

# RISK ESTIMATION

## -INTRODUCTION TO DATA SCIENCE-

ISL 5.1, 6.1.3

Lecturer: Darren Homrighausen, PhD

# Preamble:

- Discuss how the training error is **optimistic** and how to correct for this optimism
- Directly estimate the risk by resampling

# RISK ESTIMATION

Reminder: Prediction risk is

$$R(f) = \mathbb{E}\ell(f(X), Y) \leftrightarrow \text{Bias} + \text{Variance}$$

The overriding theme is that we would like to add a judicious amount of bias to get **lower** risk

As  $R$  isn't known, we need to estimate it

As discussed,  $\hat{R} = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i)$  isn't very good  
(In fact, one tends to not add bias when estimating  $R$  with  $\hat{R}$ )

# RISK ESTIMATION: A GENERAL FORM

The problem is that  $\hat{R}$  is overly **optimistic**

The **average optimism** is

$$\text{opt} =^* \mathbb{E}[R - \hat{R}]$$

Typically,  $\text{opt}$  is positive as  $\hat{R}$  will underestimate the risk

(\* See ESL, Chapter 7 for details for a more precise statement)

## RISK ESTIMATION: A GENERAL FORM

It turns out for a variety of  $\ell$  (such as squared error and 0-1)

$$\text{opt} = \frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{f}(X_i), Y_i)$$

This is related intimately with **degrees of freedom**

$$\text{df} = \frac{1}{\sigma^2} \sum_{i=1}^n \text{Cov}(\hat{f}(X_i), Y_i) = \frac{n}{2\sigma^2} \text{opt}$$

$$(\sigma^2 = \mathbb{V}Y_i)$$

**EXAMPLE:** For multiple regression (i.e.  $\hat{f}(X) = \hat{\beta}_{LS}^\top X$ ),

$$\text{df} = \text{trace}(\mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top) = \text{rank}(\mathbb{X})$$

# A RISK ESTIMATE

Therefore, we get the following expression of risk

$$\text{GIC} = \hat{R} + \widehat{\text{opt}}$$

(Writing GIC indicates **generalized information criterion**)

Differing  $\widehat{\text{opt}}$  lead to different risk estimators

$\widehat{\text{opt}}$  depends on:

- a variance estimator  $\hat{\sigma}$
- a scaling term

# VARIOUS FORMS OF RISK ESTIMATES

Akaike's information criterion:  $AIC = \hat{R} + \frac{2}{n} \cdot \text{df} \cdot \hat{\sigma}^2$

Mallow's  $C_p = \hat{R} - \hat{\sigma}^2 + \frac{2}{n} \cdot \text{df} \cdot \hat{\sigma}^2$

Schwarz information criterion:  $BIC = \hat{R} + \frac{\log(n)}{n} \cdot \text{df} \cdot \hat{\sigma}^2$

Including more parameters leads to:

- a smaller  $\hat{R}$
- a larger  $\widehat{\text{opt}}$

**GOAL:** Now, we can use one of the GIC procedures to tell us which model to use

(As long as  $\log n \geq 2$ , BIC picks a **smaller** model than AIC)

# VARIOUS FORMS OF RISK ESTIMATES: AIC AND BIC

Akaike's Information Criterion (AIC) and the Schwarz/Bayesian Information Criterion (BIC) have alternative formulations:

$$\begin{array}{ll} \text{AIC} = \hat{R} + \frac{2}{n} \cdot \text{df} \cdot \hat{\sigma}^2 & \text{or} \quad n \log(\hat{R}) + 2 \cdot \text{df} \\ \text{BIC} = \underbrace{\hat{R} + \frac{\log(n)}{n} \cdot \text{df} \cdot \hat{\sigma}^2}_{\text{Use whenever}} & \text{or} \quad \underbrace{n \log(\hat{R}) + \log(n) \cdot \text{df}}_{\text{Only use when } n \geq p} \end{array}$$



# Cross-validation

# A DIFFERENT APPROACH TO RISK ESTIMATION

Let  $(X_0, Y_0)$  be a test observation, identically distributed as an element in  $\mathcal{D}$ , but also **independent** of  $\mathcal{D}$ .

$$R(f) = \ell(f(X_0), Y_0) \underbrace{=}_{\text{regression}} \mathbb{E}(Y_0 - f(X_0))^2$$

Of course, the quantity  $(Y_0 - f(X_0))^2$  is an unbiased estimator of  $R(f)$  and hence we could use it to estimate  $R(f)$

However, **we don't have any such new observation**

Or do we?

# AN INTUITIVE IDEA

Let's set aside one observation and predict it

**For example:** Set aside  $(X_1, Y_1)$  and fit  $\hat{f}^{(1)}$  on  $(X_2, Y_2), \dots, (X_n, Y_n)$

(The notation  $\hat{f}^{(1)}$  just symbolizes leaving out the first observation before fitting  $\hat{f}$ )

$$R_1(\hat{f}^{(1)}) = (Y_1 - \hat{f}^{(1)}(X_1))^2$$

As the left off data point is **independent** of the data points used for estimation,

$$\mathbb{E}R_1(\hat{f}^{(1)}) \approx R(\hat{f})$$

# LEAVE-ONE-OUT CROSS-VALIDATION

Cycling over all observations and taking the average produces  
leave-one-out cross-validation

$$\text{CV}_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^n R_i(\hat{f}^{(i)}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}^{(i)}(X_i))^2.$$

# MORE GENERAL CROSS-VALIDATION SCHEMES

**SOME NOTATION:** Suppose  $v$  is a **set**. Then  $|v|$  is the number of elements in that set

Example: if  $v = \{1, 4, 10, -\pi\}$  then  $|v| = 4$

Let  $\mathcal{N} = \{1, \dots, n\}$  be the index set for  $\mathcal{D}$

**K-FOLD:** Fix  $V = \{v_1, \dots, v_K\}$  such that  $v_j \cap v_k = \emptyset$  and  $\bigcup_j v_j = \mathcal{N}$

$$CV_K(\hat{f}) = \frac{1}{K} \sum_{v \in V} \frac{1}{|v|} \sum_{i \in v} (Y_i - \hat{f}^{(v)}(X_i))^2$$

(There are others, mainly a bootstrap version)

Here,  $|v| \approx \frac{n}{K}$

(Example, choosing  $K = 2$  splits the data in half and hence there are  $|v| = \frac{n}{2}$  observations in each fold)

# MORE GENERAL CROSS-VALIDATION SCHEMES: A COMPARISON

- $CV_K$  gets more computationally demanding as  $K \rightarrow n$   
(As we have to train  $\hat{f}^{(v)}$   $|K|$  times and  $(v)$  has fewer observations)
- The bias (as a risk estimator) of  $CV_K$  goes down, but the variance increases as  $K \rightarrow n$
- The bootstrap version isn't commonly used

## SUMMARY TIME

CV	+	Good at selecting models that make good predictions
	+/-	Generally selects a model larger than necessary
	-	Is computationally demanding, especially if $K$ is large
AIC	+	Good at selecting models that make good predictions (and is asymptotically equivalent to CV)
	+/-	Generally selects a model larger than necessary
	-	
BIC	+	Good at selecting the correct model (if this exists)
	-	Generally selects model with poor prediction risk

Aside: There exist impossibility theorems stating that risk estimation procedures good at prediction are bad at model selection (and vice-versa)

# RISK ESTIMATION IN A DATA RICH ENVIRONMENT

If we have a large amount of data, we can split into three parts:

- **TRAINING:** Used to fit (or **train**) the considered procedures
- **VALIDATION:** Used to score these trained procedures
- **TESTING:** Used to estimate the prediction risk for the selected procedure

(A typical split might be 50%/25%/25%)



# Postamble:

- Discuss how the training error is **optimistic** and how to correct for this optimism  
(This generates AIC, BIC, Mallows, GCV, ...)
- Directly estimate the risk by resampling  
(Most commonly done with K-Fold CV or the bootstrap)