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

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

Typically, 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 ℓ (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

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

(Writing GIC indicates generalized information criterion)

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

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

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

Various forms of risk estimates

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

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

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

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:

$$\mathrm{AIC} = \hat{R} + \frac{2}{n} \cdot \mathrm{df} \cdot \hat{\sigma}^2 \qquad \text{or} \qquad n \log(\hat{R}) + 2 \cdot \mathrm{df}$$

$$\mathrm{BIC} = \underbrace{\hat{R} + \frac{\log(n)}{n} \cdot \mathrm{df} \cdot \hat{\sigma}^2}_{\text{Use whenever}} \qquad \text{or} \qquad \underbrace{n \log(\hat{R}) + \log(n) \cdot \mathrm{df}}_{\text{Only use when } n \geq p}$$

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 $(X_2, Y_2), \ldots, (X_n, Y_n)$

(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

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

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

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

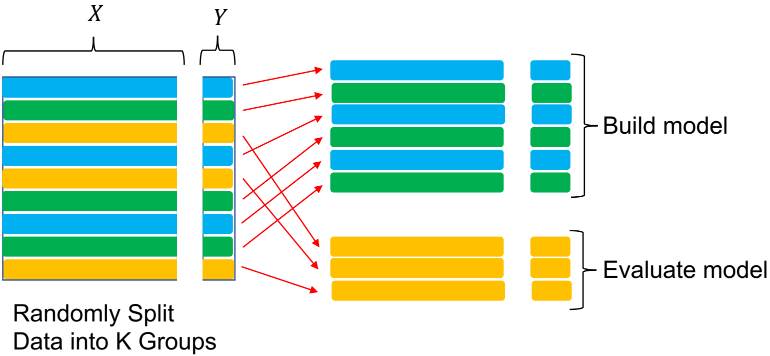
$$\mathrm{CV}_K(\hat{f}) = \frac{1}{K} \sum_{v \in V} \frac{1}{|v|} \sum_{i \in v} (Y_i - \hat{f}^{(v)}(X_i))^2$$

(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



More general cross-validation schemes: A comparison

- CV_K gets more computationally demanding as $K \to n$
  (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 $K \to n$
- 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)