Cross-Validation

Dimitris Fouskakis

Department of Mathematics
School of Applied Mathematical and Physical Sciences
National Technical University of Athens

fouskakis@math.ntua.gr

Spring Semester

Introduction

- Method for measuring the goodness-of-fit of a model to our data.
  Extension of this idea: compare different models and choose the best.
  Example: application to kernels
- Given some data:
  1. we could fit our model to the data (estimate the parameters) and
  2. use the coefficient of determination to measure how good it is

  Problem: we use the data twice!
- As a solution, we could leave some observations out of the estimation process and use them later to test if our model is good.

  Disadvantages: 1) not all the data is used
  2) our result w.r.t. the goodness-of-fit of the model might depend on which observations were left out.

Method description

Alternatively $\rightarrow$ cross-validation (CV).

We leave one observation out each time and then make a prediction for this (left-out) observation based on the model we fitted (using the rest of the observations). Afterwards, we repeat the procedure leaving another observation out till we cover the whole space. $\rightarrow$ we compute an overall score from every choice.

e.g. Random sample (X_i, Y_i) $i = 1, 2, \ldots n$

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \epsilon_i = g(x_i) + \epsilon_i \\ Y_i &= \alpha' + \beta' X_i + \gamma X_i^2 + \epsilon_i' = h(x_i) + \epsilon_i' \end{aligned}$$

$\hat{g}_{-i}(x_i) \rightarrow$ prediction for x_i based on the model $g(x_i) + \epsilon_i$ when we have left the i observation out.
$\hat{h}_{-i}(x_i) \rightarrow$ same but for the second model.

Then $y_i - \hat{g}_{-i}(x_i)$ and $y_i - \hat{h}_{-i}(x_i)$ are the prediction errors for each model (studentized or deleted residuals).

Criterion

$$CV(g) = \frac{\sum_{i=1}^{n} \left(y_i - \hat{g}_{-i}(x_i)\right)^2}{n}$$

$$CV(h) = \frac{\sum_{i=1}^{n} \left(y_i - \hat{h}_{-i}(x_i)\right)^2}{n}$$

The quantity $nCV(\cdot) \equiv \mathrm{PRESS} \rightarrow$ prediction error sum of squares

If $CV(g) < CV(h)$ then g has a better fit than h since it has a smaller PRESS.

PRESS vs RSS

PRESS vs classical residual sum of squares ($RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$)

1. CV involves many computations
2. CV is robust against outliers
3. Nested models

$$\alpha + \beta X_{1i} + \epsilon, \quad \alpha + \beta X_{1i} + \gamma X_{2i} + \epsilon$$

RSS becomes smaller each time we add a new variable independently of how good this is. So, the more complex models are better which is not correct (parsimony principle). With PRESS this does not happen.

4. The choice of the function in the CV criterion is free, e.g. $\frac{\sum |y_i - \hat{g}_{-i}(x_i)|}{n}$
5. PRESS also shows us the predictive power of our model
6. PRESS can also be used in generalized linear models

Cross-validation in linear regression

In the linear model case, the computations can be reduced to the minimum.
Let $\mathbf{Y} = \tilde{\mathbf{X}}\beta$ where $\mathbf{Y}$ is a $(n \times 1)$ vector with the values of the response variable, $\tilde{\mathbf{X}}$ is a $n \times (p+1)$ design matrix with the p explanatory variables (1st column has 1's) and β is a $((p+1) \times 1)$ vector with the coefficients.

$$\begin{aligned} \hat{\beta} &= \left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{Y} \\ \hat{\mathbf{Y}} = \tilde{\mathbf{X}}\hat{\beta} &= \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{Y} = \mathbf{A}\mathbf{Y}\,, \end{aligned}$$

where $\mathbf{A} = (\alpha_{ij}) = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T \in \mathbb{M}_{nxn}$: hat matrix

It can be shown that $y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1-\alpha_{ii}}$.

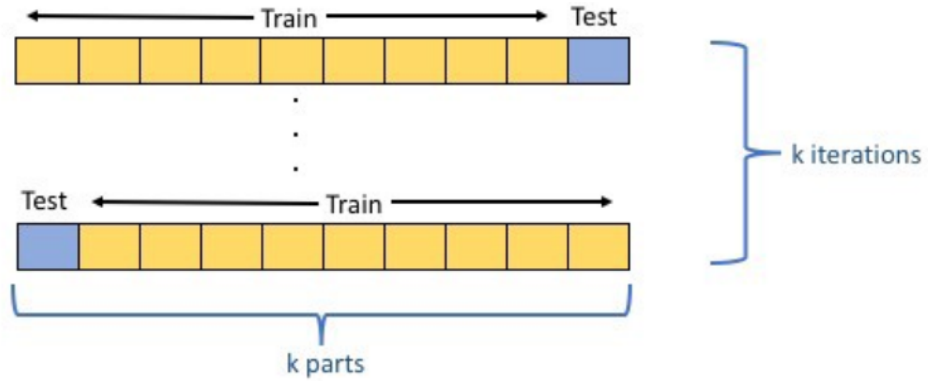
It thus suffices to fit the model only once and save the values of $\mathbf{A}$.

α_{ii}: leverages (measures how influential one observation is in the model estimation).

In a similar way, the matrix $\mathbf{A}$ also appears in the non-parametric regression (kernels), so the CV can also be used there as well without big computational cost.

k-fold cross-validation

- Method just described: leave-one out CV
- Generalization: leave-k out CV
  use $n - k$ observations (e.g. 2/3) for modeling and the remaining k for validation (e.g. 1/3). There are $\binom{n}{k}$ groups in total. Perform N repetitions so that Monte Carlo s.e. is small.
- k-fold CV
  1. Divide data into k equal sized folds ($\approx n/k$ obs in each)
  2. $(k-1)$ folds used for modeling (or training)
  3. 1 fold used for validation (or testing)
  4. Repeat steps 1-3, k times. In each time calculate the average (squared/absolute) prediction error in the validation fold.



Average or combine k results. Usually $k = 10$.