# The variance of the cross-validation error

Felix Held

MSA220/MVE441 Statistical Learning for Big Data
Started: 2020-04-13, Last update: 2021-03-26

## 1 Introduction

In the course we have dealt with the framework of statistical learning. Given a loss function $L(y, f(\boldsymbol{x}))$ and some data $\mathcal{T} = \{(y_l, \boldsymbol{x}_l) : l = 1, \dots, n\}$ we are attempting to estimate some function $f(\boldsymbol{x})$. The goals of learning the function $f(\boldsymbol{x})$ are (1) to describe the training data, but also, *more importantly*, (2) to predict future observations. Formally, we would like to minimize the expected prediction error (EPE)

$$J(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[L(y, f(\boldsymbol{x}))\right] = \mathbb{E}_{p(\boldsymbol{x})}\left[\mathbb{E}_{p(y|\boldsymbol{x})}\left[L(y, f(\boldsymbol{x})\right]\right]. \tag{1.1}$$

However, as was mentioned in the lectures, this is usually not possible directly. Instead, we learn the optimal solution (i.e. the optimal $f$) to an approximation of $J(f)$.

One such approximation is the training, or emperical EPE,

$$R^{\text{tr}}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)). \tag{1.2}$$

The optimal function $\widehat{f}$ is then found as

$$\widehat{f} = \arg\min_f R^{\text{tr}}(f). \tag{1.3}$$

Often, the minimization is restricted to a subclass of functions. As an example, in linear regression $f$ is restricted to all linear functions.

The best solution is often connected to the selection of one or multiple hyper-parameters (for example the value of $k$ in kNN) or we might want to compare the best model by comparing different classes of models such as LDA vs kNN. $c$-fold cross-validation is often the method of choice for selection of hyperparameters or choice between models.

## 2 The variance of the cross-validation error

Given the training data $\mathcal{T} = \{(y_l, \boldsymbol{x}_l) : l = 1, \dots, n\}$, we **randomly split the data** into $c$ folds $\mathcal{F}_1, \dots, \mathcal{F}_c$ of about equal size. The cross-validation error can then be calculated as

$$R_{\text{CV}}(\lambda) = \frac{1}{c} \sum_{j=1}^{c} \frac{1}{|\mathcal{F}_j|} \sum_{l \in \mathcal{F}_j} L(y_l, \widehat{f}(\boldsymbol{x}_l; \mathcal{F}_{-j}, \lambda)) \tag{2.1}$$

where $|\mathcal{F}_j|$ is the number of elements in the set $\mathcal{F}_j$, $\lambda$ denotes a single or multiple hyperparameters or is an indicator for the chosen model and $\mathcal{F}_{-j} = \{(y, x) : (y, x) \in \mathcal{T}, (y, x) \neq \mathcal{F}_j\}$ is the training data when fold $j$ is used for testing. $\widehat{f}(\cdot; \mathcal{F}_{-j}, \lambda)$ is the optimal function that was chosen after training on all folds except fold $j$ subject to the hyperparameters $\lambda$.

Note that **the cross validation error itself is random**, since the folds have been created randomly and the training data can be considered to be drawn randomly from the data distribution. Once training data has been collected and the folds have been created, the cross validation error acts as a point estimate for the EPE $J(f)$ in Eq. (1.1).

However, since the CV error is random and therefore varies, it is often of interest to estimate the variance of the cross validation error as well. In other words, what we would like to estimate is how much the point estimate of the EPE would vary if we were to collect new training sets.

In the following, denote

$$R_j(\lambda) = \frac{1}{|\mathcal{F}_j|} \sum_{l \in \mathcal{F}_j} L(y_l, \widehat{f}(\boldsymbol{x}_l; \mathcal{F}_{-j}, \lambda)). \tag{2.2}$$

This way, $R_{\text{CV}}(\lambda) = \frac{1}{c} \sum_{j=1}^{c} R_j(\lambda)$ is the average of $c$ random variables. Note that the $R_j$ are random variables with respect to the randomness in the training data and the allocation of the folds. ***If*** *the* $R_j(\lambda)$ ***were*** *independent and identically distributed (iid),* then

$$\text{Var}(R_{\text{CV}}(\lambda)) = \frac{1}{c^2} \sum_{j=1}^{c} \text{Var}(R_j(\lambda)) = \frac{\text{Var}(R_1(\lambda))}{c} \tag{2.3}$$

using the formulas $\text{Var}(aX) = a^2 \text{Var}(X)$ for a random variable $X$ and a constant $a$, as well as $\text{Var}(X_1 + X_2 + \cdots + X_N) = \sum_{l=1}^{N} \text{Var}(X_l)$ for independent random variables $X_1, \dots, X_N$. In addition, we used the fact that the $R_j(\lambda)$ are assumed to be identically distributed. If we use the standard unbiased estimator for the sample variance

$$\text{Var}(R_1) \approx \frac{1}{c-1} \sum_{j=1}^{c} (R_j - R_{\text{CV}})^2. \tag{2.4}$$

we then get

$$\mathrm{Var}(R_{\mathrm{CV}}(\lambda)) \approx \frac{1}{c}\left(\frac{1}{c-1}\sum_{j=1}^{c}(R_j(\lambda) - R_{\mathrm{CV}}(\lambda))^2\right), \tag{2.5}$$

which is the standard variance estimator for means of iid random variables. To get the sample standard error[1], we can use

$$\mathrm{se}(R_{\mathrm{CV}}(\lambda)) \approx \frac{1}{\sqrt{c}}\sqrt{\frac{1}{c-1}\sum_{j=1}^{c}(R_j(\lambda) - R_{\mathrm{CV}}(\lambda))^2}. \tag{2.6}$$

*However,* since the training data sets in cross validation overlap, the assumption of independance is not always valid and the independence estimator will over- or underestimate variance (most likely underestimate).

Assuming the $R_j(\lambda)$ are identically distributed but *not* independent leads to the following decomposition of the variance

$$\begin{aligned}\mathrm{Var}(R_{\mathrm{CV}}(\lambda)) &= \frac{1}{c^2}\sum_{j=1}^{c}\mathrm{Var}(R_j(\lambda)) + \frac{1}{c^2}\sum_{j\neq j'}\mathrm{Cov}(R_j(\lambda), R_{j'}(\lambda)) \\ &= \frac{\mathrm{Var}(R_1(\lambda))}{c} + \frac{1}{c^2}\sum_{j\neq j'}\mathrm{Cov}(R_j(\lambda), R_{j'}(\lambda)),\end{aligned} \tag{2.7}$$

where the second equality holds due to the $R_j$ being identically distributed. The amount of over- or underestimation is determined by the amount of correlation between the $R_j$ and whether it is positive or negative on average.

Usually it is not possible to reasonably estimate the correlation in Eq. 2.7. Some more explicit but complicated attempts are Nadeau and Bengio (2003), Bengio and Grandvalet (2004), and Markatou et al. (2005). Empirically however, the covariance is typically considered small as long as $c$, the number of folds, is small. When $c$ approaches $n$, correlations (and thereby covariances) between folds grow larger and the approximation of the standard error in Eq. (2.6) becomes worse.

Two other ways how the variance of $R_{\mathrm{CV}}(\lambda)$ can be computed in the independence setting are based on element-wise error (see also Jiang and Wang (2017)). Define

$$e_l = L(y_l, \widehat{f}(\boldsymbol{x}_l; \mathcal{F}_{-j_l}, \lambda)) \tag{2.8}$$

where $j_l$ is the index of the fold which the pair $(y_l, \boldsymbol{x}_l)$ belongs to. With this definition and the assumption that all folds are equally large (i.e. the number of samples $n$ is divisible by the number of folds $c$), the cross-validation error can be written as

$$R_{\mathrm{CV}}(\lambda) = \frac{1}{n}\sum_{l=1}^{n} e_l. \tag{2.9}$$

[1] Recall that the standard deviation of an estimator around its expected mean is called the standard error. The cross-validation error is an estimator of the EPE in Eq. (1.1) and its estimated mean is the value $R_{\mathrm{CV}}(\lambda)$.

Assuming that the $e_l$ are *iid* it then follows that

$$\mathrm{Var}(R_{\mathrm{CV}}(\lambda)) = \frac{\mathrm{Var}(e_1)}{n} \approx \frac{1}{n}\left(\frac{1}{n-1}\sum_{l=1}^{n}(e_l - R_{\mathrm{CV}}(\lambda))^2\right). \tag{2.10}$$

The estimate in Eq. (2.10) is just another way of writing Eq. (2.5) if all folds are exactly the same size.

In the following, consider the special case where the loss function is the 0-1 loss. In this case, the cross validation estimate in Eq. (2.9) is an estimate of the misclassification rate or more formal, assuming that the $e_l$ are identically distributed, $R_{\mathrm{CV}}(\lambda) \approx P(e_1 = 1)$. Since $e_l \in \{0, 1\}$, the element-wise losses can be considered Bernoulli distributed random variables with parameter $\widehat{p} = R_{\mathrm{CV}}(\lambda)$. It then holds that

$$\mathrm{Var}(R_{\mathrm{CV}}(\lambda)) = \frac{\mathrm{Var}(e_1)}{n} \approx \frac{\widehat{p}(1-\widehat{p})}{n}. \tag{2.11}$$

# References


Bengio, Y and Grandvalet, Y (2004) "No Unbiased Estimator of the Variance of K-Fold Cross-Validation". In: *Advances in Neural Information Processing Systems 16*. Ed. by S Thrun, LK Saul, and B Schölkopf. MIT Press. 513–520 (p. 3).

Jiang, G and Wang, W (2017) Error estimation based on variance analysis of *k*-fold cross validation. *Pattern Recognition* 69:94–106. DOI 10.1016/j.patcog.2017.03.025 (p. 3).

Markatou, M, Tian, H, Biswas, S, and Hripcsak, G (2005) Analysis of Variance of Cross-Validation Estimators of the Generalization Error. *Journal of Machine Learning Research* 6:1127–1168 (p. 3).

Nadeau, C and Bengio, Y (2003) Inference for the Generalization Error. *Machine Learning* 52:239–281. DOI 10.1023/A:1024068626366 (p. 3).