CS 524: Introduction to Optimization
Lecture 31 : SVM & cross-validation

Michael Ferris

Computer Sciences Department
University of Wisconsin-Madison

November 15, 2023

Choosing a value of the tuning parameter

Each regularization method has an associated tuning parameter: e.g., this was λ in the smoothing spline problem, and λ for lasso and ridge regression in the penalized forms (or t in the constrained forms)

The tuning parameter controls the amount of regularization, so choosing a good value of the tuning parameter is crucial. Because each tuning parameter value corresponds to a fitted model, we also refer to this task as model selection

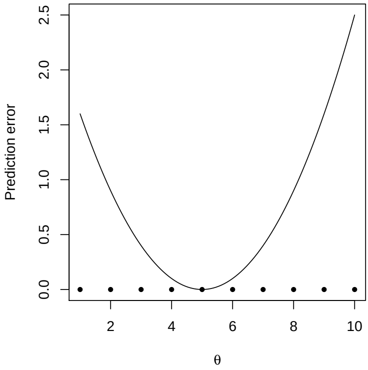
What we might consider a good choice of tuning parameter, however, depends on whether our goal is prediction accuracy or recovering the right model for interpretation purposes. We'll cover choosing the tuning parameter for the purposes of prediction; choosing the tuning parameter for the latter purpose is a harder problem

Cross-validation

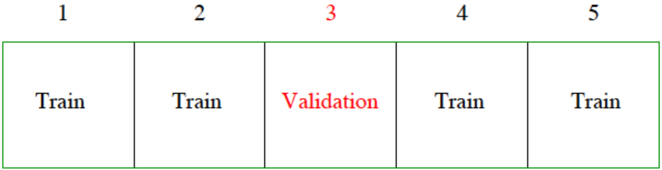
Cross-validation is a simple, intuitive way to estimate prediction error

Given training data (a_i, y_i), $i = 1, \ldots, m$ and an estimator ϕ_θ depending on a tuning parameter θ

Even if θ is a continuous parameter, it's usually not practically feasible to consider all possible values of θ, so we discretize the range and consider choosing θ over some discrete set $\{\theta_1, \ldots, \theta_p\}$



For a number K, we split the training pairs into K parts or "folds" (commonly $K = 5$ or $K = 10$)



K-fold cross validation considers training on all but the kth part, and then validating on the kth part, iterating over $k = 1, \ldots K$

(When $K = m$, we call this leave-one-out cross-validation, because we leave out one data point at a time)

K-fold cross validation procedure:

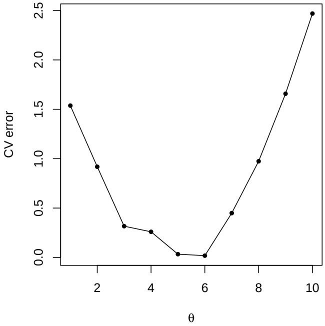
- Divide the set $\{1, \ldots, m\}$ into K subsets (i.e.,folds) of roughly equal size, $F_1, \ldots, F_K$
- For $k = 1, \ldots, K$:
  - Consider training on (a_i, y_i), $i \notin F_k$, and validating on (a_i, y_i), $i \in F_k$
  - For each value of the tuning parameter $\theta \in \{\theta_1, \ldots, \theta_p\}$, compute the estimate ϕ_θ^{-k} on the training set, and record the total error on the validation set:

$$e_k(\theta) = \sum_{i \in F_k} (y_i - \phi_\theta^{-k}(a_i))^2$$

- For each tuning parameter value θ, compute the average error over all folds,

$$CV(\theta) = \frac{1}{m} \sum_{k=1}^{K} e_k(\theta) = \frac{1}{m} \sum_{k=1}^{K} \sum_{i \in F_k} (y_i - \phi_\theta^{-k}(a_i))^2$$

Having done this, we get a cross-validation error curve $\mathrm{CV}(\theta)$ (this curve is a function of θ), e.g.,



and we choose the value of tuning parameter that minimizes this curve,

$$\hat{\theta} = \arg \min_{\theta \in \{\theta_1, \ldots, \theta_p\}} CV(\theta)$$

Recap: cross validation

Training error, the error of an estimator as measured by the data used to fit it, is not a good surrogate for prediction error. It just keeps decreasing with increasing model complexity

Cross-validation, on the other hand, much more accurately reflects prediction error. If we want to choose a value for the tuning parameter of a generic estimator (and minimizing prediction error is our goal), then cross-validation is the standard tool

We usually pick the tuning parameter θ that minimizes the cross-validation error curve. Sometimes called "Hyper-parameter estimation"

Cross-validation is a general tool

So far we've looked at cross-validation for estimation under squared error loss, but it applies much more broadly than this.
For an arbitrary loss $\ell(y_i - \phi(a_i))$, the cross-validation estimate of prediction error under ℓ is

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i \in F_k} \ell(y_i - \phi^{-k}(a_i))$$

E.g., for classification, each $y_i \in \{0, 1\}$, and we might want to use the 0-1 loss

$$\ell(y_i - \phi(a_i)) = \begin{cases} 0 & \text{if } y_i = \phi(a_i) \\ 1 & \text{if } y_i \neq \phi(a_i) \end{cases}$$

Cross-validation now gives us an estimate of misclassification error for a new observation. Usually in cross-validation for classification we try to balance the folds.