Generalization and Validation

ORIE 4741

October 21, 2020

Table of Contents

1. Bias-Variance Tradeoff
2. Model Selection versus Model Assessment
3. Cross Validation

Recap: the bias-variance decomposition

- Assume $Y = f(X) + \epsilon$ where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\varepsilon^2$
- For a fit $\hat{f}(X)$ at an input point $X = x_0$, the prediction error using squared-error loss is:

Recap: the bias-variance decomposition

- Assume $Y = f(X) + \epsilon$ where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\varepsilon^2$
- For a fit $\hat{f}(X)$ at an input point $X = x_0$, the prediction error using squared-error loss is:

$$
\begin{aligned}
\mathrm{Err}\left(x_0\right) &= \mathrm{E}\left[\left(Y - \hat{f}\left(x_0\right)\right)^2 \mid X = x_0\right] \\
&= \sigma_\varepsilon^2 + \left[\mathrm{E}\hat{f}\left(x_0\right) - f\left(x_0\right)\right]^2 + \mathrm{E}\left[\hat{f}\left(x_0\right) - \mathrm{E}\hat{f}\left(x_0\right)\right]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2\left(\hat{f}\left(x_0\right)\right) + \mathrm{Var}\left(\hat{f}\left(x_0\right)\right) \\
&= \text{ Irreducible Error } + \text{ Bias }^2 + \text{ Variance.}
\end{aligned}
$$

Recap: the bias-variance decomposition

- Assume $Y = f(X) + \epsilon$ where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\varepsilon^2$
- For a fit $\hat{f}(X)$ at an input point $X = x_0$, the prediction error using squared-error loss is:

$$\begin{aligned}
\mathrm{Err}\left(x_0\right) &= \mathrm{E}\left[\left(Y - \hat{f}\left(x_0\right)\right)^2 \mid X = x_0\right] \\
&= \sigma_\varepsilon^2 + \left[\mathrm{E}\hat{f}\left(x_0\right) - f\left(x_0\right)\right]^2 + \mathrm{E}\left[\hat{f}\left(x_0\right) - \mathrm{E}\hat{f}\left(x_0\right)\right]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2\left(\hat{f}\left(x_0\right)\right) + \mathrm{Var}\left(\hat{f}\left(x_0\right)\right) \\
&= \text{ Irreducible Error } + \text{ Bias }^2 + \text{ Variance.}
\end{aligned}$$

- Increasing model complexity will decrease bias but increase variance.

Recap: the bias-variance decomposition

- Assume $Y = f(X) + \epsilon$ where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\varepsilon^2$
- For a fit $\hat{f}(X)$ at an input point $X = x_0$, the prediction error using squared-error loss is:

$$\begin{aligned}
\mathrm{Err}\left(x_0\right) &= \mathrm{E}\left[\left(Y - \hat{f}\left(x_0\right)\right)^2 \mid X = x_0\right] \\
&= \sigma_\varepsilon^2 + \left[\mathrm{E}\hat{f}\left(x_0\right) - f\left(x_0\right)\right]^2 + \mathrm{E}\left[\hat{f}\left(x_0\right) - \mathrm{E}\hat{f}\left(x_0\right)\right]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2\left(\hat{f}\left(x_0\right)\right) + \mathrm{Var}\left(\hat{f}\left(x_0\right)\right) \\
&= \text{ Irreducible Error } + \text{ Bias }^2 + \text{ Variance.}
\end{aligned}$$

- Increasing model complexity will decrease bias but increase variance.
  - Linear regression with/without quadratic regularization

Lasso regression: minimize $\|y - Xw\|^2 + \lambda||w||_1$

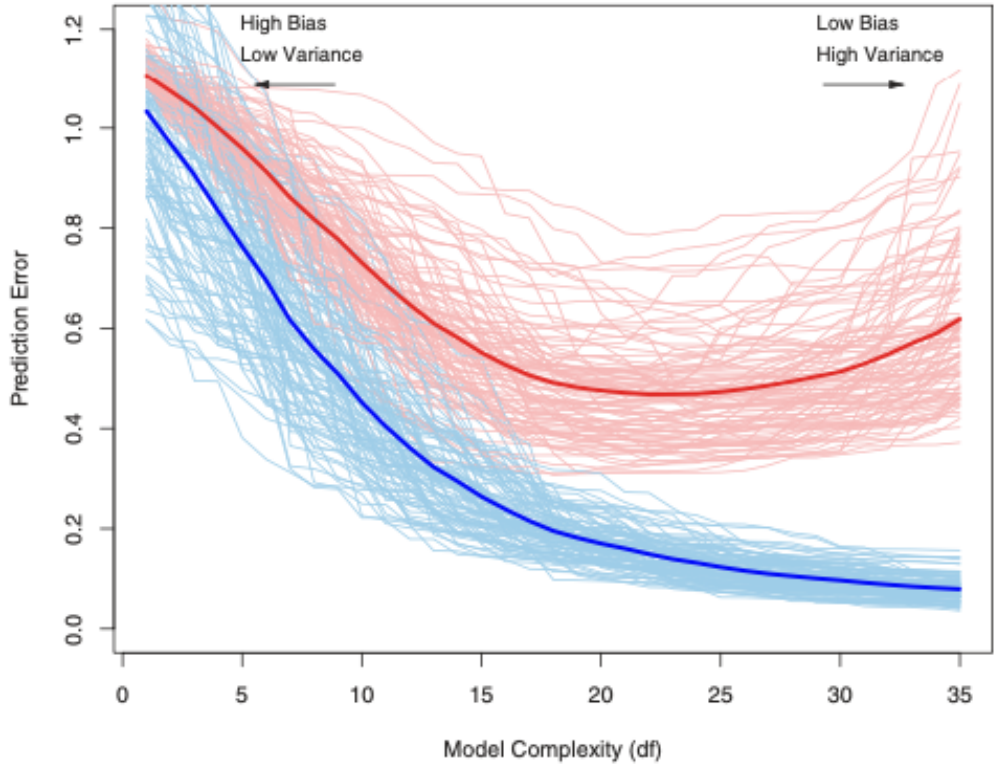


Figure 1: The light blue curves show the training error err, while the light red curves show the corresponding test error for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error and the expected training error.

Model Selection versus Model Assessment

- Model selection: estimating the performance of different models in order to choose the best one.
- Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.
- Both needs a different set from the training set: validation set and the test set.

When do you need model selection?

When do you need model selection?

- There are different models to choose from:
  - Regularization parameter λ: minimize $\|y - Xw\|^2 + \lambda\|w\|^2$.
  - Linear model with different feature transformation.
  - k in k-nearest neighbors
  - Linear model with quadratic regularization or k-nearest neighbors?

When do you need model selection?

- There are different models to choose from:
  - Regularization parameter λ: minimize $\|y - Xw\|^2 + \lambda\|w\|^2$.
  - Linear model with different feature transformation.
  - k in k-nearest neighbors
  - Linear model with quadratic regularization or k-nearest neighbors?
    - Select validation set
    - Selection some values of λ and some values of k
    - Find the model with smallest prediction error on the validation set

Different behaviour for different loss function

For 80 observations and 20 predictors uniformly distributed in the hypercube $[0,1]^{20}$, consider the following two situation:

1. Y is 0 if $X_1 \leq 1/2$ and 1 if $X_1 > 1/2$, and we apply k-nearest neighbors.
2. Y is 0 if $\sum_{j=1}^{10} X_j > 5$ and 0 otherwise, and we use best subset linear regression of size p.

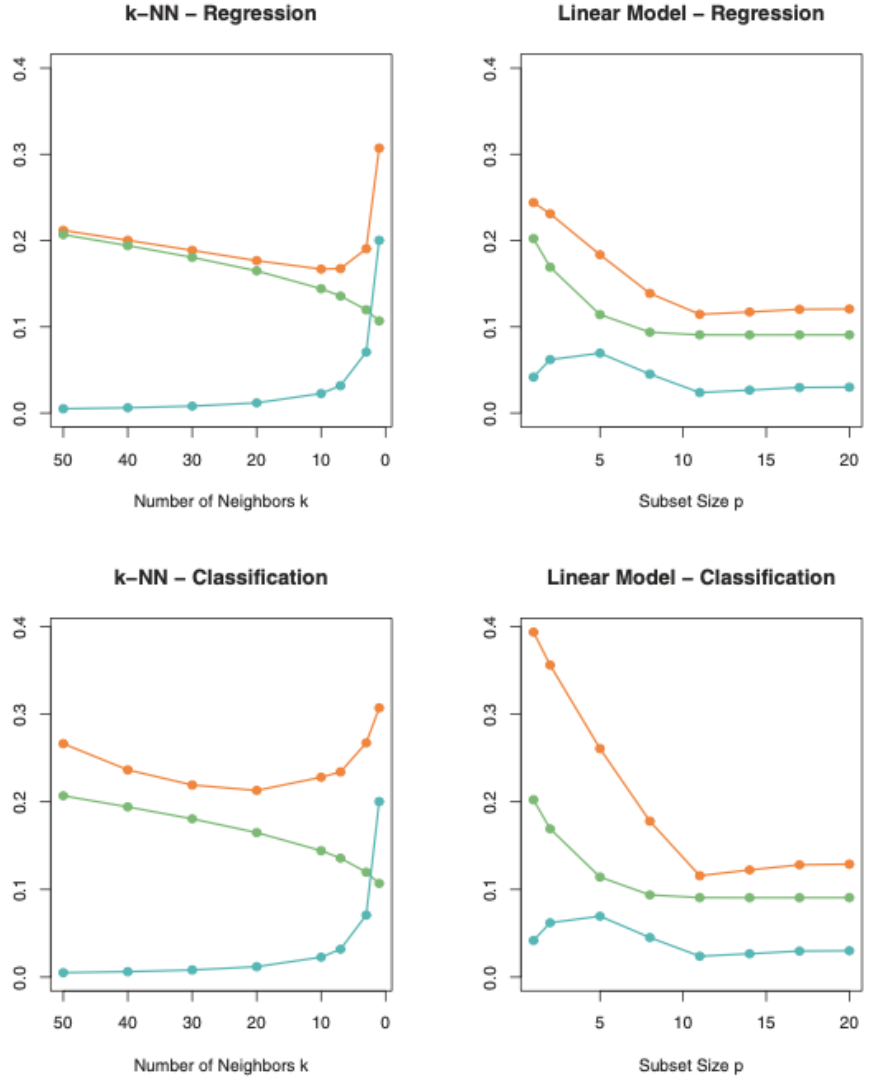


Figure 2: Expected prediction error (orange), squared bias (green) and variance (blue). The top row is regression with ℓ_2 loss; the bottom row is classification with 0–1 loss. Left: $Y = 0$ when $X_1 < .5$. Right: $Y = 0$ when $\sum_{j=1}^{10} X_j > 5$.

Goal of using validation set

- Estimate the prediction error on new data (or test set).

Goal of using validation set

- Estimate the prediction error on new data (or test set).
- If we had enough data, we would set aside a validation set.
- If the data is scarce, K-fold cross validation uses part of the data to fit the model, and a different part to test it.
- Using cross validation, we can evaluate on every available data point.

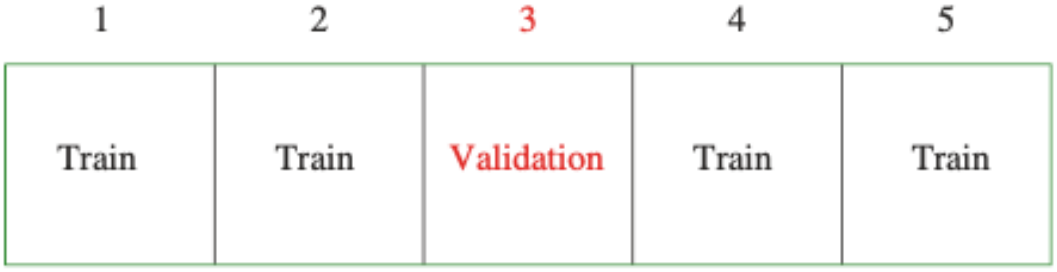


Figure 3: 5-fold cross validation split.

K-Fold Cross Validation

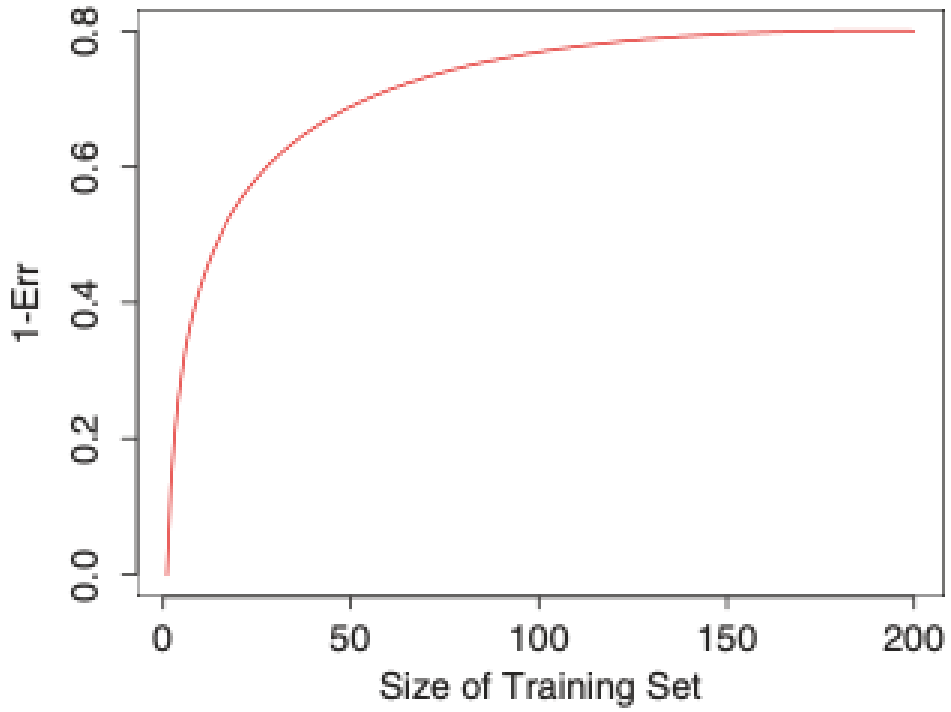
- Let $\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by the randomization.
- Denote by $\hat{f}^{-k}(x, \alpha)$ the model with "tuning parameter" α, fitted with the kth part of the data removed.
- The cross validation estimate of prediction error:

$$\mathrm{CV}(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i, \alpha\right)\right)$$

- Find the tuning parameter $\hat{\alpha}$ that minimizes it.
- The final chosen model is $f(x, \hat{\alpha})$ should be fitted to all the data.

Size of Training Set

- With K-folds cross validation, we have training set of size $N \times \frac{K-1}{K}$.
- For large K, the computation is too expensive. For small K, the model may underfit due to insufficient data points. Usually use $K = 5$ or $K = 10$.



One-Standard-Error Rule

Choose the most parsimonious model whose error is no more than one standard error above the error of the best model.

- For 80 observations and 20 predictors uniformly distributed in the hypercube $[0, 1]^{20}$
- Y is 0 if $\sum_{j=1}^{10} X_j > 5$ and 0 otherwise, and we use best subset linear regression of size p
- Apply best subset linear regression.

One-Standard-Error Rule

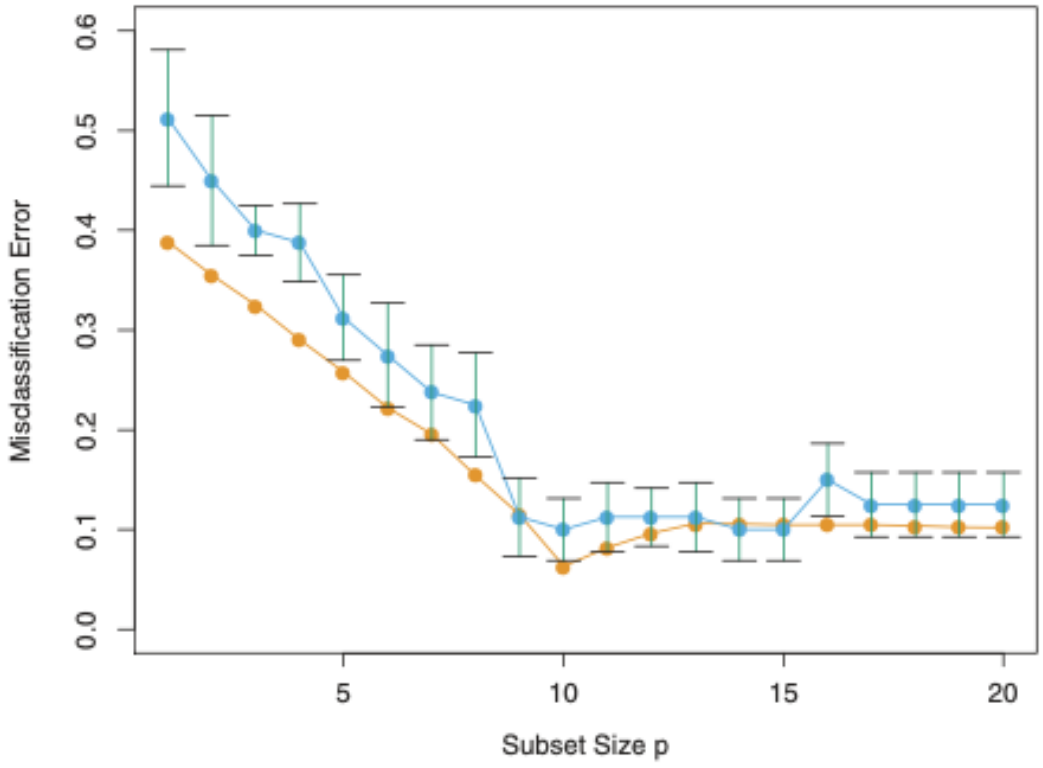


Figure 4: Prediction error (orange) and tenfold cross-validation curve (blue) estimated from a single training set

- If using one-standard-error rule , choose $p = 9$; otherwise, choose $p = 10$.

A cross validation strategy

Consider a classification problem with a large number of predictors. A typical strategy for analysis might be as follows:

1. Screen the predictors: find a subset of "good" predictors that show fairly strong (univariate) correlation with the class labels.
2. Using just this subset of predictors, build a multivariate classifier.
3. Use cross-validation to estimate the unknown tuning parameters and to estimate the prediction error of the final model.

What is wrong here?

- Consider $N = 50$ samples in two equal-sized classes, and $p = 5000$ quantitative predictors (standard Gaussian) that are independent of the class labels.
- The true (test) error rate of any classifier is 50%.

What is wrong here?

- Consider $N = 50$ samples in two equal-sized classes, and $p = 5000$ quantitative predictors (standard Gaussian) that are independent of the class labels.
- The true (test) error rate of any classifier is 50%.

1. choosing the 100 predictors having highest correlation with the class labels
2. using a 1-nearest neighbor classifier, based on just these 100 predictors
3. Over 50 simulations from this setting, the average CV error rate was 3%. This is far lower than the true error rate of 50%.

Right Way to Do Cross Validation

- The predictors were chosen on the basis of all of the samples and thus had unfair advantage.
- These predictors "have already seen" the left out samples.

Right Way to Do Cross Validation

- The predictors were chosen on the basis of all of the samples and thus had unfair advantage.
- These predictors "have already seen" the left out samples.

1. choosing the 100 predictors having highest correlation with the class labels, using all of the samples except those in fold k.
2. using a 1-nearest neighbor classifier, based on just these 100 predictors using all of the samples except those in fold k.
3. Use the classifier to predict the class labels for the samples in fold k.

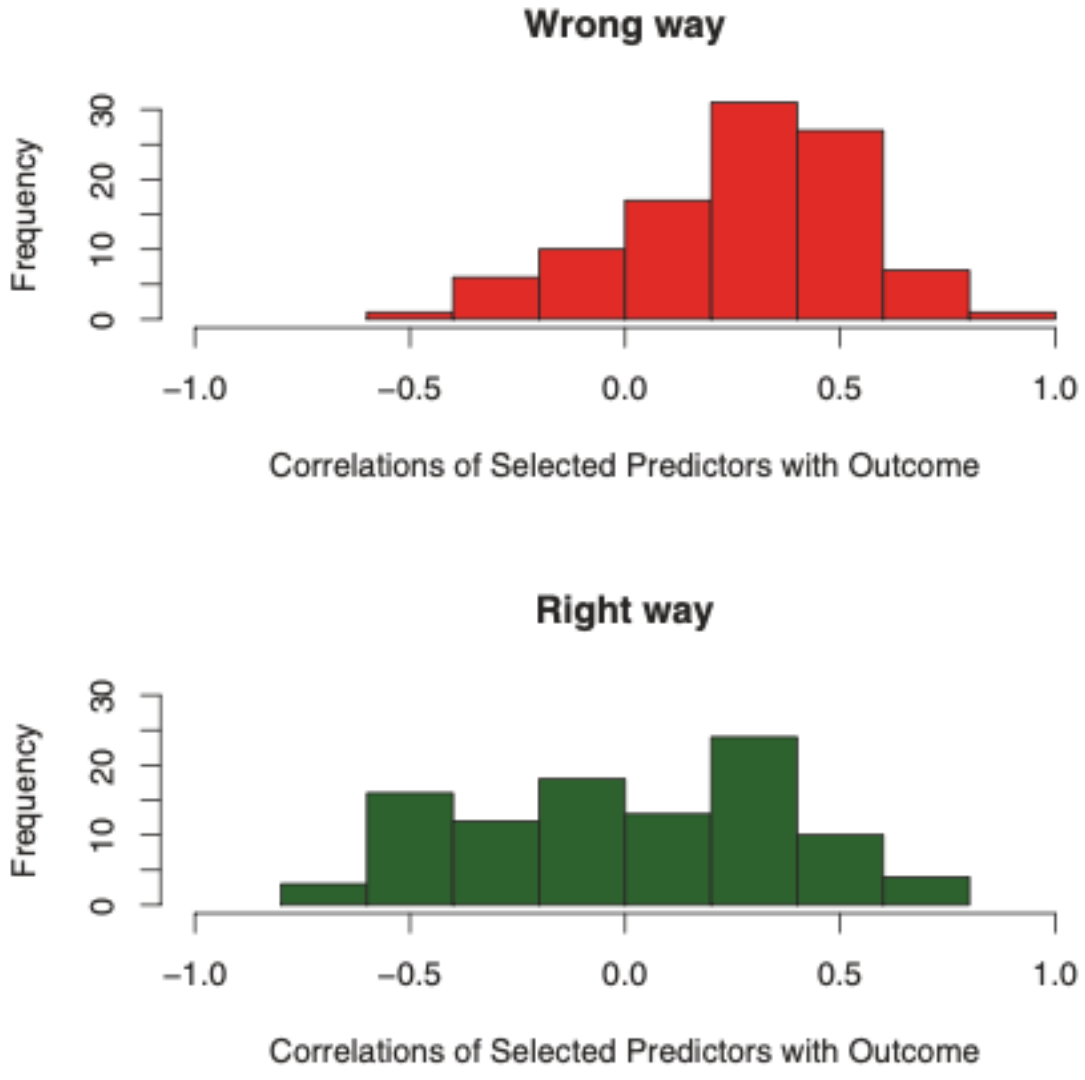


Figure 5: Histograms shows the correlation of class labels, in 10 randomly chosen samples, with the 100 predictors chosen using the wrong and correct versions of cross-validation.