Model Selection in Regression Problems Based on the Modulus of Continuity

Imhoi Koo and Rhee Man Kil

Neural Information Processing Research Laboratory
KAIST

1. Learning System

- training data \( \{(x_1, y_1), \cdots, (x_N, y_N)\} \).
- \( \left\{ \begin{array}{ll} 
\text{Classification problem} & \text{if } y \in \{-1, 1\} \\
\text{regression} & \text{if } y \text{ is real-value.} 
\end{array} \right. \)
- The learning algorithm means that algorithm takes the training data as input and selects a hypothesis \( f_n \) from the hypothesis space \( \mathcal{H} \).

\[ \text{Algorithm : } \mathcal{D} \rightarrow \mathcal{H} \]
1. Learning System

• Expected risk

\[ R(f_n) = \int_{X \times Y} Q(y, f_n(x)) dP(x, y) \]  

• Empirical risk

\[ R_{\text{emp}}(f_n) = \frac{1}{N} \sum_{i=1}^{N} Q(y_i, f_n(x_i)) \]  

• Loss function \( Q(y, f_n(x)) \)

\[
\begin{cases}
0, & \text{if } y = f_n(x), \\
1, & \text{otherwise.}
\end{cases} \]

Pattern Recognition Regression Density estimation

1. Learning System

• Bias-variance problem

\[ \text{Err}(x) = \sigma^2 + [E\hat{f}(x) - f(x)]^2 + E[\hat{f}(x) - E\hat{f}(x)]^2 \]

Square Error loss = Noise Error + Bias\(^2\) + Variance

• If model complexity \( \leq \) optimal complexity

\[ \implies \text{large bias and small variance.} \]

• If model complexity \( \geq \) optimal complexity

\[ \implies \text{small bias and large variance.} \]
2.1. Model Selection Criteria for Regression Models

AIC, BIC (Akaike information criteria, Bayesian information criteria)

- Let us represent the form of the estimate of expected risks as

\[ \hat{R}(f_n) = R_{emp}(f_n)T(n, l) \quad (3) \]

- The penalty factors of AIC and BIC are the followings:

\[ T_{AIC}(n, l) = \frac{1 + \frac{n}{l}}{1 - \frac{n}{l}} \quad \text{and} \quad (4) \]
\[ T_{BIC}(n, l) = 1 + \frac{\ln l}{2} \left( 1 - \frac{n}{l} \right) \quad (5) \]

- It does not well work with small training data.

SEB (Smallest eigenvalue bound)

- The estimated expected risk can be determined by

\[ E[R(f_n)] = E[R_{emp}(f_n)] \left( 1 + \frac{E[\sum_{i=1}^{n} 1/\lambda_i]}{l} \right) \left( 1 - \frac{n}{l} \right)^{-1} \quad (6) \]

where \( \{\lambda_1, \cdots, \lambda_n\} \) are the eigenvalues of the \( n \times n \) covariant matrix \( C \) in which the \( pq \)th entry \( \frac{1}{l} \sum_{i=1}^{l} \phi_p(x_i)\phi_q(x_i) \).

- With the confidence \( \delta = 0.1 \):

\[ E[R(f_n)] \leq E[R_{emp}(f_n)]T_{SEB}(n, l) \quad (7) \]

\[ T_{SEB}(n, l) = \frac{1 + \frac{n}{lk}}{1 - \frac{n}{l}} \quad \text{and} \quad k = \left( 1 - \sqrt{\frac{n}{l}} \left( 1 + \ln \frac{2l}{n} \right) + \frac{4}{l} \right)^+. \quad (8) \]
2.1. Model Selection Criteria for Regression Models

Comments on Model Selection Criteria:
- AIC and BIC are good model selection criteria when we use linear regression models for large number of samples.
- VC dimension based model section criteria are good for nonlinear models even for small number of samples. However, the accurate estimation of VC dimension of nonlinear regression models is difficult.
- AIC, BIC, and VC dimension bases criteria are obtained from the view point of the complexity of the hypothesis space (not the given hypothesis). As a result, the risk estimate is too loose.
- We need an alternative criteria from the view point of the given hypothesis. Furthermore, it is desirable that the suggested criteria be obtained from the given samples and the trained regression models.

2.2. Model Selection Criteria based on the Modulus of Continuity

- the measure of continuity of a function \( f \in C(X) \) can be described by the following form:

\[
\omega(f, h) = \max_{x, x+t \in X, |t| \leq h} |f(x + t) - f(x)|
\]

• This modulus of continuity of \( f \) has the following properties:
  - \( \omega(f, h) \) is continuous at \( h \geq 0 \) for each \( f \),
  - \( \omega(f, h) \) is positive and increasing for \( h > 0 \), and
  - \( \omega(f, h_1 + h_2) \leq \omega(f, h_1) + \omega(f, h_2) \) for each \( f \).
2.2. Model Selection Criteria based on the Modulus of Continuity

- In general regression problem, the observed value \( y \) is summation of target(regression) function \( f \) and noise \( \epsilon \):

\[
y = f(x) + \epsilon
\]

(10)

where assume the noise \( \epsilon \) has the zero mean and variance \( \sigma^2 \).
- \((x_1, y_1), \ldots, (x_N, y_N)\) i.i.d. randomly drawn samples
- expected risk, empirical risk are defined by the following \( L_1 \) measure:

\[
R(f_n)_{L_1} = \int_{X \times \mathbb{R}} |y - f_n(x)|dP(x, y) \quad \text{and}
\]

\[
R_{emp}(f_n)_{L_1} = \frac{1}{N} \sum_{i=1}^{N} |y_i - f_n(x_i)|.
\]

(11) (12)

**Theorem.** Let us consider that the target function \( f \) is approximated by the estimation function \( f_n(x) = \sum_{i=1}^{n} w_i \phi_i(x) \) for the given samples \((x_i, y_i), i = 1, \ldots, N\). Then, the expected risk in \( L_1 \) sense is bounded by the following inequality with probability at least \( 1 - \delta_1 - \delta_2 \):

\[
R(f_n)_{L_1} \leq R_{emp}(f_n)_{L_1} + \text{Variation} + (\omega(f, h_0) + \omega(f_n, h_0) + C) \sqrt{\frac{1}{2N} \ln \frac{2}{\delta_1}}
\]

where \( \text{Variation} = \frac{1}{N^2} \sum_{i,j=1}^{N} (|y_i - y_j| + |f_n(x_i) - f_n(x_j)|) \)

\[
C = |f_n(x_0) - f_n(x'_0)| + |f(y_0) - f(y'_0)| + 2\sigma \sqrt{\frac{1}{\delta_2}} \quad \text{for}
\]

\( x_0, y_0, x'_0, y'_0 \in \{x_1, \ldots, x_N\} \)
Sketch of Proof

\[ \int_X |f(x) - f_n(x)|dP(x) - \frac{1}{N} \sum_{i=1}^N |f(x_i) - f_n(x_i)| \]
\[ \leq \frac{1}{N} \sum_{i=1}^N \int_X |f(x) - f_n(x) - f(x_i) - f_n(x_i)|dP(x) \]
\[ \leq \frac{1}{N} \sum_{i=1}^N \int_X (|f(x) - f(x_i)| + |f_n(x) - f_n(x_i)|)dP(x). \]

\[ \max_{x \in X} \frac{1}{N} \sum_{i=1}^N |f_n(x) - f_n(x_i)| \leq \omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')| \]

\[ \frac{1}{N} \sum_{i=1}^N \int_X |f(x_i) - f_n(x_i)|dP(x) \leq \frac{1}{N^2} \sum_{i,j=1}^N |f_n(x_i) - f_n(x_j)| \]
\[ + \sqrt{\frac{1}{2N} \ln \frac{2}{\delta_1} (\omega(f_n, h_0) + |f_n(x_0) - f_n(x_0')|)}. \]

and

\[ \frac{1}{N} \sum_{i=1}^N \int_X |f(x) - f(x_i)|dP(x) \leq \frac{1}{N^2} \sum_{i,j=1}^N |f(x_i) - f(x_j)| \]
\[ + \sqrt{\frac{1}{2N} \ln \frac{2}{\delta_1} (\omega(f, h_0) + |f(y_0) - f(y_0')|)}, \]

respectively
Corollary. Let $H_y$ be the $N \times N$ matrix in which the $ij$-th entry is given by $|y_i - y_j|$. Then, the following inequality holds with probability at least $1 - \delta_1 - \delta_2$:

$$R(f_n)_{L_1} \leq 3R_{emp}(f_n)_{L_1} + \frac{2}{N} \max\{\lambda_i\} + \left(\omega(f_n, h_0) + C\right) \sqrt{\frac{1}{2N} \ln \frac{2}{\delta_1}}$$

(13)

where $\lambda_i$ represents the $i$th eigenvalue of the matrix $H_y$.

From this point of view, we can consider the model selection criteria referred to as the modulus of continuity information criteria (MCIC) as follows:

$$MCIC(n) = R_{emp}(f_n)_{L_1} + \frac{\omega(f_n, h_0)}{3} \sqrt{\frac{1}{2N} \ln \frac{2}{\delta_1}}.$$  

(14)

The distinctive characteristics of the suggested model selection criteria $MCIC(n)$ can be described as follows:

- The suggested MCIC can be applied to nonlinear models with arbitrary distribution of samples while the AIC and BIC are good measures of linear models with large number of samples.
- The suggested MCIC depends on the modulus of continuity for a specific hypothesis generated by a learning algorithm and also on the sample distribution while the VC dimension or degree of freedoms based methods consider the structural information only, that is, the complexity of a hypothesis space and the number of samples, and does not depend on a specific hypothesis and the sample distribution.
2.3. Modulus of Continuity for Estimation Functions

**Remark.** If $f \in C^1(X)$, then

$$\omega(f, h) \leq \|f\|_{\infty} \cdot h, \quad \forall h > 0.$$  \hfill (15)

**Example.** Consider the case of estimation function $f_n$ with trigonometric polynomials:

$$\left\{ \frac{1}{2}, \cos px, \sin px \right\}.$$  \hfill (16)

Then, the modulus of continuity for $f_n$ can be determined by

$$\omega(f_n, h) \leq h \sum_{k=0}^{n} |w_k| \cdot \left\lfloor k \cdot \frac{1}{2} \right\rfloor.$$  \hfill (17)

2.4. Model Selection Using $R(f_n)_{L_1}$ versus $R(f_n)_{L_2}$

**Theorem.** Let $E[|g|] = \int |g|d\mu \leq E[|h|] = \int |h|d\mu$. Then,

$$E[|g|^2] = \int |g|^2d\mu \leq E[|h|^2] = \int |h|^2d\mu$$  \hfill (18)

under the following condition:

(a) $\text{Var}(|g|) \leq \text{Var}(|h|)$ \quad or \quad (19)

(b) $\text{Var}(|g|) \geq \text{Var}(|h|)$ \quad and \quad (20)

$$E[|h|] - E[|g|] \geq -\int |g|d\mu + \sqrt{\left(\int |g|d\mu\right)^2 + \text{Var}(|g|) - \text{Var}(|h|)}.$$
2.4. Model Selection Using $R(f_n)_{L_1}$ versus $R(f_n)_{L_2}$

Let us assume that

\[ g = y - f_n \sim N(0, \sigma_n^2) \quad \text{and} \quad \tag{21} \]

\[ h = y - f_m \sim N(0, \sigma_m^2). \quad \tag{22} \]

Then, the expectations for $|g|$ and $|h|$ are given by

\[
E[|g|] = \int_{-\infty}^{+\infty} \frac{|x|}{\sqrt{2\pi}\sigma_n} \exp(-x^2/2\sigma_n^2) \, dx = \frac{2\sqrt{2}\sigma_n}{\sqrt{\pi}} \quad \text{and} \quad \tag{23}
\]

\[
E[|h|] = \int_{-\infty}^{+\infty} \frac{|x|}{\sqrt{2\pi}\sigma_m} \exp(-x^2/2\sigma_m^2) \, dx = \frac{2\sqrt{2}\sigma_m}{\sqrt{\pi}} \quad \tag{24}
\]

respectively. Therefore, if $E[|g|] \leq E[|h|]$, then $\sigma_n \leq \sigma_m$.

\[ E[|g|^2] \leq E[|h|^2] \quad \tag{24} \]

3. Simulation

1. the step function defined by

\[ f(x) = \begin{cases} 
1 & \text{if } x \geq 0 \\
0 & \text{otherwise,} 
\end{cases} \quad \tag{25} \]

2. the sine square function defined by

\[ f(x) = \sin^2(x), \quad \tag{26} \]

3. the sinc function defined by

\[ f(x) = \frac{\sin(\pi x)}{\pi x}, \quad \tag{27} \]

4. and the combined function defined by

\[ f(x) = 2x \exp(-\frac{x^2}{2}) \cos(2\pi x). \quad \tag{28} \]
3. Simulation

- estimation function \( f_n(x) = \sum_{i=1}^{n} w_i \phi_i(x) \)
- trigonometric poly.

\[
\phi_0(x) = \frac{1}{2}, \quad \phi_{2j-1}(x) = \sin jx, \quad \text{and} \quad \phi_{2j}(x) = \cos jx,
\]

- the model selection criteria based on the modulus of continuity for the trigonometric polynomial function was determined by

\[
MCIC(n) = R_{emp}(f_n) + \frac{h_0}{3} \sqrt{\frac{1}{2N} \ln \frac{2}{\delta}} \sum_{i=0}^{n} |\hat{w}_i| \left[ \frac{i}{2} \right]
\]

where \( h_0 \) is set as \( \pi/N \) and \( \delta \) is set as 0.05.
- fixed 200 test samples, 1000 sets of 50 training samples
- \( \sigma = 0.0, 0.025, 0.05, \) and 0.1.
3. Simulation

- estimated optimal model complexity with estimated expected risk $\hat{R}(f_n)$ as AIC, BIC, SEB and MCIC
  \[\hat{n} = \arg\min_n \hat{R}(f_n).\] (30)

- log ratio of two risks
  \[r_R = \log \frac{R(f_{\hat{n}})}{\min_n R(f_n)}\] (31)
  where $R(f_n)$ uses the square error loss $(y - f_n(x))^2$.

- log ratio of two model complexities
  \[r_n = \log \frac{\hat{n}}{\arg\min_n R(f_n)}.\] (32)
4. Conclusions

- We suggest MCIC method
  - characteristic property $\omega(f_n, h)$ of specific hypothesis.
  - density quantity $h$ of training input sample.
  - better performance than previous methods such as AIC, BIC, SEB, etc.
- Applications to machine learning models such as kernel methods, Boosting, etc are easy to implement.

$$f(x) = \frac{\sin(\pi x)}{\pi x}$$

$$f(x) = 2x \exp\left(-\frac{x^2}{2}\right) \cos(2\pi x)$$