Generalized Cross Validation

Mårten Marcus

26 November 2009
Plan

1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
Plan

1 Generals

2 What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3 Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4 Discussion
References

- Chapter 4 of *Spline models for Observational Data (1990)* - Grace Wahba

- *Optimal Estimation of Contour Properties by Cross-Validated Regularization (1989)* - Behzad Shahraray, David Anderson

- *Smoothing Noisy Data with Spline Function (1979)* - Peter Craven, Grace Wahba
Conditional Expectations

From probability theory, we have for \( X \in \mathcal{L}^2\{\Omega, \mathcal{F}, \mathbb{P}\} \)

\[
\mathbb{E}[X] = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}[(X - \theta)^2]
\]  

(1)

The least squares estimate therefore gives discretized estimate of \( \theta \) and a natural norm to choose when searching for data disturbed by white noise.
Ill-posedness

Consider the model \( y_i = g(t_i) + \epsilon_i, \ i = 1, 2, \ldots, n, \ t_i \in [0, 1] \) where \( g \in W_2^{(m)} = \{f | f', f'', \ldots, f^{(m-1)} \} \text{abs.cont., } f^{(m)} \in L_2[0, 1] \) and \( \{\epsilon_i \sim WN(0, \sigma^2)\} \) where \( \sigma \) is unknown.

The least squares estimate gives

\[
\min_{\hat{g}_i \in W_2^{(m)}} \sum_{i=1}^{n} (y_i - \hat{g}_i)^2 = 0 \quad \forall \{y_i\}_{i=1}^{n} \quad \tag{2}
\]

The minimum does not depend on data and is clearly ill-posed.
Regularization

We may choose to regularize, or “smooth”, the data as

$$\hat{g}_{n,\lambda} = \arg\min_{\hat{g}_i \in W_2^{(m)}} \sum_{i=1}^{n} (y_i - \hat{g}(t_i))^2 + \lambda \int_{0}^{1} (\hat{g}^{(m)}(t))^2 dt \quad \lambda \geq 0$$

which has a unique solution.
Plan

1 Generals

2 What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3 Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4 Discussion
Properties of $\hat{g}_{n,\lambda}$

The reason for the term *smoothing spline* is the following:

- For $\lambda = 0$, $\hat{g}_{n,\lambda}$ can be seen as an interpolating spline.
- For $\lambda = \infty$, $\hat{g}_{n,\lambda}$ is a single polynomial of degree $m - 1$ (optimal in the least squares sense).
- For $0 < \lambda < \infty$, it can be shown that $\hat{g}_{n,\lambda}$ is composed by polynomials of degree at most $2m - 1$ on the intervals $[t_i, t_{i+1}]$, $i \in \{1, 2, \ldots, n - 1\}$ such that the function and its derivatives up to and including the $2m - 2$ derivative are continuous at the knots, and $f^{(k)}(t_1) = f^{(k)}(t_n) = 0$ for $k = \{m, m + 1, \ldots, 2m - 2\}$, i.e. natural conditions in the end points.

From which we can see that there is a strong dependence between the quality of the result and a good choice of $\lambda$. 
Plan

1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
Example of impact of different $\lambda$ on $\hat{g}_{n,\lambda}$ and $\hat{g}'_{n,\lambda}$ for $y_i = g(t_i) + \epsilon_i$ where $g(t) = \sin\left(\frac{\pi t}{180}\right)$, $t \in [0, 360]$
Defining the Optimal $\lambda$

Plan

1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
Defining the Optimal $\lambda$

True Mean Square Error $R(\lambda)$ and Optimal $\lambda$

Using the notation above we define the true mean square error, $R(\lambda)$, as

$$R(\lambda) := \frac{1}{n} \sum_{i=1}^{n} (\hat{g}_{n,\lambda}(t_i) - g_{t_i})^2$$  \hspace{1cm} (4)$$

The optimal $\lambda$ is then defined as

$$\lambda^* = \arg\min_{\lambda \in \mathbb{R}^+} R(\lambda)$$  \hspace{1cm} (5)$$
Plan

1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
Cross Validation

Plan

1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
Given our sample of $n$ measurements, we leave one measurement out and predict that data point by the use of the $n - 1$ remaining measurements.

The idea is now that the best model for the measurements is the one that best predicts each measurement as a function of the others.
The Ordinary Cross-Validation (OCV)

**Definition**

Let \( \hat{g}_{n,\lambda}^k \), \( k \in \{1, 2, \ldots, n\} \) be defined as

\[
\hat{g}_{n,\lambda}^{[k]} = \arg\min_{\hat{g}_i \in W_2^{(m)}} \sum_{\substack{i=1 \\, i \neq k}}^{n} (y_i - \hat{g}(t_i))^2 + \lambda \int_{0}^{1} (\hat{g}^{(m)}(t))^2 dt \quad \lambda \geq 0 \quad (6)
\]

Then we define the OCV mean square error as

\[
V_0(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\hat{g}_{n,\lambda}(t_i) - y_i)^2 \quad (7)
\]
The Ordinary Cross-Validation (OCV)

Definition

Let $V_0(\lambda)$ be defined as in the previous definition, the Ordinary Cross-Validation Estimate of $\lambda$ is

$$
\lambda_0 = \arg\min_{\lambda \in \mathbb{R}^+} V_0(\lambda) 
$$

(8)
Why we need more

The OCV estimate have proven good accuracy for certain periodic functions, but in general it lacks a feature shown by the following: Consider the function \( \{g(t) \in W^2, \quad g(t) = g(t + 1)\} \) sampled equidistantly e.g. \( \{t_j = \frac{j}{n}, \quad j \in \{1, 2, \ldots, n\}\} \) adding some uniform noise.

In this case all data are treated symmetrically. On the other hand, dropping the conditions of periodicity and equidistant sampling, the points interact differently, giving rise to the need of weighting the samples differently.
1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
Generalized Cross Validation (GCV)

Rewriting $V_0$

There exists an $n \times n$ matrix $A(\lambda)$, the *influence matrix*, with the property

$$
\begin{bmatrix}
\hat{g}_{n,\lambda}(t_1) \\
\hat{g}_{n,\lambda}(t_2) \\
\vdots \\
\hat{g}_{n,\lambda}(t_n)
\end{bmatrix} = A(\lambda)y
$$

(9)

Such that $V_0(\lambda)$ can be rewritten as

$$
V_0(\lambda) = \frac{1}{n} \sum_{k=1}^{n} \left( \sum_{i=1}^{n} (a_{kj}y_j - y_k)^2 \right) \frac{1}{(1-a_{kk})^2}
$$

(10)

Where $a_{kj}, k, j \in 1, 2, \ldots, n$ is element $\{k,j\}$ of $A(\lambda)$
Generalized Cross Validation (GCV)

The Generalized Cross Validation (GCV)

Definition

Let $A(\lambda)$ be the influence matrix defined above, then the GCV function is defined as

$$V(\lambda) = \frac{1}{n} \left\| (I - A(\lambda))y \right\|^2 \frac{1}{n} tr(I - A(\lambda))^{-2}$$ (11)

We say that the Generalized Cross-Validation Estimate of $\lambda$ is

$$\bar{\lambda} = \arg\min_{\lambda \in \mathbb{R}^+} V(\lambda)$$ (12)
Generalized Cross Validation (GCV)

Similarity to OCV

By rewriting $V(\lambda)$ we get

$$V(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\hat{g}_{n,\lambda}(t_i) - y_i)^2 w_k(\lambda)$$

(13)

where $w_k(\lambda)$ are given by

$$w_k(\lambda) = \left( \frac{1 - a_{kk}(\lambda)}{\frac{1}{n} \text{tr}(\mathbb{I} - A(\lambda))} \right)^2$$

(14)

$w_k(\lambda)$ have been shown to adjust $V_0(\lambda)$ for periodicity and non-equidistant samples.
Comparison

Comparison of the true mean square error with the GCV function for the example $\lambda$ on $\hat{g}_{n,\lambda}$ and $\hat{g'}_{n,\lambda}$ for $y_i = g(t_i) + \epsilon_i$ where $g(t) = sin(\frac{\pi t}{180})$, $t \in [0, 360]$
Plan

1 Generals

2 What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3 Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4 Discussion
Convergence Result

Theorem

Let $g(\cdot) \in W_2^{(m)}$ and let $\{t_i\}_{i=1}^n$ satisfy $\int_0^{t_i} w(u)du = \frac{i}{n}$, where $w(u)$ is a strictly positive continuous weight function. Then there exists sequences $\{\bar{\lambda}_n\}_{n=1}^\infty$ and $\{\lambda_n^*\}_{n=1}^\infty$ of minimas of $E[V(\lambda)]$ and $E[R(\lambda)]$ such that the expectation inefficiency $I^*$,

$$I^* = \frac{E[V(\bar{\lambda}_n)]}{E[R(\lambda_n^*)]} \to 1, \quad \text{as} \quad n \to \infty \quad (15)$$
Plan

1. Generals

2. What Regularization Parameter $\lambda$ is Optimal?
   - Examples
   - Defining the Optimal $\lambda$

3. Generalized Cross-Validation
   - Cross Validation
   - Generalized Cross Validation (GCV)
   - Convergence Result

4. Discussion
An implicit assumption that the true function $g(t)$ is smooth

GCV was applied effectively to problems such as:
- Finding the right order of splines in regression
- Regularized solution of the Fredholm integral equations of the first kind

It provides an estimate of the regularization parameter from general assumptions on the data.