Various Bootstrap-Cross-Validation Methods for Model Selection in Density Estimator

Kiyotake KISHI

Abstract
Bayesian evidence is deduced from metrized-sample-space as shown in JSTAT9 (refer to [1]), where we introduce the method of bootstrap-cross-validation, based on Bayes formula, however we were not concerned with whether the bootstrap-cross-validation contained other methods than the Bayes formula. We now exhibit the several methods of bootstrap-cross-validation, based on Bayesian and not on Bayesian.

1. Introduction
Let us denote $y=(y_1, y_2, \ldots, y_N)^T$ be the given data consists with sample size $N$, $\theta$ be the density of unknown population distribution, and $v$ be the model or hyper-parameter for density estimator. Then, we represents $\theta(y,v)$ be the density which estimated from $y$ and $v$, and ideal choice of $v=V_{OPT}$ will enable adjustment to $\theta(y,v_{OPT})=\theta^*$.

Make good use of “bootstrap-resampling” that is, same-size resamples may be drawn repeatedly from the original sample. Let denote $y(k)$ be the bootstrap-resample from $y$ with size $N$, repeatedly $K$ times as $y(k),(k=1,2,\ldots,K)$, these re-samples construct the union of pseudo-data named “pseudo-data-space” symbolized as $Y(y)$, and to use this pseudo-data $y(k)\in Y(y)$, we denote $\theta(y(k),v)$ be the density estimator conditional on $y(k)$ and $v$. Now we can denote “parameter-space” symbolized as $\Theta(v) \equiv \{ \theta(y(k),v) : k=1,2,\ldots,K \}$.

Now, again make good use of “smoothed-bootstrap-resampling” that is, same-size resamples may be drawn repeatedly from the absolutely continuous distribution of the estimator $\theta(y(k),v)\in \Theta(v)$. Use this smoothed-bootstrap-resampling to construct the union of finite number of pseudo-data $y(j,y(k),v)$ which is drawn repeatedly $N$ times from $\theta(y(k),v)$, and also repeatedly same procedure $J$ times iteration as $j=1,2,\ldots,J$ and obtain the union of $J$-th number of pseudo-data named “sample-space” symbolized as $Y(y(k),v)\equiv \{ y(j,y(k),v) : j=1,2,\ldots,J \}$.

Same Group of Empirical Distribution
Now presume $y(j,y(k),v)$ and $\theta(y(k),v)$ to be empirical distributions each other, then absolutely continuous distribution as $\theta(y(k),v)$ and purely discontinuous distribution as pseudo-data $y(j,y(k),v)$ are the same distribution from the empirical viewpoint because of only a resampling relation between the two.

2. Definition of Distance between Two Distributions
An empirical distribution of the estimator $\theta(y(k),v)$ is an absolutely continuous distribution; on the other hand an empirical distribution of the pseudo-data as like $y(t)$ and $y(j,y(k),v)$ are both purely discontinuous distributions.

Therefore, let denote $\mu_1$ be a purely discontinuous distribution, $\mu_2$ be an absolutely continuous distribution, and $\mu_3 \equiv \lambda \mu_1 + (1-\lambda) \mu_2$ be a new drawing distribution, where $\lambda$ be an arbitrary constant within the range of $0<\lambda<1$. Then new deduced measure $\mu_3$ guarantee that $\mu_3$ satisfy an absolutely continuous for both $\mu_1$ and $\mu_2$ measures simultaneously. Consequently, there exists a Radon-Nikodym differential $\mu_1(dx)/\mu_3(dx)$ and $\mu_2(dx)/\mu_3(dx)$, these are called the density of $\mu_1$ and $\mu_2$ with respect to $\mu_3$ on separable measure space $X$, corresponding to the density of absolutely continuous distribution with respect to Lebesgue measure. From this, we can defined the metric between $\mu_1$
and \( \mu_2 \) as follows (refer to [2]):

\[
\mathbf{d}(\mu_1, \mu_2) = \sqrt{\int_X \left( \mu_1(dx) - \mu_2(dx) \right)^2} = \sqrt{\int_X \left( \mu_1(dx) - \sqrt{\mu_2(dx)} \right)^2} 
\]

where, fortunately both the new drawing measure \( \mu_3 \) and the arbitrary constant \( \lambda \) are disappeared.

3. **Bootstrap-Cross-Validation, Not Based on Bayes’ Formula - (1).**

Simple method of bootstrap-cross-validation is as analogous to the MISE: make use of two distributions \( y \) the given-data and the estimator \( \theta(y(k),v) \) which belong to \( \Theta(v) \) as below:

\[
e_1(v) = \frac{1}{K} \sum_{k=1}^{K} d(y, \theta(y(k),v)) \]

where \( y \) is a purely discontinuous and \( \theta(y(k),v) \) is an absolutely continuous distribution, hence we set to \( \mu_1 \equiv y \) and \( \mu_2 \equiv \theta(y(k),v) \) in Eq.(1) for the construction of drawing distance \( d(y, \theta(y(k),v)) \).

4. **Bootstrap-Cross-Validation, Not Based on Bayes’ Formula - (2).**

The basic principle for the choice of optimal model with stable is that: The optimal model adapted to the given data \( y \) is essentially influenced by the character proper to given only one data set with size \( N \). Therefore it might be better to adapt to the expectation of given data as \( E[y] \) and not to \( y \) only. Bootstrap-resampling is effective in diminishing to the character proper to only one given data, and the diminishing technique is composed of two steps average as shown below. First-average to be the average of all the pseudo-data \( y(t) = Y(y_t), (t = 1,\ldots,K) \) except \( t = k \), and the second-average to be the average of all the parameter \( \theta(y(k),v) \equiv \Theta(v), (k = 2,\ldots,K) \) where \( v \) fixed that is to say,

\[
e_2(v) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{K-1} \sum_{t=1(t\neq k)}^{K} d(y(t), \theta(y(k),v)) \]

5. **Bootstrap-Cross-Validation, Based on Bayes’ Formula - (1).**

Let’s denote \( \Theta(v) = \{ \theta(y(k),v) | (k = 1,\ldots,K) \} \) be the parameter-space: the union of probability density estimators, conditional model \( v \). Then Bayes’ formula represents as below.

\[
P[\theta(y(k),v) | \Theta(v)] = \frac{P[\Theta(v) | \theta(y(k),v)]P[\theta(y(k),v)]}{\sum_{\theta(y(k),v)\in\Theta(v)} P[\Theta(v) | \theta(y(k),v)]P[\theta(y(k),v)]} \]

On parameter-space \( \Theta(v) \) all points \( \theta(y(k),v) (k = 1,\ldots,K) \), are equally likely a priori then we have:

\[
P[\theta(y(k),v) | \Theta(v)] = 1/K, \quad P[\Theta(v)] = 1, \quad P[\Theta(v) | \theta(y(k),v)] = 1, \quad P[\theta(y(k),v)] = 1/K \]

Let took total summation respect to \( k \) with both sides of Eq.(4), then we obtain as follows:

\[
P[\theta(y(k),v) | \Theta(v)] = \frac{P[\Theta(v) | \theta(y(k),v)]P[\theta(y(k),v)]}{\sum_{\theta(y(k),v)\in\Theta(v)} P[\Theta(v) | \theta(y(k),v)]P[\theta(y(k),v)]} \]

Here, we attempt to construct a likelihood function as \( L[\theta(y(k),v) | \Theta(v)] \) which corresponding to the probability for \( P[\Theta(v) | \theta(y(k),v)] \), where parameter-space \( \Theta(v) \) plays a role of “Bayesian-given-data”, and \( P[\theta(y(k),v)] \) indicates a prior-probability as equally a priori as Eq.(5): \( P[\theta(y(k),v)] = 1/K = \text{constant} \). Therefore we obtain next formula:

\[
P_{\text{POST}}[\theta(y(k),v) | \Theta(v)] = \frac{L[\theta(y(k),v) | \Theta(v)]}{\sum_{\theta(y(k),v)\in\Theta(v)} L[\theta(y(k),v) | \Theta(v)]} \]

**Metrizable for \( \Theta(v) \) Space**

The central point of parameter-space \( \Theta(v) \) is equal to the expectation of parameter-points \( \theta(y(k),v), (k = 1,\ldots,K) \).

Then, let’s denote \( E[\theta(y(k),v)] = \bar{\theta}(v) = \frac{1}{K} \sum_{k=1}^{K} \theta(y(k),v) \) be the central point of \( \Theta(v) \),
and $d(\theta(y(k),v), \overline{\theta}(v))$ be the distance between $\theta(y(k),v)$ and $E[\theta(y(k),v)] = \overline{\theta}(v)$ on $\Theta(v)$ space.

**Definition of a likelihood function**

A necessary condition for a property of the likelihood function is monotone decreasing as increasing the distance from the parameter-point $\theta(y(k),v)$ to the central point of parameter-space $\Theta(v)$. For example we defined the likelihood function as below:

$$L[\theta(y(k),v) | \Theta(v)] = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} d(\theta(y(k),v), \overline{\theta}(v))^2 \right]$$  \(\text{(8)}\)

Or, if purely discontinuous case, $\theta(y(k),v)$ in Eq.(8) may be replaced by $y(t)$ as $d(y(t), \overline{\theta}(v))$.

Now we consider the extension of the parameter-space which added one pseudo-data-point as $y(t)$ to the parameter-space $\Theta(v)$ denoted by $\overline{\Theta}(v) = \Theta(v) + y(t) = \{ \theta(1,v), \theta(2,v), \theta(3,v), \ldots, \theta(K,v), y(t) \}$.

Then, the probability of occurrence of the $y(t)$ on $\overline{\Theta}(v)$ would be next formula which adding one sample-point to denominator for Eq.(7).

$$P_{\text{POST}}[y(t) | \overline{\Theta}(v)] = \frac{L[y(t) | \Theta(v)]}{\sum_{\theta(y(k),v) \in \Theta(v)} L[\theta(y(k),v) | \Theta(v)] + L[y(t) | \Theta(v)]}$$  \(\text{(9)}\)

We are ready to define the expectation of Bayesian evidence as follows:

$$E[P[y | v]] = E_T[P_{\text{POST}}[y(t) | \overline{\Theta}(v)]] = \frac{1}{K} \sum_{i=1}^{K} P_{\text{POST}}[y(t) | \overline{\Theta}(v)]$$  \(\text{(10)}\)

6. **Bootstrap-Cross-Validation, Based on Bayes’ Formula -(2) (refer to [1]).**

Let’s denote $\Omega = \{Y(y(k),v):(k=1,2,\ldots,K)\} = \{Y(j,y(k),v); (j=1,2,\ldots,J; k=1,2,\ldots,K)\}$ be the universal event space and then Bayes’ formula represents as below.

$$P[y(j,y(k),v) | Y(y(k),v)] = P[Y(y(k),v) | Y(j,y(k),v)]P[y(j,y(k),v)]$$  \(\text{(11)}\)

Where all points $Y(j,y(k),v)$ and $Y(y(k),v)$; $(j=1,2,\ldots,J; k=1,2,\ldots,K)$, are equally likely a priori then we have:

$$P[y(j,y(k),v) | Y(y(k),v)] = \frac{1}{J}, \quad P[Y(y(k),v) | y(j,y(k),v)] = 1, \quad P[y(j,y(k),v)] = 1/K, \quad P[Y(y(k),v)] = 1/J$$  \(\text{(12)}\)

Let took total summation respect to $j$ with both sides we obtain as below:

$$P[Y(y(k),v)] = \sum_{j,y(k),v \in Y(y(k),v)} P[Y(y(k),v) | y(j,y(k),v)]P[y(j,y(k),v)]$$  \(\text{(13)}\)

Substitute (13) for (11), then we obtain as follows:

$$P[y(j,y(k),v) | Y(y(k),v)] = \frac{P[Y(y(k),v) | y(j,y(k),v)]P[y(j,y(k),v)]}{\sum_{y(j,y(k),v) \in Y(y(k),v)} P[Y(y(k),v) | y(j,y(k),v)]P[y(j,y(k),v)]}$$  \(\text{(14)}\)

Here, we attempt to construct a likelihood function as $L[y(j,y(k),v) | Y(y(k),v)]$ which corresponding to the probability for $P[Y(y(k),v) | y(j,y(k),v)]$, and substitute this likelihood for (14), then we obtain posterior probability as follows:

$$P_{\text{POST}}[y(j,y(k),v) | Y(y(k),v)] = \frac{L[y(j,y(k),v) | Y(y(k),v)]P[y(j,y(k),v)]}{\sum_{y(j,y(k),v) \in Y(y(k),v)} L[y(j,y(k),v) | Y(y(k),v)]P[y(j,y(k),v)]}$$  \(\text{(15)}\)

Where sample-space $Y(y(k),v)$ plays a role of “Bayesian-given-data”, and $P[y(j,y(k),v)]$ indicates a prior-probability as equally likely a priori then from Eq.(12), we have $P[y(j,y(k),v)] = 1/(J \times K) = \text{constant}$. Hence we obtain next formula:
\[
P_{\text{post}}[y(j,y(k),v)|Y(y(k),v)] = \frac{L[y(j,y(k),v) | Y(y(k),v)]}{\sum_{y(j,y(k),v) \in Y(y(k),v)} L[y(j,y(k),v) | Y(y(k),v)]} \quad \ldots(16)
\]

**Metrizable for \(Y(y(k),v)\) Space**

Let's new attempt to make use of a distance from sample-point \(y(j,y(k),v)\) to central-point of the sample-space \(Y(y(k),v)\). The central point of sample-space \(Y(y(k),v)\) is equal to the expectation of sample-points \(y(j,y(k),v)\) \((j=1,2,\ldots,J)\) which drawing from \(\theta(y(k),v)\) by \(J\) times. Therefore, \(E[y(j,y(k),v)] = \theta(y(k),v)\) is the central point of sample-space \(Y(y(k),v)\). Then, let's denote \(d(j,y(k),v)\) be the distance between \(y(j,y(k),v)\) and \(\theta(y(k),v)\) on \(Y(y(k),v)\) space. We denote \(\sigma\) be the standard deviation metric that contains 68.3\% sample-points within the sphere \(|d(j,y(k),v)| < \sigma\) on the region of sample-space \(Y(y(k),v)\), and we redefine \(d(j,y(k),v) = d(j,y(k),v) / \sigma\) as a standardized metric. Now, we are ready for construct the likelihood function as \(L(d(j,y(k),v))\) equivalent to the \(L[y(j,y(k),v) | Y(y(k),v)]\) described above, and replaced in Eq.(16) by this, then we can rewrite the posterior-probability given the metric space \(Y(y(k),v)\) after measuring to the sample-space, as below:

\[
P_{\text{post}}[y(j,y(k),v)|Y(y(k),v)] = \frac{L[y(j,y(k),v) | Y(y(k),v)]}{\sum_{y(j,y(k),v) \in Y(y(k),v)} L[y(j,y(k),v) | Y(y(k),v)]} = \frac{L(d(j,y(k),v))}{\sum_{j=1}^{J} L(d(j,y(k),v))} \quad \ldots(17)
\]

Where

\[
L(d(j,y(k),v)) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( d(j,y(k),v) \right)^2 \right] \quad \ldots(18)
\]

**Definition of a Posterior-Probability for Arbitrary Pseudo-Data**

Now, make use of the same \(\sigma\) above, we denote \(d(y(k),v)\) be a standardized metric between arbitrary pseudo-data \(y(t)\) and \(\theta(y(k),v)\) the central point of sample-space \(Y(y(k),v)\). And we consider the extension of the sample-space which added one sample-point as \(y(t)\) to the sample-space \(Y(y(k),v)\). Then, the probability of occurrence of the \(y(t)\) on this extension of sample-space would be next formula which adding one sample-point to denominator for Eq.(17).

\[
P_{\text{post}}[y(t) | Y(y(k),v)] = \frac{L(d(t,y(k),v))}{\sum_{j=1}^{J} L(d(j,y(k),v)) + L(d(t,y(k),v))} \quad \ldots(19)
\]

Where \(Y(y(k),v) \equiv \{ y(1,y(k),v), y(2,y(k),v), \ldots, y(J,y(k),v) \}\)

First-average to be the average of all the pseudo-data \(y(t) \in Y(y(k) ,\{t=1, \ldots, K\})\) except \(t=k\), that is to say,

\[
E[P_{\text{post}}[y(t) | Y(y(k),v)]] = \frac{1}{(K-1)} \sum_{t=1, t\neq k}^{K} \frac{L(d(t,y(k),v))}{\sum_{j=1}^{J} L(d(j,y(k),v)) + L(d(t,y(k),v))} \quad \ldots(20)
\]

Second-average to be the average of all the sample-spaces \(Y(y(k),v) \in \Omega,\{k=1, 2, \ldots, K\}\) where \(v\) fixed that is to say,

\[
E[P_{\text{post}}[y(t) | Y(y(k),v)]] = E_k \left[ E_t \left[ P_{\text{post}}[y(t) | Y(y(k),v)] \right] \right] = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{(K-1)} \sum_{t=1, t\neq k}^{K} \frac{L(d(t,y(k),v))}{\sum_{j=1}^{J} L(d(j,y(k),v)) + L(d(t,y(k),v))} \quad \ldots(21)
\]

Now, we define Eq.(21) to be the expectation of Bayesian evidence.

**References**
