Cross-Validation

Statistics · Module 10 — Model Selection Learning

Node 10.3

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.3.

Cross-validation is the nonparametric, model-agnostic approach to estimating generalization error. Where information criteria (Node 10.2) rely on parametric assumptions and closed-form penalties, cross-validation directly simulates the train–test split on the observed data. It applies to any estimator and any loss function, making it the universal model selection tool for complex or nonparametric pipelines. It feeds into high-dimensional methods (Node 10.5) as the standard tuning procedure.

Formal Definition

Let $\hat{f}_n(\,\cdot\,;\mathcal{D})$ denote an estimator trained on dataset $\mathcal{D}$. The **generalization error** under loss ℓ is

$$R_n = \mathbb{E}_{(X,Y)\sim P_0}\big[\ell(Y, \hat{f}_n(X;\mathcal{D}_n))\big],$$

where the outer expectation is over a new test point independent of $\mathcal{D}_n$.

K-Fold Cross-Validation. Partition $\mathcal{D}_n$ into K disjoint folds $\mathcal{D}_1, \ldots, \mathcal{D}_K$ of approximately equal size n/K. For each $k = 1, \ldots, K$, train on $\mathcal{D}_{-k} = \mathcal{D}_n \setminus \mathcal{D}_k$ to obtain $\hat{f}^{(-k)}$ and evaluate on fold $\mathcal{D}_k$. The CV estimator is

$$\widehat{R}^{K\text{-CV}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{(X_i,Y_i)\in\mathcal{D}_k} \ell\big(Y_i,\, \hat{f}^{(-k)}(X_i)\big).$$

Leave-One-Out Cross-Validation (LOOCV). Set $K = n$:

$$\widehat{R}^{\text{LOO}} = \frac{1}{n}\sum_{i=1}^{n} \ell\big(Y_i,\, \hat{f}^{(-i)}(X_i)\big).$$

Intuition

Cross-validation answers the question: how well does this estimator perform on data it has not seen? Rather than using a single held-out test set (which wastes data), it rotates through all possible partitions so every observation serves as both training and validation data. The average held-out loss is an estimate of the generalization error.

The key tradeoff in choosing K is between bias (training on fewer points overestimates the error) and variance (correlated fold estimates inflate the variance of the CV estimator).

Structural Role

Dependencies: Builds on the bias–variance tradeoff (Node 10.1) and estimation theory (Node 3.1). **Enables:** Tuning parameter selection for regularized estimators (Node 10.4), evaluation of high-dimensional methods (Node 10.5), and weight selection in stacking (Node 10.6).

Cross-validation is the empirical complement to the analytic criteria of Node 10.2. Where AIC and BIC require parametric structure, CV applies to any estimator—random forests, neural networks, kernel methods—making it indispensable for modern statistical practice.

Mathematical Development

Bias of K-Fold CV

The k-th fold trains on $n(1-1/K)$ observations. Since the generalization error typically decreases with training set size, the CV estimator targets the risk of an estimator trained on $n(1-1/K)$ points, not n points. Therefore:

$$\mathbb{E}[\widehat{R}^{K\text{-CV}}] = R_{n(1-1/K)} \geq R_n.$$

The CV estimator has **upward bias** (pessimistic) for R_n. LOOCV ($K = n$) trains on $n-1$ points, so its bias is minimal: $R_{n-1} - R_n = O(1/n^2)$ under smoothness conditions.

Variance of the CV Estimator

The n leave-one-out residuals $e_i = \ell(Y_i, \hat{f}^{(-i)}(X_i))$ are not independent: each pair of models $\hat{f}^{(-i)}$ and $\hat{f}^{(-j)}$ shares $n-2$ training points. The variance of LOOCV is

$$\mathrm{Var}(\widehat{R}^{\mathrm{LOO}}) = \frac{1}{n^2}\left[n\,\mathrm{Var}(e_i) + n(n-1)\,\mathrm{Cov}(e_i, e_j)\right].$$

When the estimator is unstable (high variance), the covariance term is large and LOOCV has high variance. Smaller K reduces this covariance at the cost of increased bias.

Empirical recommendation: $K = 5$ or $K = 10$ provides a reasonable bias–variance balance. Repeated K-fold CV (averaging over multiple random partitions) reduces the variance from the particular fold assignment.

Stone's Theorem: LOOCV and AIC

Theorem (Stone, 1977). For linear models under squared-error loss, LOOCV and AIC are asymptotically equivalent:

$$\widehat{R}^{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2 \approx \frac{\mathrm{RSS}}{n} + \frac{2\hat{\sigma}^2 k}{n} + o(1/n),$$

where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix and $k = \mathrm{Tr}(H)$.

Proof sketch. Expand $(1 - H_{ii})^{-2} \approx 1 + 2H_{ii} + O(H_{ii}^2)$. Then

$$\widehat{R}^{\mathrm{LOO}} \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2(1 + 2H_{ii}) = \frac{\mathrm{RSS}}{n} + \frac{2}{n}\sum_i e_i^2 H_{ii}.$$

Under the Gaussian linear model, $\mathbb{E}[e_i^2 H_{ii}] \approx \sigma^2 H_{ii}$, so $\sum_i e_i^2 H_{ii} \approx \sigma^2 \mathrm{Tr}(H) = \sigma^2 k$. This gives

$$\widehat{R}^{\mathrm{LOO}} \approx \frac{\mathrm{RSS}}{n} + \frac{2\sigma^2 k}{n} = \frac{1}{n}\left(\mathrm{RSS} + 2\sigma^2 k\right),$$

which is AIC (up to a constant and the factor $1/n$) for the Gaussian linear model.

LOOCV Shortcut for Linear Models

For any linear smoother $\hat{Y} = HY$, the LOOCV loss under squared error is computed without refitting:

$$\widehat{R}^{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2.$$

Derivation. By the Sherman–Morrison formula, removing observation i from OLS gives

$$\hat{Y}_i^{(-i)} = \hat{Y}_i - \frac{H_{ii}(Y_i - \hat{Y}_i)}{1 - H_{ii}},$$

so the leave-one-out residual is

$$Y_i - \hat{Y}_i^{(-i)} = \frac{Y_i - \hat{Y}_i}{1 - H_{ii}}.$$

This reduces the cost of LOOCV from n model fits to a single fit plus computation of the hat matrix diagonal.

Generalized Cross-Validation (GCV)

For large n, the individual leverages H_{ii} can be replaced by their average $\mathrm{Tr}(H)/n = k/n$:

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - k/n}\right)^2 = \frac{\mathrm{RSS}/n}{(1 - k/n)^2}.$$

GCV is invariant under orthogonal transformations of the response and is computationally cheaper than LOOCV when H_{ii} values are expensive to compute.

Worked Examples

Example 1: LOOCV for Polynomial Selection

Data: $n = 15$ observations from $Y = X^2 + \varepsilon$, $X \sim \mathrm{Uniform}(-1, 1)$, $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. Fit polynomials of degree 1, 2, 4.

For the degree-p polynomial, $k = p + 1$, and $H = X_p(X_p^\top X_p)^{-1}X_p^\top$ where X_p is the $n \times (p+1)$ Vandermonde matrix.

Suppose $\mathrm{RSS}_1 = 4.20$, $\mathrm{RSS}_2 = 3.15$, $\mathrm{RSS}_4 = 2.98$, and the average leverages are $\bar{H}_1 = 2/15$, $\bar{H}_2 = 3/15$, $\bar{H}_4 = 5/15$.

Using GCV: $\mathrm{GCV}_1 = (4.20/15)/(1 - 2/15)^2 = 0.280/0.751 = 0.373$; $\mathrm{GCV}_2 = (3.15/15)/(1 - 3/15)^2 = 0.210/0.640 = 0.328$; $\mathrm{GCV}_4 = (2.98/15)/(1 - 5/15)^2 = 0.199/0.444 = 0.448$.

GCV selects degree 2. The degree-4 model reduces RSS by only 0.17 but the GCV denominator penalizes the higher leverage heavily, reflecting the variance cost of extra parameters.

Example 2: CV for Regularization Parameter Selection

Consider ridge regression with $p = 20$ predictors and $n = 50$. Use 5-fold CV to select $\lambda \in \{0.01, 0.1, 1, 10, 100\}$.

For each λ, the ridge estimator is $\hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1} X^\top Y$. Partition the data into 5 folds of 10 observations each.

Suppose the average held-out MSE values are: $\lambda = 0.01$: 2.85; $\lambda = 0.1$: 2.41; $\lambda = 1$: 2.12; $\lambda = 10$: 2.35; $\lambda = 100$: 3.90.

The minimum is at $\lambda = 1$. The one-standard-error rule: if $\mathrm{SE}(\widehat{\mathrm{MSE}}_{\lambda=1}) = 0.30$, select the largest λ with CV error within $2.12 + 0.30 = 2.42$. This is $\lambda = 10$ (CV error 2.35), yielding a sparser, more regularized model.

Cross-Links

AI. Cross-validation is the standard method for hyperparameter tuning in machine learning. In deep learning, the computational cost of K-fold CV for large models is prohibitive; single validation splits or learning-curve extrapolation are common surrogates. Stacking (Node 10.6) uses CV predictions as meta-features.

Economics. Out-of-sample forecasting evaluation is the time-series analogue of CV. Rolling-window and expanding-window designs respect temporal dependence, replacing random fold assignment with forward-chaining to avoid information leakage.

Linear Algebra. The hat matrix $H = X(X^\top X)^{-1} X^\top$ is the orthogonal projector onto the column space of X. The diagonal entries $H_{ii} \in [1/n, 1]$ are the leverages, and the LOOCV shortcut exploits the algebraic structure of rank-one updates to projectors.

Quantum Mechanics. In quantum process tomography, cross-validation selects the complexity of the process matrix reconstruction (e.g., the rank of the Choi matrix) by holding out subsets of measurement data and evaluating predictive log-likelihood on the held-out measurements.

Structural Compression

Invariant

The CV estimator approximates the generalization error by averaging hold-out losses over rotated train–test splits; as the number of folds increases, bias decreases but variance increases due to correlated fold estimates.

Mechanism

Hold-out data provides an unbiased estimate of test error for the model trained on the complement; averaging over folds reduces variance of the estimate. The LOOCV shortcut for linear smoothers exploits the Sherman–Morrison formula to avoid explicit refitting.

Cross-System Echo

Cross-validation is the standard generalization error estimator in machine learning. LOOCV is algebraically related to the jackknife (a variance estimation method via leave-one-out resampling). Stone's theorem establishes the asymptotic equivalence of LOOCV and AIC for linear models, unifying the model-based and model-free approaches to model selection.

Exercises

1. Derive the LOOCV shortcut for linear regression: starting from $\hat{\beta}^{(-i)} = (X_{-i}^\top X_{-i})^{-1} X_{-i}^\top Y_{-i}$, use the Sherman–Morrison formula to show that $Y_i - X_i^\top \hat{\beta}^{(-i)} = (Y_i - \hat{Y}_i)/(1 - H_{ii})$.

2. Prove that LOOCV is exactly unbiased for the expected prediction error of an estimator trained on $n-1$ observations, i.e., $\mathbb{E}[\widehat{R}^{\mathrm{LOO}}] = R_{n-1}$.

3. For a dataset with $n = 100$ and a linear model with $k = 10$ parameters, compute the GCV score given $\mathrm{RSS} = 85.0$. Then compute the exact LOOCV score if the leverage values are approximately constant at $H_{ii} \approx k/n = 0.1$.

4. Show that the variance of the LOOCV estimator satisfies $\mathrm{Var}(\widehat{R}^{\mathrm{LOO}}) \leq \frac{1}{n}\mathrm{Var}(e_i) + \frac{n-1}{n}|\mathrm{Cov}(e_i, e_j)|$, and explain when the covariance term dominates.

5. Implement Stone's theorem: for a Gaussian linear model, show that $n \cdot \widehat{R}^{\mathrm{LOO}}$ and $\mathrm{RSS} + 2\hat{\sigma}^2 k$ differ by $O_p(1)$, so that model rankings under LOOCV and AIC agree asymptotically.

6. In a time-series context, explain why random K-fold CV is invalid and describe the forward-chaining (time-series CV) procedure. State the bias–variance implications of expanding-window vs. fixed-window designs.

7. Argue, structurally, why cross-validation must have higher variance than AIC for model selection in correctly specified parametric models, and identify the settings where this variance cost is justified by the elimination of model-specification assumptions.

Statistics · Module 10 — Model Selection Learning · Node 10.3
Concepts: bias-variance-tradeoff, cost-function
Prerequisites: 10.1, 3.1
Enables: 10.5

hbar.university — own.your.cognition