BIOS 835: Model Assessment

Alexander McLain

October 5, 2023

Outline

Introduction

Recall two of our goals in subset selection of linear models

1. *Prediction accuracy:* to make reasonable predictions or estimations, we need
   - **accuracy** $\rightarrow$ on average, what we estimate is equal to what we expect (e.g., $\hat{Y}$ in the long run is equal to the population mean of Y)
   - **precision** $\rightarrow$ small variation in prediction/estimation
2. *Interpretation:* if we can limit the number of variables, we can get a better idea of what are the "main factors" that are driving the outcome.

So far we've been using (rather blindly) validation and cross-validation measure prediction accuracy.

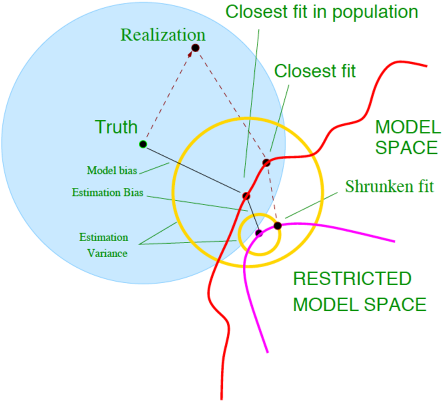


Figure: From ESL (online version) page 225.

Loss and expected loss

- Assume we have an outcome Y which we are predicted with $\hat{f}(X)$ which has been fitted with a training set $\mathcal{T}$.
- The loss function measures the error in $\hat{f}(X)$ when predicting Y.
- Common loss function are:

$$L\{Y, \hat{f}(X)\} = \begin{cases} (Y - \hat{f}(X))^2 & \text{squared error} \\ |Y - \hat{f}(X)| & \text{absolute error} \end{cases}$$

Loss and expected loss

- *Test error*, also referred to as *generalization error*, is the prediction error over an independent test sample

$$\mathrm{Err}_{\mathcal{T}} = E[L\{Y, \hat{f}(X)\}|\mathcal{T}]$$

where X and Y are considered to be random.

- A related (but different) quantity is the *expected prediction (or test) error*

$$\mathrm{Err} = E[L\{Y, \hat{f}(X)\}] = E(\mathrm{Err}_{\mathcal{T}})$$

where we don't condition on $\mathcal{T}$.

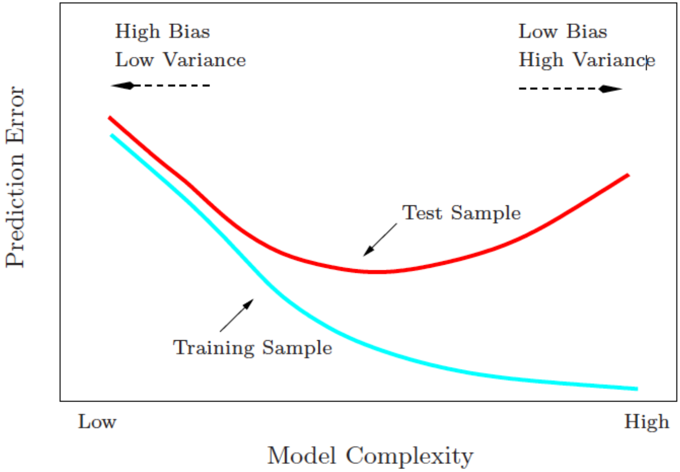
Loss and expected loss

- *Training error*, as we have discussed, is a less desirable quantity defined as

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L\{y_i, \hat{f}(x_i)\}.$$

- *Training error* will under estimate both $\text{Err}_{\mathcal{T}}$ and Err, and will choose overly complex models.

Bias-Variance Tradeoff



Loss and expected loss for categorical outcomes

- Suppose our outcome G takes one of K values in a set $\mathcal{G}$ labeled $1, 2, \ldots, K$.
- Some common loss functions for categorical outcomes are:

$$\begin{aligned} L\{G, \hat{G}(X)\} &= I\{G = \hat{G}(X)\} && \text{0-1 loss} \\ L\{G, \hat{p}(X)\} &= -2 \log \hat{p}_G(X) && -2 \times \text{log-likelihood} \end{aligned}$$

  where $p_k(X) = \Pr(G = k|X)$.
- $\text{Err}_{\mathcal{T}}$, Err and $\overline{\text{err}}$ are all defined analogously.
- The bias–variance tradoff behaves differently for 0-1 loss versus squared error loss.

In-sample error

- The $\overline{\text{err}}$ will underestimate $\text{Err}_{\mathcal{T}}$ or Err, but by how much and can we estimate it?
- For the moment let's consider the x values as fixed, and the observed x's are the only ones of interest.
- In this case, $\text{Err}_{\mathcal{T}}$ is known as the **in-sample error** defined as

$$\text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} E_{Y_0}[L\{Y_i^0, \hat{f}(x_i)\}|\mathcal{T}]$$

where Y^0 is used to denote that we observe new y values at each of the training points x_i.

Optimism

- **Optimism** is then defined as

$$\text{op} = \text{Err}_{\text{in}} - \overline{\text{err}}$$

  which is estimated with the average optimism $\omega = E_y(op)$ that is averaged over all training sets (though X is still fixed).

- For squared error, 0–1, and other loss functions, one can show that

$$\omega = \frac{2}{N} \sum_{i=1}^{N} \text{Cov}(\hat{y}_i, y_i).$$

- When would this be large?

Optimism

- Summarizing the above, we have

$$E_y(\text{Err}_{\text{in}}) = E_y(\overline{\text{err}}) + \frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i).$$

- If $\hat{y}_i$ can be expressed as a function of d linear inputs we can simplify ω to

$$\omega = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i) = \frac{2d\sigma_\epsilon^2}{N},$$

  where $Y = f(X) + \epsilon$
- Further, $E_y(\text{Err}_{\text{in}}) = E_y(\overline{\text{err}}) + 2\frac{d}{N}\sigma_\epsilon^2$.

Using Optimism to estimate Err_{in}

- Thus, if we have an estimate of ω we can estimate the in-sample prediction error via
$$\widehat{\text{Err}}_{\text{in}} = \overline{\text{err}} + \hat{\omega}.$$
- Motivated by the results above, one estimate is
$$C_p = \overline{\text{err}} + 2\frac{d}{N}\hat{\sigma}_\epsilon^2$$
which is referred to as Mallows C_p.

Using Optimism to estimate *AIC*

- If we were to use a log-likelihood loss function, similar results to above can show that
$$-2E\{\log \Pr_{\theta}(Y)\} \approx -\frac{2}{N}E(\text{loglik}) + 2\frac{d}{N}.$$
- Which motivates *Akaike information criteria* (AIC)
$$AIC = -\frac{2}{N}\text{loglik} + 2\frac{d}{N}$$
- For the Gaussian model the AIC statistics is equivalent to C_p
- AIC can be used to select tuning parameters without CV or validation.

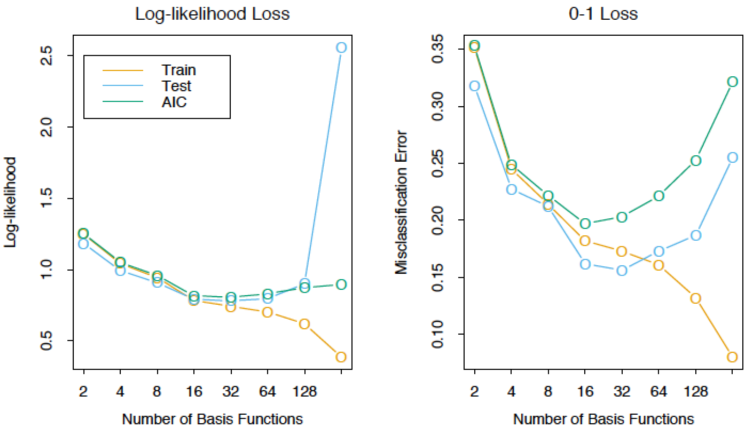


Figure: From ESL (online version) page 232.

Introduction

- Cross-validation is a technique we've used frequently to estimate tuning parameters.
- Cross-validation typically only estimates the expected prediction error Err.
- The effectiveness of CV depends on the size of your data set and the relationship between Err and the training sample size.

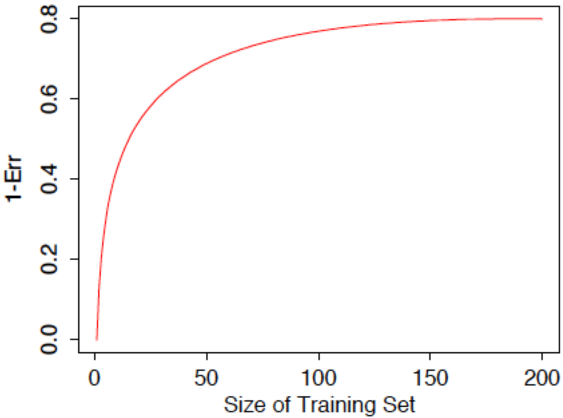


Figure: From ESL (online version) page 243.

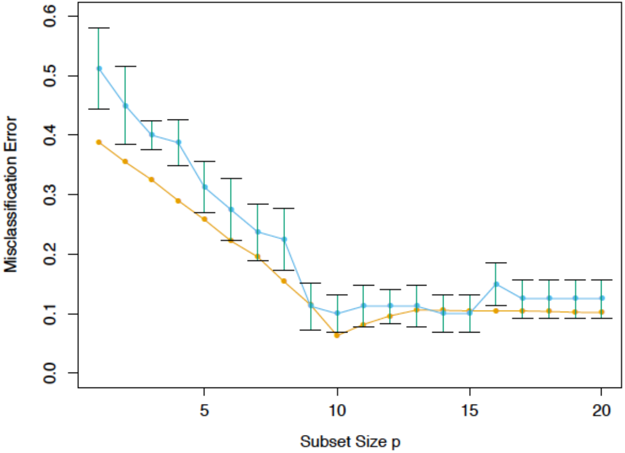


Figure: From ESL (online version) page 244.

Generalized cross-validation

- **Generalized cross-validation** is an approximation to leave one out cross-validation.
- Assume that $\hat{y} = \boldsymbol{S}\boldsymbol{y}$ (i.e., a linear model).
  - Note that $\boldsymbol{S}$ is commonly denoted by $\boldsymbol{H}$ and called the **Hat Matrix**.
- For many linear fitting methods,

$$\frac{1}{N}\sum_{i=1}^{N}\{y_i - \hat{f}^{-i}(x_i)\}^2 = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}\right\}^2$$

where S_{ii} is the ith diagonal element of $\boldsymbol{S}$.

Generalized cross-validation

- *Generalized cross-validation* (GVC) approximates this quantity with

$$\mathrm{GVC}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{y_i - \hat{f}(x_i)}{1 - \mathrm{trace}(\boldsymbol{S})/N}\right\}^2.$$

Recall that $\mathrm{trace}(\boldsymbol{S})$ is the *effective number of parameters*.

Cross-validation done wrong

Read the following and tell me what you think. This sequence can be done in genomic or proteomic applications.

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.

Cross-validation done right

1. Divide the data into K samples (folds).
2. For each fold k:
   2.1 Screen the predictors: find a subset of "good" predictors that show fairly strong (univariate) correlation with the class labels (leaving out fold k)
   2.2 Using just this subset of predictors, build a multivariate classifier (leaving out fold k).
   2.3 Estimate the prediction error of the final model for fold k.

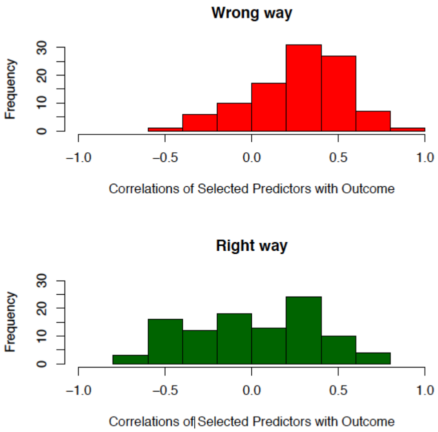


Figure: From ESL (online version) page 246.

Bootstrapping

- The bootstrap is a general tool for assessing statistical accuracy.
- We can use the bootstrap to estimate prediction error, and optimism.