Resampling and Evaluation Methods

Tianhang Zheng

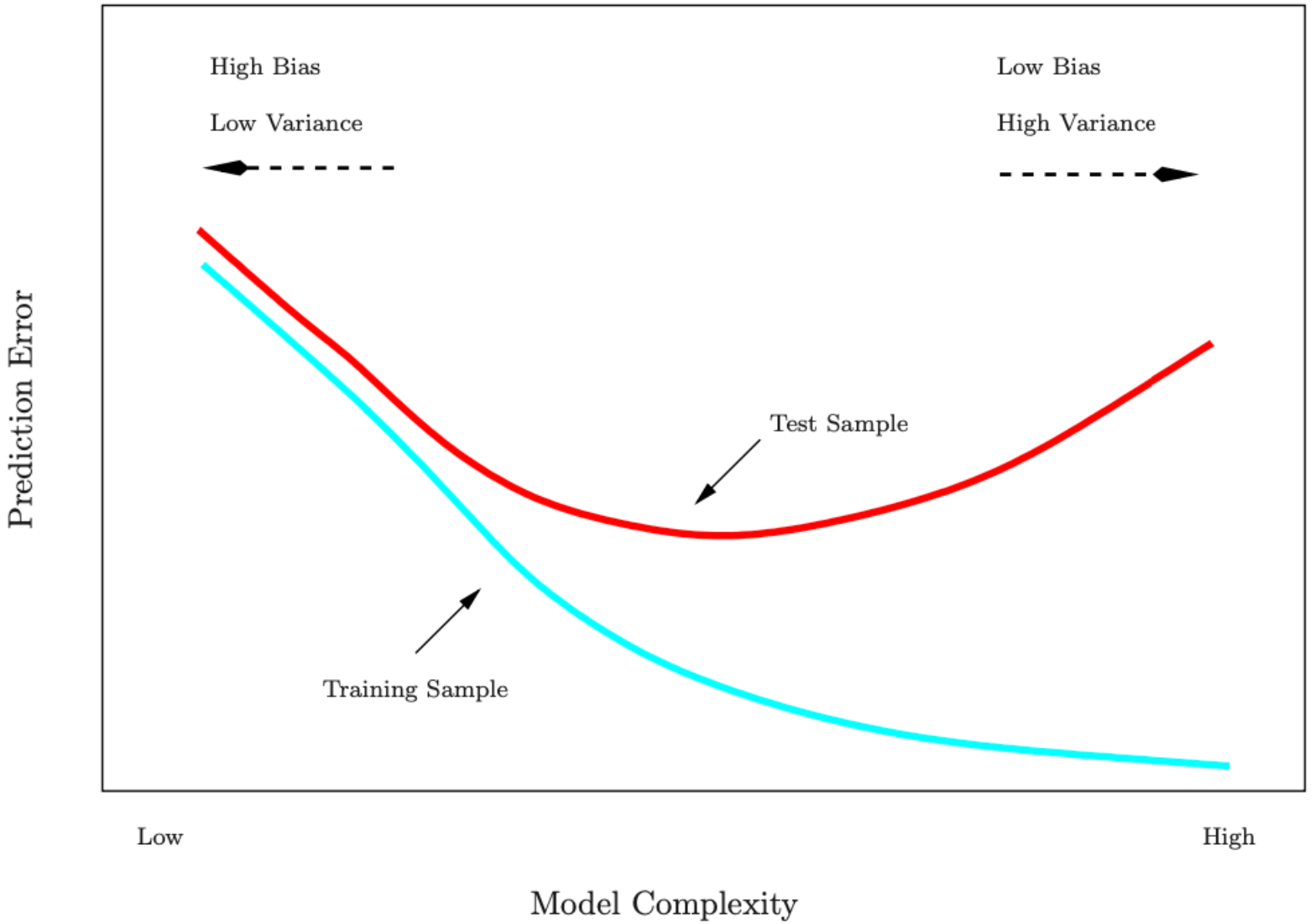
https://tianzheng4.github.io

Cross-validation and the Bootstrap

These methods refit a model of interest to samples formed from the training set, to obtain additional information about the fitted model.

For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

Training Error versus Test error



Prediction-error estimates

Cannot use training error to estimate prediction error

Best solution: a large designated test set **but often not available**

Hold out a subset of the training observations from the fitting process (**validation set**), and then applying the statistical learning method to those held out observations (validation samples).

Validation-Set Approach

We randomly divide the available set of samples into two parts: a training set and a validation or hold-out set.

The model is fit on the training set, and the fitted model is used to **predict the responses for the observations in the validation set.**

The resulting validation-set error provides an estimate of the test error.

Drawbacks of Validation-Set Approach

The validation estimate of the test error can vary a lot

In the validation approach, only a subset of the observations are used to fit the model

Thus, the validation set error may tend to **overestimate** the test error for the model fit on the entire data set.

K-Fold Cross-validation

Randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other K − 1 parts (combined), and then obtain predictions for the left-out kth part.

This is done in turn for each part k = 1,2,...K, and then the results are combined.

K-Fold Cross-validation

1	2	3	4	5
Validation	Train	Train	Train	Train

Denote K parts by $C_1, C_2, C_3 \ldots, C_K$

$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k \qquad \mathrm{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$$

Leave-One Out Cross-Validation (LOOCV)

K=n K-fold cross validation

LOOCV is sometimes useful, but the estimates are highly correlated, so the average has high variance

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$$

A better choice is K=5 or 10

Bias vs Variance

The estimates of prediction error will typically be biased since the training data size is n(K-1)/K

LOOCV is almost unbiased but incurs high variance

K=5 or 10 provides a good compromise for this bias-variance tradeoff.

Cross-Validation for Classification Problems

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{Err}_k \qquad \mathrm{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$$

$$\widehat{\mathrm{SE}}(\mathrm{CV}_K) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \frac{(\mathrm{Err}_k - \overline{\mathrm{Err}_k})^2}{K-1}}$$

Cross-Validation: Right and Wrong

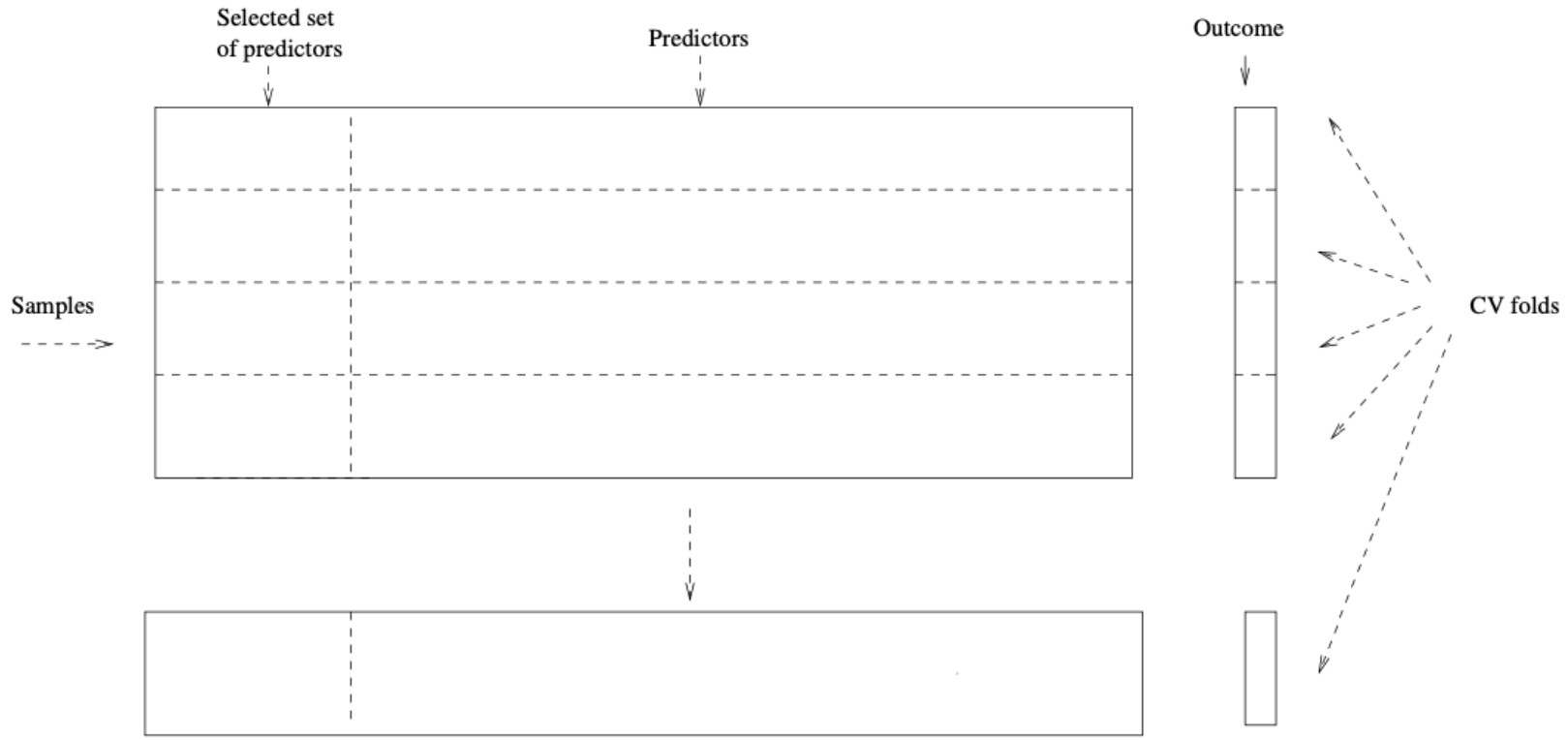
Consider a simple classifier applied to some two-class data:

1. Starting with 5000 predictors and 50 samples, find the 100 predictors having the largest correlation with the class labels.
2. We then apply a classifier such as logistic regression, using only these 100 predictors.

How should we conduct cross validation?

Cross-Validation: Right and Wrong

Apply cross-validation to both step 1 and 2



BootStrap

We obtain distinct datasets by repeatedly sampling observations from the original dataset with replacement.

Each “bootstrap dataset” is created by sampling with replacement and is the same size as our original dataset.

Use the bootstrap to get an estimate of a parameter

BootStrap

Primarily used to obtain standard errors of an estimate.

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2}.$$

Provide approximate confidence intervals for a population parameter.

BootStrap to Estimate Prediction Error?

To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.

But there is large overlap between training and validation set, which will underestimate the error

Q & A