Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

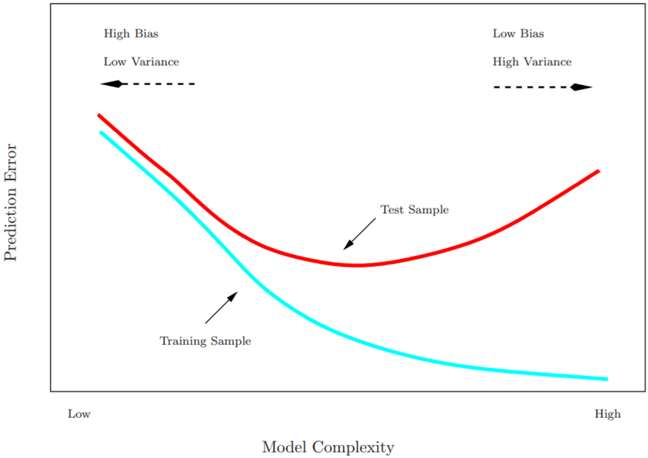
Contents

Outline

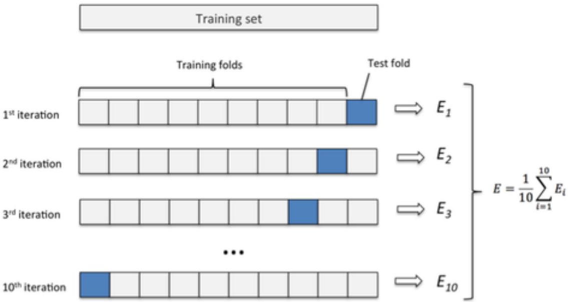
- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

Outline

Training- versus Test-Set Performance



- Widely used approach for estimating test error.
- Idea is to 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.



in general $k = 5$ or 10

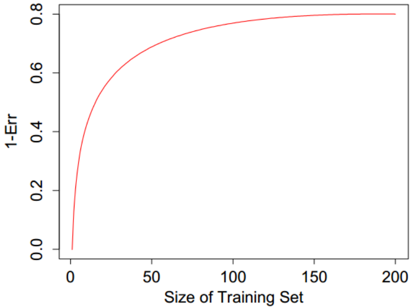
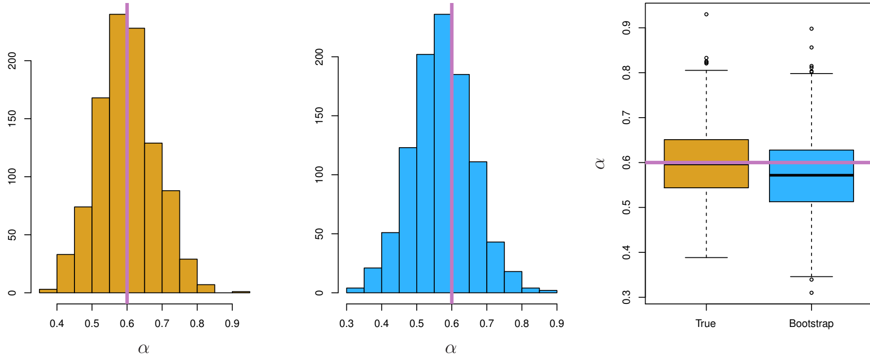


FIGURE 7.8. *Hypothetical learning curve for a classifier on a given task: a plot of* $1 - \mathrm{Err}$ *versus the size of the training set* N. *With a dataset of* 200 *observations, 5-fold cross-validation would use training sets of size* 160, *which would behave much like the full set. However, with a dataset of* 50 *observations fivefold cross-validation would use training sets of size* 40, *and this would result in a considerable overestimate of prediction error.*

Outline

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.
- see example in the book



Estimating α

In practice (back to the real world)

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets **by repeatedly sampling observations from the original data set with replacement**.

- Each of these "bootstrap data sets" is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Contents

Outline

We could improve simple least square regression by variable selection and feature engineering (precisely here it's feature selection).

- Despite its simplicity, the linear model has distinct advantages in terms of its interpretability and often shows good predictive performance.
- Hence we discuss in this lecture some ways in which the simple linear model can be improved, by replacing ordinary least squares fitting with some alternative fitting procedures.

Outline

How to select among the p variables / predictors ?

1. **Subset Selection** : We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
2. **Shrinkage** : We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as **regularization**) has the effect of reducing variance and can also perform variable selection.
3. **Dimension Reduction (PCA)** : We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Outline

- Best Subset selection , exhaustive method ;
- Forward Stepwise selection, start with no predictor and had one by one ;
- Backward Stepwise selection, start will all the predictors and eliminate one by one.

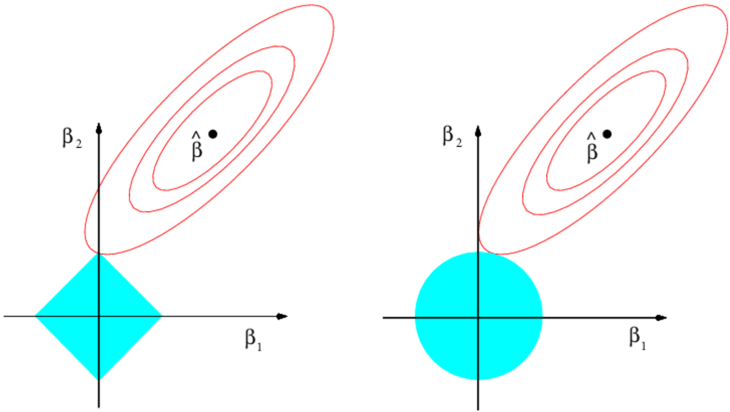
Outline

Ridge regression and Lasso

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

We know define our cost function as

- $J(\beta) = \sum_{i=1}^{N}(h(X^i) - Y_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$
  $= \sum_{i=1}^{N}(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_i X_{i,p} - Y_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$
  $= \sum_{i=1}^{N}(\beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} - Y_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$ as written in the book, with $\lambda \geq 0$;
- $\lambda \sum_{j=1}^{p} \beta_j^2$ is a penalization parameter used in the Ridge Regression (it is an L^2 norm) ;
- $\lambda \sum_{j=1}^{p} |\beta_j|$ is a penalization parameter used in the Lasso (Regression) (it is an L^1 norm) ;
- The greater is λ , the greater is the amount of shrinkage ;
- With the Lasso some β coefficients can goes to zero (fig 6.7) ;
- In general, λ is chosen by cross-validation


β_2
$\hat{\beta}$
β_1
β_2
$\hat{\beta}$
β_1

Outline

- The methods that we have discussed so far in this chapter have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors, $X_1, X_2, ..., X_p$.
- We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as dimension reduction methods

- Here we apply **Principal Components Analysis (PCA)** (discussed in Chapter 10 of the text) to define the linear combinations of the predictors, for use in our regression.
- The first principal component is that (normalized) linear combination of the variables with the largest variance.
- The second principal component has largest variance, subject to being uncorrelated with the first. And so on.
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

Outline

- Cross-validation is a generic tool to measure accuracy and avoid overfitting ;
- Model selection, shrinkage, PCA may be included in the feature engineering works