N\}$.
- Repeat this procedure for $j = 1, \ldots, J$, with $J \gg 1$.

For x_i, the same options as for fully parametric Bootstrap are available.

Randomisation/permutation tests

When assessing differences between two statistical populations, the hypotheses often take the form of some kind of *exchangeability structure*.

Exchangeability test example

- Let $Y_A = \{y_1^A, \ldots, y_{N_A}^A\}$ and $Y_B = \{y_1^B, \ldots, y_{N_B}^B\}$, and we want to test the hypotheses

  H_0 : The A and B come from the same distribution

  H_1 : The A and B do not come from the same distribution

- Given a test statistic $T(Y_A, Y_B)$, such as $\overline{y_\cdot^A} - \overline{y_\cdot^B}$, we need the distribution of T under H_0.
- Under H_0, the joint sample $Y_{A\cup B} = \{y_1^A, \ldots, y_{N_A}^A, y_1^B, \ldots, y_{N_B}^B\}$ is a collection of exchangeable variables.
- Each random permutation of $Y_{A\cup B}$ has the same distribution as $Y_{A\cup B}$, under H_0 (but not under H_1).

Randomisation/permutation tests

Assume that large test statistics T indicate deviations from H_0.

Permutation tests

- For $j = 1, \ldots, J$, draw $Y_{A\cup B}^{(j)}$ as a random permutation of $Y_{A\cup B}$, and split the result into subsets $Y_A^{(j)}$ and $Y_B^{(j)}$ of size N_A and N_B, respectively.
- Compute the test statistics $T^{(j)} = T(Y_A^{(j)}, Y_B^{(j)})$.
- The average $\frac{1}{J}\sum_{j=1}^{J} \mathbb{I}\{T^{(j)} \geq T(Y_A, Y_B)\}$ is an unbiased estimator of the p-value w.r.t. T for the hypothesis H_0 : the elements of Y_A and Y_B are mutually exchangeable.
- If $N_A + N_B$ is small, the limit $J \to \infty$ can be obtained by using all possible permutations instead of independent random permutations.

- In other exchangeability situations, similar tests can be designed.

Randomisation/permutation test for proper scores

- In a collection of prediction scores for two models A and B, each data point has its own predictive distribution, so we don't have full exchangeability.
- We do have *pairwise exchangeability* if the two model predictions are equivalent.
- To construct a formal test, we randomise the scores within each pair. If we investigate the difference between scores, $S_i^\Delta = S(F_i^A, y_i) - S(F_i^B, y_i)$, $i = 1, \ldots, N$, that means swapping the sign of the difference. For testing the average difference, we can use test statistic

$$T(\{S_i^\Delta\}) = \frac{1}{N} \sum_{i=1}^{N} S_i^\Delta$$

- For each $j = 1, \ldots, J$ and $i = 1, \ldots, N$, draw $S_i^{\Delta(j)} = S_i^\Delta$ with probability 0.5, and $-S_i^\Delta$ with probability 0.5.
- Compute the test statistics, $T^{(j)} = T(\{S_i^{\Delta(j)}, i = 1, \ldots, N\})$.
- The average $\frac{1}{J} \sum_{j=1}^{J} \mathbb{I}\{T^{(j)} \geq T(\{S_i^\Delta\})\}$ is an unbiased estimator of the one-sided p-value w.r.t. T for the hypothesis
  H_0 : the scores of the two models are pairwise exchangeable vs.
  H_1 : Model B is better than A