Optimal Cross-Validation Split Ratio: Experimental Investigation

Cyril Goutte and Jan Larsen
Department of Mathematical Modelling
Technical University of Denmark, Lyngby, Denmark
E-mail: cg,jl@imm.dtu.dk

Abstract

Cross-validation is a widespread method for assessing the generalisation ability of a model in order to tune a regularisation parameter or other hyper-parameters of a learning process. The use of cross-validation requires to set yet an additional parameter, the *split ratio*. Few texts have investigated theoretically the asymptotic setting of this ratio, and no consensus has emerged. In this contribution, we investigate the sensitivity and optimal setting of the split ratio on a particular model, a non-parametric kernel estimator with adaptive metric.

1 Cross-validation

Most efficient learning procedures require the setting of an extra learning parameter, or "hyper-parameter". Neural networks typically use a regularisation parameter weighting a weight decay [1], or the extent of pruning [2]. Estimating the "optimal" hyper-parameter is the topic of active current research in the statistical learning community [3]. Let us consider a typical learning problem: modelling an input-output relationship based on some empirical data $\mathcal{D} = \left(x^{(i)}, y^{(i)}\right)_{i=1,\ldots N}$ sampled from an unknown, but fixed joint probability $p(x, y)$. The performance of a candidate model f is measured by the generalisation error or expected risk:

$$G(f) = \int \ell\left(x, y, f\right)\ p(x, y)\,\mathrm{d}x \tag{1}$$

where the risk $\ell\left(x, y, f\right)$ measures how "close" the model is to the data. For regression, the usual choice is the quadratic difference: $\ell\left(x, y, f\right) = \left(y - f(x)\right)^2$. A principled approach to setting the hyper-parameter is to minimise the generalisation error, or an estimate thereof, as (1) can usually not be calculated due to the unknown input-output distribution. Apart from the algebraic estimators in line with Akaike's FPE [4], several data-intensive alternatives are commonly used. The *split sample* method sets aside V of our N data points to provide an empirical estimator of (1):

$$\widehat{G}_{SS} = \frac{1}{V}\sum_{i=1}^{V} \ell\left(x^{(i)}, y^{(i)}, f^{-V}\right) \tag{2}$$

where for notational purpose (and without loss of generality) the first V data are left out for validation. f^{-V} is the model estimation based on the remaining data, $i = V+1, \ldots N$. Large values of the split ratio $\gamma = V/N$ leave few data for model estimation (eg neural network training) while small values lead to an unreliable generalisation estimator. *Cross-validation*[1] [5] attempts to increase this reliability by averaging over several splits. In *L-fold cross-validation*, $\mathcal{D}$ is split in L disjoint subsets $(S_j)_{j=1,\ldots L}$ of roughly equivalent size, $\bigcup_{j=1}^{N} S_j = \mathcal{D}$. The validation error (2) calculated on each S_j, using the remaining data for model estimation, is averaged over the subsets:

$$\widehat{G}_{CV} = \frac{1}{N} \sum_{j=1}^{L} \sum_{i \in S_j} \ell\left(x^{(i)}, y^{(i)}, f^{-j}\right) \tag{3}$$

where f^{-j} is the model using $\bigcup_{i \neq j} S_i$. Equation (3) estimates the *average* generalisation error. Cross-validation supposedly enhances the reliability of the generalisation estimator at a cost of increased computational requirements. The split ratio becomes $\gamma = 1/L$, yet another parameter that potentially influences the performance of the resulting model. Little theoretical analysis has been devoted to its setting: [6] shows that for linear model selection, γ should asymptotically tend to 1, a surprising result that indicates that the relative amount of data left for model estimation should tend to 0. A similar result was obtained by [7] in a different setting, while a recent communication [8] suggests that small split ratios should be favoured for the split sample method. Accordingly, we expect the optimal γ to either approach 1 or 0 when N increases. The experiments presented in section 3 address this important issue.

2 Adaptive metric kernel regression

Adaptive metric kernels are a natural extension to the standard non-parametric kernel estimator [9]. Estimation is performed on the basis of the available data:

$$\widehat{y} = f_{\mathbf{H}}(x) = \frac{\sum_{k=1}^{N} K\left(d_{\mathbf{H}}\left(x, x^{(k)}\right)\right) y^{(k)}}{\sum_{k=1}^{N} K\left(d_{\mathbf{H}}\left(x, x^{(k)}\right)\right)} \tag{4}$$

where $d_{\mathbf{H}}^2(u,v) = (u-v)^\top \mathbf{H}(u-v)$ defines the metric in the input space. For P-dimensional data, $\mathbf{H}$ is a $P \times P$ positive definite *smoothing matrix*. $\mathbf{H}$ can be parameterised by its Cholesky decomposition, or the more parsimonious diagonal parameterisation presented in [10] and adopted here: $\mathbf{H} = \operatorname{diag}(h_1^2, h_2^2, \ldots h_P^2)$. The smoothing parameters h_j are adapted by minimising the cross-validation error:

$$\widehat{G}_{CV} = \frac{1}{N} \sum_{j=1}^{L} \left[\sum_{i \in S_j} \left(y^{(i)} - f_{\mathbf{H}}^{-j}\left(x^{(i)}\right)\right)^2 \right] \tag{5}$$

[1] The split sample method is sometimes referred to as "cross-validation" in the neural networks literature. This is inconsistent with the definition of [5] and we will here reserve the term exclusively to the averaging method (3).

where $\mathcal{D}$ is split into L subsets S_j and $f_{\mathbf{H}}^{-j}$ is the estimation excluding S_j:

$$f_{\mathbf{H}}^{-j}(x) = \frac{\sum_{k \notin S_j} K\left(d_{\mathbf{H}}\left(x, x^{(k)}\right)\right) y^{(k)}}{\sum_{k \notin S_j} K\left(d_{\mathbf{H}}\left(x, x^{(k)}\right)\right)} \tag{6}$$

and the derivatives of (5) with respect to smoothing parameter h_p is [10]:

$$\frac{\partial \widehat{G}_{CV}}{\partial h_p} = -\frac{2}{N} \sum_{j=1}^{L} \sum_{i \in S_j} \varepsilon_{ii} \left(\frac{\sum_{k \notin S_j} \varepsilon_{ki} \dfrac{\partial K_{ik}}{\partial h_p}}{\sum_{k \notin S_j} K_{ik}} \right) \tag{7}$$

with $\varepsilon_{ki} = (y^{(k)} - f_{\mathbf{H}}^{-j}(x^{(i)}))$ and $K_{ik} = K(d_{\mathbf{H}}(\mathbf{x}^{(i)}, \mathbf{x}^{(k)}))$. The smoothing matrix is adapted by first order multi-dimensional minimisation of $\widehat{G}_{CV}$, using a conjugate gradient algorithm with approximate line search [11, appendix B].

3 Experiments

Let us first consider a system inspired from [12] and described by:

$$y = 10 \sin(\pi x_1 x_2) + 20 \left(x_3 - \frac{1}{2} \right)^2 + 10 x_4 + 5 x_5 + \epsilon \tag{8}$$

with iid noise $\epsilon \sim \mathcal{N}(0,1)$. The input vector x contains 10 uniformly distributed values, $x_1 \ldots x_{10}$, $x_i \sim \mathcal{U}([0,1])$, so that x_6 to x_{10} are irrelevant. 10 dimensions is rather large in our context: in order to uniformly sample the input space in 5 evenly spaced locations on each dimensions (eg 0, 0.25, 0.5, 0.75 and 1), nearly 10^7 data points are required. We have shown in previous experiments that the adaptive metric handles the irrelevant inputs efficiently [10].

We consider 4 increasing sample sizes: 100, 200, 500 and 1000 points. A large, 2^{13} points test set is generated independently from (8) in order to assess the generalisation ability. For each sample size N, we generate a number of datasets and investigate the following split ratios: $1/N$ (aka *leave-one-out* (LOO) cross-validation), 0.1 (the popular 10-fold cross-validation), 0.2, 0.3, 0.4 and 0.5. For each sample and each split ratio, the smoothing parameters are adapted by minimising (5). We estimate the resulting generalisation on the test set. Though only one value of γ yields minimum generalisation, the differences might not be significant: this effect is assessed using a paired t-test on the squared residuals obtained from the test set. Accordingly, for each experiment, we generate 6 p-values assessing the significance of the difference in generalisation between each split ratio and the optimum.

Figure 1 summarises the results: for each sample size, we show the empirical distribution of the optimal split ratio. This suggests that as the number of available data grows, the optimal split ratio tends to decrease: the average optimal split ratio decreases from 27% for $N = 100$ to 11% for $N = 1000$. A χ^2-test shows that the empirical distribution departs from uniform (at a 10^{-3} level) for $N > 100$. Another way of comparing the different split ratios is to

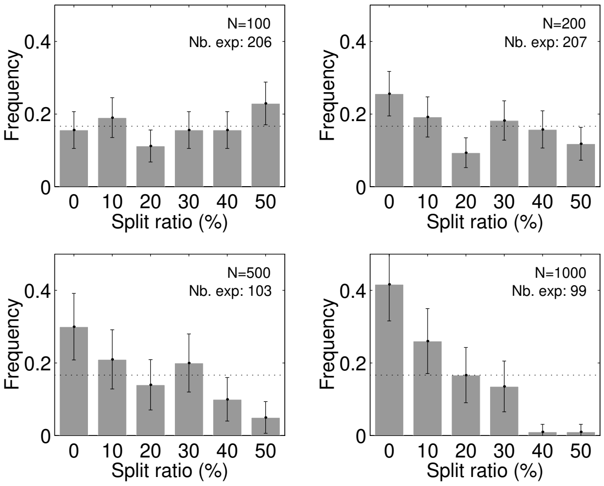

Figure 1: Empirical distribution of the generalisation-optimal split ratio for four sample sizes. Error bars indicate 2 standard deviations, the x-axis label '0' is the LOO case ($\gamma = 1/N$) and the dotted line corresponds to the uniform distribution. Read: for $N = 1000$, LOO was optimal in more than 40% of the experiments.

consider the *average* generalisation error (AGE). Figure 2 shows the percentage by which the AGE for each γ exceeds the minimum AGE for this sample size. Eg for $N = 500$, $\gamma = 0.4$ yields on average 1% higher AGE than $\gamma = 1/N$. While figure 1 shows that optimal generalisation is achieved for a fair number of $\gamma > 0.3$ for law sample sizes, figure 2 suggests that small split ratios minimise the AGE in all cases.

Figure 4 sheds some light over this intriguing effect. In each sub-plot we consider the experiments for which the corresponding split ratio is either optimal, or non significantly worse at a 5% level according to the p-values mentioned above. The curve shows the excess generalisation for each split ratio compared to the optimal one (which obviously has a 0 excess). This plot was obtained for $N = 200$ but the other sample sizes display similar effects. This shows that when small split ratios are optimal (top row), choosing a large γ is heavily penalised, while choosing a small ratio when a large one is optimal (bottom row) yields a limited excess in generalisation error. While large split ratios might lead to the smallest generalisation error in a number of cases, the will perform badly on average.

These experiments were carried out for a specific model and a specific (artificial) problem, raising the question of how general the results are. We could argue that as theoretical work suggests a clear asymptotic behaviour (towards 1 or 0), one experiment where γ tends towards 0 is sufficient evidence against the alternative hypothesis. However, in order to show that this behaviour is not dependent on this particular artificial problem or experimental conditions (noise level, variable interaction), we have carried out the same set of experiments on 5 problems studied in [13]. Contrary to (8), which combines modelling and input selection, we will now tackle pure modelling problems (we refer the reader

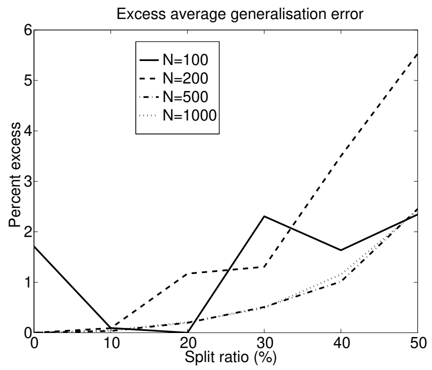

Figure 2: Excess (in percents) average generalisation error, for the 4 sample sizes. Label '0' is the LOO ($\gamma = 1/N$).

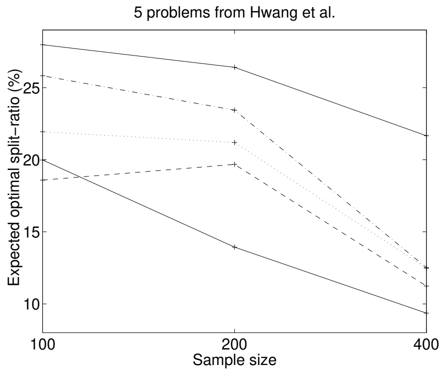

Figure 3: Evolution of the average optimal split-ratio with increasing sample size for 5 the problems of [13].

to [13] and [14] for a description). One aspect of the results is presented in figure 3, where we plot the expected optimal γ, ie the average of the empirical distribution of the generalisation optimal split ratio (cf figure 1) for our five problems (hence 5 curves). Figure 3 shows that on these problems again the optimal γ tends to decrease with increasing sample sizes. This suggests that the pattern exhibited with the systematic study of our first problem is highly reproducible. Due to the limits imposed on this paper, we will reserve more detailed results and comments to the conference.

4 Conclusion

This contribution addresses the use of cross-validation for optimising various hyper-parameters in non-linear modelling. Optimal setting of hyper-parameters, eg regularisation parameters for neural networks or smoothing parameters in kernel regression, is crucial for obtaining high generalisation ability. Cross-validation involves an additional parameter, the split ratio or fraction of data used for validation. As theoretical studies indicate that the optimal split ratio is very dependent on the modelling scenario and other assumptions, we pursue an experimental investigation. We model a number of artificial problems of various difficulties using adaptive metric kernel regression. The results of these experiments suggest the following conclusions: 1) The optimal split ratio tends to zero as the amount of data grows; 2) Small split ratios seem to be a safe bet, even when the optimal ratio is fairly large; 3) The results depend on the measure of optimality, eg generalisation error or average thereof.

Acknowledgements This work was supported by BIOMED II grant number BMH4-CT97-2775 and by the Danish Research Councils through the Computational Neural Network Center and Center for Neuroinformatics. JL furthermore acknowledges the Radio Parts Foundation for financial support.

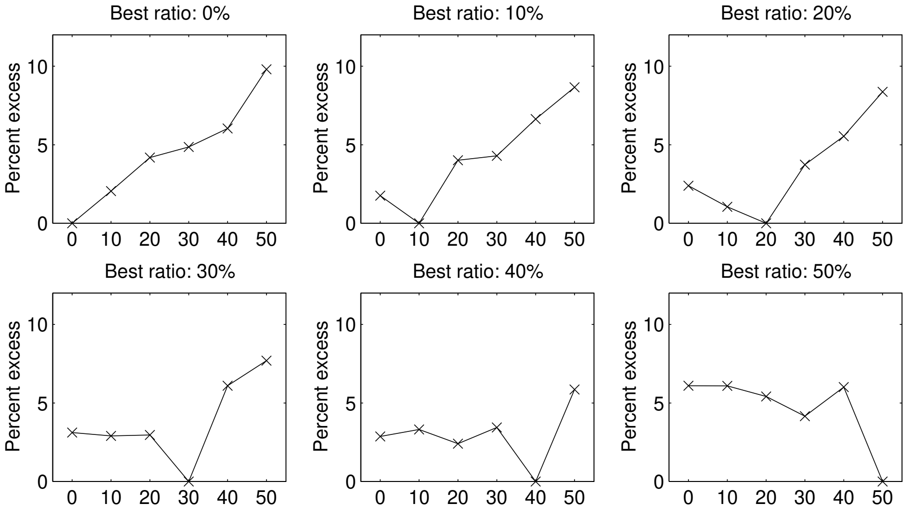

Figure 4: Average excess generalisation errors associated with each γ compared to the optimal choice ($N = 200$). '0' stands for LOO. Read: when $\gamma = 30\%$ is optimal (bottom left), LOO yields 3% excess generalisation, compared to 8% for $\gamma = 50\%$.

References

[1] Krogh A, Hertz JA, A simple weight decay can improve generalization. In: Moody et al. (eds), Advances in Neural Inf Proc Systems 4, 1992, pp 950–957

[2] Le Cun Y, Denker JS, Solla SA, Optimal brain damage. In: Touretzky (ed), Advances in Neural Information Processing Systems 2, 1990, pp 598–605

[3] Larsen J, Svarer C, Andersen LN, Hansen LK, Adaptive regularization in neural network modeling. Orr et al. (eds), The Book of Tricks, Springer-Verlag, 1998

[4] Akaike H, Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics, 1969, 21:243–247

[5] Stone M, Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society B, 1974, 36:111–147, with discussion

[6] Shao J, Linear model selection by cross-validation. Journal of the American Statistical Association, 1993, 88:486–494

[7] Larsen J, Hansen LK, Empirical generalization assessment of neural network models. In: Girosi et al. (eds), NNSP V – Proc IEEE Workshop, 1995, pp 42–51

[8] Kearns M, A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. Neural Computation, 1997, 9:1143–1161

[9] Härdle W, Applied nonparametric regression. Cambridge University Press, 1990

[10] Goutte C, Larsen J, Adaptive metric kernel regression. In: NNSP VIII – Proceedings of the IEEE Workshop, 1998, in press

[11] Rasmussen CE, Evaluation of Gaussian processes and other methods for non-linear regression. PhD thesis, Dept Computer Science, Univ of Toronto, 1996

[12] Friedman J, Multivariate adaptive regression splines. Ann Stat 1991, 19:1–141

[13] Hwang JN, Lay SR, Maechler M et al., Regression modelling in back-propagation and projection pursuit learning. IEEE Trans Neural Networks, 1994, 5:342–353

[14] Neal R, `http://www.cs.utoronto.ca/~delve/data/hwang/hwangDetail.html`