Lecture Note
COS 702
Choosing a good shape parameter using cross validation

Let us consider, in R^2, a set of interpolating points $S = \{(x_i, y)\}_{i=1}^{N}$. For any function $b(x, y)$, it may be approximate by MQ

$$b(x,y) \simeq \sum_{i=1}^{N} a_i \varphi_i(r) = \sum_{i=1}^{N} a_i \sqrt{r_i^2 + c^2} \tag{1}$$

where $r_i = \|(x, y) - (x_i, y_i)\|$ and the free parameter $c^2 \geq 0$ is referred as the shape parameter. We would like to find a good shape parameter using cross validation.

Cross validation has been used for many years as a standard technique for model selection and the determination of model performance in statistics and our algorithm for finding the optimal shape parameter c is based on the exclusion (cross validation) algorithm of Milroy et. al. with modifications to improve the efficiency of the computation. Their approach is quite reasonable and reliable and yet computational inefficient. We will show how to improve their algorithm by applying a standard statistical scheme.

First, we set aside a point (x_i, y_i) from S and use the remaining $N - 1$ points to form the basis functions $\left\{\sqrt{r_j^2 + c^2}\right\}_{j=1, j\neq i}^{N-1}$ and estimate the coefficients $\{a_j\}_{j=1,j\neq i}^{N-1}$ in (1). We then use this estimate to approximate the function value at the deleted point and compute the error between the predicted and actual value of b at (x_i, y_i); i.e.,

$$e_{i,-i} = b(x_i, y_i) - \widehat{b}_{i,-i}(x_i, y_i);$$

where $e_{i,-i}$ and $\widehat{b}_{i,-i}$ are the prediction error and the approximate value of b at (x_i, y_i), respectively, when (x_i, y_i) is removed from the fitting set. We then repeat the above procedure N times. Thus for a given shape parameter c, the model will be fitted N times.

In the statistical literature, PRESS (Prediction Residual Error Sum of Squares) is defined as

$$\text{PRESS} = \sum_{i=1}^{N} (e_{i,-i})^2. \tag{2}$$

It is reasonable, in the statistical sense, to choose the shape parameter c giving the smallest PRESS value.

Using collocation, one can see that the above algorithm may be tedious and inefficient. For each given shape parameter c, there are N prediction errors $e_{i,-i}$ to be calculated and each time a system of equations $N - 1$ needs to be solved. When N becomes large, this approach is not practical.

To remedy this drawback of the above approach, we propose the use of the method of least-squares. Instead of choosing one collocation data set, we choose two distinct uniformly distributed point sets $S = \{(x_i, y_i)\}_{i=1}^{N}$ and $T = \{(\widehat{x}_j, \widehat{y}_j)\}_{j=1}^{N+m}$ where $m \geq 1$. The first set S serves to define the basis functions in (1). The second set T serves as the fitting points for cross validation. During the cross validation process, one data point at a time from T will be set aside as we have stated earlier. Thus we have the following $(N+m) \times N$ system of equations to solve;

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1N} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{(N+m)1} & \varphi_{(N+m)2} & \cdots & \varphi_{(N+m)N} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} b(\widehat{x}_1, \widehat{y}_1) \\ b(\widehat{x}_2, \widehat{y}_2) \\ \vdots \\ b(\widehat{x}_N, \widehat{y}_N) \end{bmatrix},$$

where $\varphi_{ij} = [(\widehat{x}_j - x_i)^2 + (\widehat{y}_j - y_i)^2 + c^2]^{1/2}$. In matrix notation, we have

$$D \cdot A = B. \tag{3}$$

Removing a data point $(\widehat{x}_i, \widehat{y}_i)$ in the cross validation process is equivalent to removing the ith row from D and B. By applying the Sherman-Morrison-Woodbury Theorem, it is well-known that the computation of PRESS in (2) is quite simple and does not require one to repeatedly remove a row from D and B.

If D^T denotes the transpose of the matrix D and d_i is the ith row of D, then we can calculate the PRESS residual $e_{i,-i}$ from the original matrix D without removing a row from it; i.e.,

$$e_{i,-i} = \frac{b(\widehat{x}_i, \widehat{y}_i) - \widehat{b}(\widehat{x}_i, \widehat{y}_i)}{1 - d_i (D^T D)^{-1} d_i^T} = \frac{e_i}{1 - h_{ii}}, \tag{4}$$

where $\widehat{b}(\widehat{x}_i, \widehat{y}_i) = d_i (D^T D)^{-1} D^T B$ and h_{ii} is the ith diagonal element of the HAT matrix $H = D(D^T D)^{-1} D^T$. [We remark that $D^{\dagger} = (D^T D)^{-1} D^T$ is the psedoinverse of D which can be computed by either the QR decomposition or singular value decomposition (SVD), using software such as MATHEMATICA or MATLAB.] By the definition of PRESS in (2), we have

$$\text{PRESS} = \sum_{i=1}^{N+m} \left(\frac{e_i}{1 - h_{ii}} \right)^2. \tag{5}$$

Equation (5) is quite remarkable in terms of computational cost. It enables one to carry out the algorithm efficiently and choose the optimal value (or a consistently 'good' value) for the shape parameter c from a series of tests. The extra effort needed to select the optimal c is worthwhile, as our numerical results show several orders of magnitude improvement over previous ones.

A similar approach using only one data set can be established. If $b^{[k]}$ is the radial basis function interpolant to the data $\{f_1, f_2, \cdots, f_N\}$, i.e.,

$$b^{[k]}(\mathbf{x}) = \sum_{j=1, j\neq k}^{N} a_j^{[k]} \varphi\left(\|\mathbf{x} - \mathbf{x}_j\|\right)$$

such that

$$b^{[k]}(\mathbf{x}_i) = f_i, \quad i = 1, \cdots, k-1, k+1, \cdots, N,$$

and if E_k is the error

$$E_k = f_k - b^{[k]}(\mathbf{x}_k)$$

at the one point $\mathbf{x}_k$ not used to determine the interpolant, then the quality of the fit is determined by the norm of the vectors $E = [E_1, \cdots, E_N]^T$ obtained by removing in turn one of the data points and comparing the resulting fit with the known value at the removed points.

By adding a loop over c we can compare the error norms for different values of the shape parameter, and choose, and choose that value of c that yield the minimal error norm at the optimal one.

The disadvantage of the above leave-one-out algorithm is the computational cost is rather expensive. Rippa [1] showed that the computation of the error components can be simplified to a single formula

$$E_k = \frac{a_k}{A_{kk}^{-1}}, \tag{6}$$

where a_k is the kth coefficient in the interpolant b based on the full data set, and A_{kk}^{-1} is the kth diagonal element of the inverse of the corresponding interpolation matrix. Since both a_k and A^{-1} need to be computed only once for each value of c, this results in $O(N^3)$ computational complexity.

Note that the error vector E can be computed in a single statement in MATLAB if we vectorize the component formula (6).

To further increase the efficiency, we can use "golden search" or "bisection search" to find shape parameter.

References

[1] S. Rippa, An algorithm for selecting a good value for the parameter c in radial basis function interpolation, Adv. in Compu. Math. 11, pp. 193-210, 1999.