Linear Methods for Regression: Risk estimation

-Statistical Machine Learning-

Lecturer: Darren Homrighausen, PhD

Subset selection and regularization

For now, let's assume we are doing ordinary least squares, and hence the design (feature) matrix is $\mathbb{X} \in \mathbb{R}^{n \times p}$.

We want to do model selection for at least three reasons:

- Prediction accuracy: Can essentially *always* be improved by introducing some bias
- Interpretation: A large number of features can sometimes be distilled into a smaller number that comprise the "big (little?) picture"
- Computation: A large p can create a huge computational bottleneck.

Subset selection and regularization

We will address three related ideas

- Model selection: Selection of only some of the original p features
- Dimension reduction/expansion: Creation of new features to help with prediction
- Regularization: Add constraints to optimization problems to provide stabilization

Risk estimation

Reminder: Prediction risk is

$$R(f) = \mathbb{P}_{Z,\mathcal{D}}\ell_f \leftrightarrow \text{Bias} + \text{Variance}$$

The overridding 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, $\hat{R}_{\text{train}} = \hat{\mathbb{P}}\ell_f$ isn't very good

(In fact, one tends to not add bias when estimating R with $\hat{\mathbb{P}}\ell_f$)

$\hat{R}_{\text{train}}$ tends to underestimate R, hence we can call it optimistic

Risk estimation: a general form

Assume that we get a new draw of the training data, $\mathcal{D}^0$, such that $\mathcal{D} \sim \mathcal{D}^0$ and

$$\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \quad \text{and} \quad \mathcal{D}^0 = \{(X_1, Y_1^0), \ldots, (X_n, Y_n^0)\}$$

If we make a small compromise to risk, we can form a sensible suite of risk estimators

To wit, letting $Y^0 = (Y_1^0, \ldots, Y_n^0)^\top$, define

$$R_{in} = \mathbb{E}_{Y^0|\mathcal{D}} \hat{\mathbb{P}}_{\mathcal{D}^0} \ell_{\hat{f}} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y^0|\mathcal{D}} \ell(\hat{f}(X_i), Y_i^0)$$

Then the average optimism is

$$\mathrm{opt} = \mathbb{E}_Y [R_{in} - \hat{R}_{\mathrm{train}}]$$

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

Risk estimation: a general form

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

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

Therefore, we get the following expression of risk

$$\mathbb{E}_Y R_{in} = \mathbb{E}_Y \hat{R}_{\mathrm{train}} + \frac{2}{n}\sum_{i=1}^{n} \mathrm{Cov}(\hat{f}(X_i), Y_i),$$

which has unbiased estimator (i.e. $\mathbb{E}_Y R_{\mathrm{gic}} = \mathbb{E}_Y R_{in}$)

$$R_{\mathrm{gic}} = \hat{R}_{\mathrm{train}} + \frac{2}{n}\sum_{i=1}^{n} \mathrm{Cov}(\hat{f}(X_i), Y_i)$$

Degrees of freedom

We call the term (where $\sigma^2 = \mathbb{V} Y_i$)

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

the degrees of freedom

(This is really the effective number of parameters, with some caveats)

Our task now is to either estimate or compute opt to produce $\widehat{\mathrm{opt}}$ and form:

$$\hat{R}_{\mathrm{gic}} = \hat{R}_{\mathrm{train}} + \widehat{\mathrm{opt}}$$

This leads to Mallows Cp/Stein's unbiased risk estimatior (SURE), as well as forms for AIC, BIC, and others

Degrees of freedom: Example

Sometimes the df is exactly computable.

(In other cases, it needs to be estimated)

Look at least squares regression onto $\mathbb{X}$, with $\mathbb{V}Y_i = \sigma^2$

Information criteria

Of course, this isn't the usual way to introduce/conceptualize information criteria

For me, thinking of the training error as overly optimistic and correcting for that optimism is conceptually appealing

For others, forming a metric[1] on probability measures is more appealing

Let's go over this now for completeness

[1] It will turn out to be a psuedo-metric; a small detail

Comparing probability measures

Kullback-Leibler

Suppose we have data Y that comes from the probability density function f.

What happens if we use the probability density function g instead?

Example: Suppose $Y \sim N(\mu, \sigma^2) = f$. We want to predict a new Y_*, but we model it as $Y_* \sim N(\mu_*, \sigma^2) = g$

How far away are we? We can either compare μ to μ_* or Y to Y^*

(This is the approach taken via the optimism)

Or, we can compute how far f is from g

(far indicates we need a notion of distance)

Kullback-Leibler

One central idea is Kullback-Leibler discrepancy[2]

$$\begin{aligned} KL(f,g) &= \int \log\left(\frac{f(y)}{g(y)}\right) f(y)dy \\ &\propto -\int \log(g(y))f(y)dy \qquad \text{(ignore term without } g\text{)} \\ &= -\mathbb{P}_f[\log(g(Y))] \end{aligned}$$

This gives us a sense of the loss incurred by using g instead of f.

[2] This has many features of a distance, but is not a true distance as $KL(f,g) \neq KL(g,f)$.

Kullback-Leibler discrepancy

Usually, g will depend on some parameters, call them θ

Example:In regression, we can specify $f = N(X^\top \beta, \sigma^2)$ for a fixed (true)[3] β, and let $g_\theta = N(X^\top \beta, \sigma^2)$ over all $\theta \in \mathbb{R}^p \times \mathbb{R}^+$

As $KL(f, g_\theta) = -\mathbb{P}_f[\log(g_\theta(Y))]$, we minimize this over θ.

Again, $\mathbb{P}_f$ is unknown, so we minimize $-\log(g_\theta(Y))$ instead. This is the maximum likelihood value

$$\hat{\theta}_{ML} = \arg\max_\theta g_\theta(Y)$$

[3] We actually don't need to assume things about a true model nor have it be nested in the alternative models.

Kullback-Leibler discrepancy

Now, to get an operational characterization of the KL divergence at the ML solution

$$-\mathbb{P}_f[\log(g_{\hat\theta_{ML}}(Y))]$$

we need an approximation (don't know f, still)

This approximation[4] is exactly AIC:

$$\mathrm{AIC} = -\log(g_{\hat\theta_{ML}}(Y)) + |\hat\beta_{ML}|$$

Example: Let $\log(g_\theta(y)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||Y - \mathbb{X}\beta||_2^2$

σ^2 KNOWN: $\hat\beta = \mathbb{X}^\dagger Y$

$$\mathrm{AIC} \propto n\hat R_{\mathrm{train}}/(2\sigma^2) + p = \hat R_{\mathrm{train}} + 2\sigma^2 n^{-1} p$$

σ^2 UNKNOWN: $\hat\beta = \mathbb{X}^\dagger Y$, $n\hat\sigma^2 = (I - \mathbb{X}\mathbb{X}^\dagger)Y = n\hat R_{\mathrm{train}}$

$$\mathrm{AIC} \propto n\log(\hat R_{\mathrm{train}})/2 + p = \log(\hat R_{\mathrm{train}}) + 2n^{-1}p$$

[4] See “Multimodel Inference” Burnham, Anderson (2004)

Summary

For $\hat{R}_{\mathrm{gic}}$:

$$\hat{R}_{\mathrm{train}} + \widehat{\mathrm{opt}} = \hat{R}_{\mathrm{train}} + 2\sigma^2 n^{-1}\mathrm{df} = \begin{cases} \text{AIC, known } \sigma^2 & \\ \text{Mallows Cp} & \text{if } \hat{f}(X) = X^\top \hat{\beta}_{LS} \\ \text{SURE} & \text{most } \hat{f}(X) \end{cases}$$

Or

$$\mathrm{IC} = \log(\hat{R}_{\mathrm{train}}) + c_n n^{-1}\mathrm{df} = \begin{cases} \text{AIC, unknown } \sigma^2 & \text{if } c_n = 2 \\ \text{BIC} & \text{if } c_n = \log(n) \end{cases}$$

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 $\mathcal{D}$.

Prediction risk: $R(f) = \mathbb{E}(Y_0 - f(X_0))^2$

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

$$\mathbb{E}_{(X_1,Y_1)|\mathcal{D}_{(1)}} R_1(\hat{f}^{(1)}) \overset{D}{=} R(\hat{f}(\mathcal{D}_{n-1})) \approx R(\hat{f}(\mathcal{D}))$$

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

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

Define a distribution $\mathcal{V}$ over $\mathcal{N}$ with (random) variable v

Then, we can form a general cross-validation estimator as

$$\mathrm{CV}_{\mathcal{V}}(\hat{f}) = \mathbb{E}_{\mathcal{V}} \hat{\mathbb{P}}_{v} \ell_{\hat{f}^{(v)}}$$

More general cross-validation schemes: Examples

$$\mathrm{CV}_{\mathcal{V}}(\hat{f}) = \mathbb{E}_{\mathcal{V}}\hat{\mathbb{P}}_{v}\ell_{\hat{f}(v)}$$

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

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

- Bootstrap: Let $\mathcal{V}$ be given by the bootstrap distribution over $\mathcal{N}$ (that is, sampling with replacement many times)
- Factorial: Let $\mathcal{V}$ be given by all subsets (or a subset of all subsets) of $\mathcal{N}$ (that is, putting mass $1/(2^n - 2)$ on each subset)

More general cross-validation schemes: A comparison

- CV_K gets more computationally demanding as $K \to n$
- The bias of CV_K goes down, but the variance increases as $K \to n$
- The factorial version isn't commonly used except when doing a 'real' data example for a methods paper
- There are many other flavors of CV. One of them, called "consistent cross validation" [Homework] is a recent addition that is designed to work with sparsifying algorithms

Summary time

Risk estimation methods

CV	Prediction risk consistent (Dudoit, van der Laan (2005)). Generally selects a model larger than necessary (unproven)
AIC	Minimax optimal risk estimator (Yang, Barron (1998)). Model selection inconsistent*
BIC	Model selection consistent (Shao (1997) [low dimensional]. Wang, Li, Leng (2009) [high dimensional]). Slow rate for risk estimation*

(Stone (1977) shows that CV_n and AIC are asymptotically equivalent.)

(*Yang (2005) gives an impossibility theorem: for a linear regression problem it is impossible for a model selection criterion to be both consistent and achieve minimax optimal risk estimation)