A Theoretical Summary of Cross-Validation Frameworks with Closed-From Expressions

Chieh Lin
Machine Learning Department
Carnegie Mellon University
chiehl1@cs.cmu.edu

1 Introduction

For machine learning and statistical models, cross-validation (CV) method is widely used for model selection and performance estimation. The philosophy of CV is to split the limited amount of data into training set and validation set. The validation set can be the "new data" for estimating the performance of the model. Since CV only assumes that the data are identically distributed and the training set are independent from validation set, CV is universal and can be applied to almost all of frameworks such as regression, density estimation, and classification. The most popular CV schemes are leave-one-out (LOO) cross-validation, and K-fold cross-validation. However, even the most naive implementation of cross-validation with K data splits has a complexity of K times the model training time. This might be impossible to do depending on how expensive the model training is. This project aims to survey the frameworks that have closed-form cross-validation expressions and figure out the theoretical basis of each closed-form expression. To be more specific, we will focus on (LOO) and leave-p-out (LPO) cross-validation schemes where p can be any number between 1 and $n-1$ where n is the size of training set. With this information, the risks of the models built in these frameworks can be easily estimated by the studied closed-form expressions. From literature survey, it is known that closed-form formulas of LOO and LPO risk estimator have been derived for projection estimator, histogram estimator(which is a special case of projection estimator), and kernel estimators for density estimation. For least squares in linear regression, closed-form formulas for LPO estimators are also derived in regression for regressogram, kernel and projection estimators. For regressogram and kernel estimator in regression, additional assumption on data and the kernel functions need to be satisfied, which makes these two closed-forms not universal, so the detail of these two LPOs is not covered in this project.

2 Notations

Framework of density estimation with squared loss
$X_1, ..., X_n \in [0,1]$ are independent and identically distributed random variables drawn from a probability distribution P of density $s \in L^2([0,1])$ with respect to Lebesgue's measure on $[0,1]$. Set $\hat{s}$ to be any estimator belonging to a given class F. The L^2-risk of the estimator $\hat{s}$ is

$$R(\hat{s}) = E_s[||s - \hat{s}||_2^2]. \tag{1}$$

The goal is to find s^* such that

$$s^* = arg\min_{\hat{s}\in F} R(\hat{s}) = arg\min_{\hat{s}\in F} E_s[||s||_2^2 + ||\hat{s}||_2^2 - 2\int s(x)\hat{s}(x)dx] = arg\min_{\hat{s}\in F} L(\hat{s}) \tag{2}$$

Where

$$L(\hat{s}) = E_s[||\hat{s}||_2^2 - 2\int s(x)\hat{s}(x)dx] \tag{3}$$

Note that s^* is unreachable due to its dependency on s.

LPO cross-validation risk estimators for density estimators
Let $X_1, ..., X_n$ be independent identically distributed random variables. For $p \in \{1, ..., n-1\}$, Let e_p be the set of all possible size p-subsets of $\{1, ..., n\}$. For any $e \in e_p, \bar{e} = \{1, ..., n\} \backslash e$ and $X^{\bar{e}} = \{X_i / i \in \bar{e}\}$
With the same notations as before,

$$\hat{L}_p(\hat{s}) = \frac{1}{\binom{n}{p}} \sum_{e \in e_p} [||\hat{s}^{\bar{e}}||_2^2 - \frac{2}{p} \sum_{i \in e} \hat{s}^{\bar{e}}(X_i)] \tag{4}$$

Where $s^{\bar{e}}$ denotes any estimator built from $X^{\bar{e}}$. Generally, this estimator is computation-time prohibitive due to the $\binom{n}{p}$ subsamples.
Set $p = 1$, we have LOO risk estimator,

$$\hat{L}_1(\hat{s}) = \frac{1}{n} \sum_{i=1}^{n} ||\hat{s}^{(i)}||_2^2 - \frac{2}{n} \sum_{i=1}^{n} \hat{s}^{(i)}(X_i) \tag{5}$$

Where $s^{(i)}$ denotes any estimator built from X^{-i}.

Histograms density estimator
Let M be the set of all possible partitions of [0,1] in D intervals. For $m \in M$, $m = (I_k)_{k=1,...,D}$, where the intervals I_k are ordered from left to right, and for any $k \in \{1, ..., D\}$, $\omega_k = |I_k|$ denotes the length of I_k. $n_k = \sum_{i=1}^{n} I(x \in I_k)$ The histogram is defined as following:

$$\hat{s}_\omega(x) = \sum_{k=1}^{D} \frac{n_k}{n\omega_k} I(x \in I_k) \tag{6}$$

Kernel density estimator
For any kernel K, h is called the bandwith and let $K_h(x) = \frac{1}{h} K(\frac{x}{h})$. The corresponding density estimator is for any positive h:

$$\hat{s}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{7}$$

Projection estimator
Let $\{\phi_\lambda\}_{\lambda \in \Lambda_n}$ be a family of functions in $L^2([0,1], \nu)$, where ν denotes the Lebesque measure on [0,1] and Λ_n is a countable set of indices. For any $m \in M_n$, set $\Lambda(m) \subset \Lambda_n$ such that $\{\phi_\lambda\}_{\lambda \in \Lambda(m)}$ is an orthonormal family of functions. Let S_m denote the linear space of dimension D_m spanned by $\{\phi_\lambda\}_{\lambda \in \Lambda(m)}$. We call $\hat{s}$ a projection estimator of s such that

$$\hat{s}_m = \sum_{\lambda \in \Lambda(m)} \hat{\beta}_\lambda \phi_\lambda \text{ with } \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \theta_\lambda(Z_i) \tag{8}$$

where $Z_1, ..., Z_n$ denote some observations (density or regression) and θ_λ is a function independent from the observations for any λ.
In the density estimation framework, $\theta_\lambda = \phi_\lambda$ for every λ and

$$\hat{s}_m = \sum_{\lambda \in \Lambda(m)} \hat{\beta}_\lambda \phi_\lambda \text{ with } \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \phi_\lambda(Z_i) \tag{9}$$

A typical example of projection estimator is histogram for which $\phi_\lambda = \frac{I(\lambda)}{\sqrt{|I(\lambda)|}}$ where $\{I(\lambda)\}_{\lambda \in \Lambda_m}$ denotes a partition of [0,1] and $|I(\lambda)|$ represents the length of the interval $I(\lambda)$.
In the regression setting (for uniformly distributed $X_i = \frac{i}{n}$), $\theta_\lambda[(x, y)] = y\phi_\lambda((x)$ and

$$\hat{s}_m = \sum_{\lambda \in \Lambda(m)} \hat{\beta}_\lambda \phi_\lambda \text{ with } \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_\lambda(Z_i) \tag{10}$$

It is easy to check that every projection estimators $\hat{s}_m$ may be written as

$$\forall 1 \leq i \leq n, \hat{s}_m(Z_i) = \frac{1}{n}\sum_{j=1}^{n} H_m(Z_j, Z_i) \tag{11}$$

where $H_m(.,.)$ is a function which may be expressed in terms of the basis vectors. This expression is also strongly connected to kernel density estimators as well. Indeed, provided K denotes a kernel and $h > 0$ a smoothing parameter, we may define

$$\forall 1 \leq i \leq n, \hat{s}_h(X_i) = \frac{1}{n}\sum_{j=1}^{n} K_h(X_j - X_i) \tag{12}$$

Framework of regression with squared loss
We observe a sample data $D = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, where $X_i = (X_i(1), ..., X_i(d)) \in R^d$ and $Y_i \in R$. For simplicity, let $X_i(1) = 1$. The conditional prediction risk of an estimator $\hat{m}$ for a new data pair (X,Y) is

$$r(\hat{m}) = E[(Y - \hat{m}(X))^2|D] = \int (y - \hat{m}(x))^2 dP(x, y) \tag{13}$$

The prediction risk of $\hat{m}$ is

$$R(\hat{m}) = E[(Y - \hat{m}(X))^2] = E[r(\hat{m})] \tag{14}$$

The true regression function is

$$m(x) = E[Y|X = x] \tag{15}$$

linear smoothers
An estimator $\hat{m}$ of m is a linear smoother if, for each x, there is a vector $l(x) = (l_1(x), ..., l_n(x))^T$ such that

$$\hat{m}(x) = \sum_{i=1}^{n} l_i(x) Y_i = l(x)^T Y \tag{16}$$

Note that in least square linear regression, $\hat{m}(x) = x^T\hat{\beta}$ where $\hat{\beta} = x^T(X^TX)^{-1}X^TY = l(x)^TY$ is a special case of linear smoother.

LOO risk of linear smoother
The LOO risk of regression with squared loss is defined as follows:

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{(-i)}(X_i))^2 \tag{17}$$

Where $\hat{m}_{(-i)}$ is the estimator obtained by deleting the data pair (X_i, Y_i), that is

$$\hat{m}_{(-i)} = \sum_{j=1}^{n} Y_j l_{j,(-i)(x)} \tag{18}$$

and

$$l_{j,(-i)(x)} = \frac{l_{j(x)}}{\sum_{k\neq i} l_k(x)} I(j \neq i) \tag{19}$$

3 Key Results and proofs

Lemma1 for proving Lemma2:
For any $i \neq j \neq k \in \{1, ..., n\}$,

$$\sum_{e\in e_p} 1_{(j\in\bar{e})} = \binom{n-1}{p} \tag{20}$$

$$\sum_{e\in e_p} 1_{(i\in e)} = \binom{n-1}{p-1} \tag{21}$$

$$\sum_{e\in e_p} 1_{(j\in \bar{e})}1_{(k\in \bar{e})} = \binom{n-2}{p} \tag{22}$$

$$\sum_{e\in e_p} 1_{(i\in e)}1_{(j\in \bar{e})}1_{(k\in \bar{e})} = \binom{n-3}{p-1} \tag{23}$$

$$\sum_{e\in e_p} 1_{(i\in e)}1_{(j\in \bar{e})} = \binom{n-2}{p-1} \tag{24}$$

proof sketch
$\sum_{e\in e_p} 1_{(j\in \bar{e})}$ can be interpreted as the number of subsets of $\{1, ..., n\}$ of size p (denoted by e) which do not contain j, since $j \in \bar{e}$. Thus it is the number of possible choices of p non-ordered and different elements among $n-1$. Other equailties follow from a similar argument.

Lemma2 for projection estimator theorem of LPO:
Let $\hat{s}_m^{\bar{e}}$ denote a generic projection estimator based on model S_m and computed from training set $X^{\bar{e}}$, Then

$$\sum_{e\in e_p} ||\hat{s}_m^{\bar{e}}||_2^2 = \frac{1}{(n-p)^2}[\binom{n-1}{p}\sum_{k=1}^{n}||H_m(X_k,.)||_2^2 + \binom{n-2}{p}\sum_{k\neq l} < H_m(X_k,.), H_m(X_l,.) >_2] \tag{25}$$

$$\sum_{e\in e_p}\sum_{i\in e} \hat{s}_m^{\bar{e}}(X_i) = \frac{1}{n-p}\binom{n-2}{p-1}\sum_{i\neq j} H_m(X_i, X_j) \tag{26}$$

$$\sum_{e\in e_p}\sum_{i\in e}[\hat{s}_m^{\bar{e}}(X_i)]^2 = \frac{1}{n-p}[\binom{n-2}{p-1}\sum_{i\neq j}[H_m(X_i,X_j)]^2 + \binom{n-3}{p-1}\sum_{i\neq j\neq l} H_m(X_k,X_i)H_m(X_l,X_i)] \tag{27}$$

$$\sum_{e\in e_p}\sum_{i\in e} Y_i\hat{s}_m^{\bar{e}}(X_i) = \frac{1}{n-p}\binom{n-2}{p-1}\sum_{i\neq j} Y_i H_m(X_i, X_j) \tag{28}$$

proof sketch
For each $e \in e_p$, we have $\forall t \in [0, 1]$,

$$\hat{s}_m^{\bar{e}}(t) = \frac{1}{n-p}\sum_{j\in\bar{e}} H_m(X_j, t) = \frac{1}{n-p}\sum_{j=1}^{n} H_m(X_j,t)1_{(j\in\bar{e})} \tag{29}$$

$$\begin{aligned}\sum_{i\in e}\hat{s}_m^{\bar{e}}(X_i) &= \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j\in\bar{e}} H_m(X_j,X_i)1_{(i\in e)} \\ &= \frac{1}{n-p}\sum_{i\neq j} H_m(X_j,X_i)1_{(i\in e)}1_{(j\in\bar{e})}\end{aligned} \tag{30}$$

Then, this lemma follows from lemma 1.

Proposition 1 for density estimator LPO closed-form risk:
For any density $t : [0, 1] \to R_+$, for any $m \in M_n$, let $\hat{s}_m$ denote a generic projection estimator based on model S_m spanned by the orthonormal basis $\{\phi_\lambda\}_{\lambda\in\Lambda_{(m)}}$. Then the L^2-loss of projection density LPO risk estimator is

$$\hat{L}_p(\hat{s}_m) = \frac{1}{n(n-p)}\sum_{\lambda\in\Lambda(m)}[\sum_j \phi_\lambda^2(X_j) - \frac{n-p+1}{n-1}\sum_{j\neq k}\phi_\lambda(X_j)\phi_\lambda(X_k)] \tag{31}$$

Besides, let $\hat{s}_m$ denote the kernel density estimator based on a symmetric kernel K, with smoothing parameter $m \geq 0$. Then, the L^2-loss of kernel density LPO risk estimator is

$$\hat{L}_p(\hat{s}_m) = \frac{1}{n-p}||K_m||_2^2 + \frac{n-p-1}{n(n-1)(n-p)} \sum_{k \neq l} K_m^*(X_k - X_l) - \frac{2}{n(n-1)} \sum_{k \neq l} K_m(X_k - X_l) \tag{32}$$

Where $K_m^* := (K * K)_m$ and "*" denotes the convolution product.

proof sketch
The LPO estimator with squared loss is

$$\hat{L}_p(\hat{s}_m) = \binom{n}{p}^{-1} \sum_{e \in e_p} ||\hat{s}_m^{\bar{e}}||_2^2 - \frac{2}{p}\binom{n}{p}^{-1} \sum_{e \in e_p} \sum_{i \in e} \hat{s}_m^{\bar{e}}(X_i) \tag{33}$$

Also, for projection density estimator,

$$H_m(X_i, X_j) = \sum_{\lambda \in \Lambda(m)} \phi_\lambda(X_j)\phi_\lambda(X_i) \tag{34}$$

Besides, for kernel density estimator,

$$H_m(X_i, X_j) = K_m(X_j - X_i) := K[(X_j - X_i)/m] \tag{35}$$

Then, the expected result follows from equation 25 and 26 in Lemma2.

Corollary: LPO risk estimators for histograms density estimators
Let $\hat{s}_m$ denotes the histogram estimator built from the partition $I(m) = (I_1, ..., I_{D_m})$ of $[0, 1]$ in D_m intervals of length $\omega_k = I_k$. For any $p \in \{1, ..., n-1\}$

$$\hat{L}_p(\hat{s}_m) = \frac{2n-p}{(n-1)(n-p)} \sum_{k=1}^{D_m} \frac{n_k}{n\omega_k} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_{k=1}^{D_m} \frac{1}{\omega_k}(\frac{n_k}{n})^2 \tag{36}$$

proof sketch
This result follows from equation 31 by plug in $\phi_\lambda = \frac{1_{I_\lambda}}{\sqrt{\omega_\lambda}}$

Corollary: LOO risk estimators for histograms density estimators
In the case of a regular D-piece histogram($\omega = \frac{1}{D}$)

$$\hat{L}_1(\omega) = \frac{2n-1}{(n-1)^2\omega} - \frac{1}{(n-1)^2\omega} \sum_{k=1}^{D} n_k^2 \tag{37}$$

proof sketch
This result follows from equation 36 by plug in $p = 1$

Proposition 2 for regression estimator LPO closed-form risk:
For any observations $Z = (X, Y)$ and any function $t : [0, 1] \to R$, for any $m \in M_n$, let $\hat{s}_m$ denote a generic projection estimator based on model S_m spanned by the orthonormal basis $\{\phi_\lambda\}_{\lambda \in \Lambda_{(m)}}$. Then the L^2-loss of projection regression LPO risk estimator is

$$\begin{aligned}\hat{L}_p(\hat{s}_m) =& \frac{1}{n(n-1)}[\frac{1}{n-p} \sum_{i \neq j} H_m^2(X_j, X_i) + \frac{n-p-1}{(n-p)(n-2)} \sum_{i \neq j \neq k} H_m(X_j, X_i)H_m(X_k, X_i) \\ & - 2\sum_{i \neq j} Y_i H_m(X_j, X_i)] + \frac{1}{n}\sum_{i=1}^{n} Y_i^2\end{aligned} \tag{38}$$

Where $H_m(X_j, X_i) = \sum_{\lambda \in \Lambda(m)} Y_i \phi_\lambda(X_j)\phi_\lambda(X_i)$

proof sketch
The LPO risk estimator can be expressed as the sum of three terms:

$$\hat{L}_p(\hat{s}_m) =[\frac{1}{p}\binom{n}{p}^{-1}\sum_{e\in e_p}\sum_{i\in e}Y_i^2] + [\frac{1}{p}\binom{n}{p}^{-1}\sum_{e\in e_p}\sum_{i\in e}[\hat{s}_m(X_i)^2]] - 2[\frac{1}{p}\binom{n}{p}^{-1}\sum_{e\in e_p}\sum_{i\in e}Y_i\hat{s}_m(X_i)] \tag{39}$$

With the first term dealt with lemma1 and the remaining terms dealt by equation 27 and 28 in lemma2, this theorem is proved.

Proposition 3, LOO risk estimator of linear smoother
Let $\hat{m}$ be a linear smoother. Then the LOO risk $\hat{R}(h)$ can be written as

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - \hat{m}_h(X_i)}{1 - L_{ii}})^2 \tag{40}$$

where $L_{ii} = l_i(X_i)$

proof

$$\begin{aligned}
\hat{m}(X_i) &= \sum_{j=1}^{n} L_{ij}Y_j \\
\text{Set } Z_j &= \begin{cases} Y_j, & \text{if } j \neq i \\ \hat{m}_{(-i)}(X_i) & \text{if } j = i \end{cases} \\
\hat{m}_{(-i)}(X_i) &= arg\min_m \sum_{j\neq i}(Y_j - \hat{m}_{(-i)}(X_j))^2 = arg\min_m \sum_j (Z_i - \hat{m}_{(-i)}(X_j))^2 \\
&=> \hat{m}_{(-i)}(X_i) = \sum_{j=1}^{n} L_{ij}Z_j \\
(\hat{m}(X_i) - \hat{m}_{(-i)}(X_i)) &= \sum_{j=1}^{n} L_{ij}(Y_j - Z_j) = L_{ii}(Y_i - m_{(-i)}(X_i)) \\
&=> \hat{m}_{(-i)}(X_i) = \frac{\hat{m}(X_i) - L_{ii}Y_i}{1 - L_{ii}} \\
\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_{(-i)}(X_i))^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \frac{\hat{m}(X_i) - L_{ii}Y_i}{1 - L_{ii}})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i - \hat{m}(X_i)}{1 - L_{ii}})^2
\end{aligned} \tag{41}$$

4 Conclusion

We have investigated and verified that closed-form LPO and LOO are available in density estimation and regression frameworks with some well-known estimators. This project provides a summary of the exact closed-form expressions under specific estimators and frameworks.

Reference

Arlot, Sylvain, and Alain Celisse. "A survey of cross-validation procedures for model selection." Statistics surveys 4 (2010): 40-79.

Rudemo, Mats. ”Empirical choice of histograms and kernel density estimators.” Scandinavian Journal of Statistics (1982): 65-78.

Bowman, Adrian W. ”An alternative method of cross-validation for the smoothing of density estimates.” Biometrika 71.2 (1984): 353-360.

Zhang, Ping. ”Model selection via multifold cross validation.” The Annals of Statistics (1993): 299-313.

Celisse, Alain. ”Model selection in density estimation via cross-validation.” (2009).

Celisse, Alain. Model selection via cross-validation in density estimation, regression, and change-points detection. Diss. Universit Paris Sud-Paris XI, 2008.

Celisse, Alain, and Stphane Robin. ”Nonparametric density estimation by exact leave-p-out cross-validation.” Computational Statistics & Data Analysis 52.5 (2008): 2350-2368.

Craven, Peter, and Grace Wahba. ”Smoothing noisy data with spline functions.” Numerische Mathematik 31.4 (1978): 377-403.

Wahba, Grace, and Svante Wold. ”Periodic splines for spectral density estimation: The use of cross validation for determining the degree of smoothing.” Communications in Statistics-Theory and methods 4.2 (1975): 125-141.

Wahba, Grace. ”Practical approximate solutions to linear operator equations when the data are noisy.” SIAM Journal on Numerical Analysis 14.4 (1977): 651-667.