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

$$R(f) = \mathbb{P}\ell_f \leftrightarrow \text{Bias} + \text{Variance}$$

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, the training error $\hat{R} = \hat{\mathbb{P}}\ell_f$ isn't very good

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

$\hat{R}$ 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

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

Then the average optimism is

$$\text{opt} = \mathbb{P}_Y[R_{in} - \hat{R}]$$

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

Risk estimation: a general form

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

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

Therefore, we get the following expression of risk

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

which has unbiased "estimator"

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

(i.e. $\mathbb{P}_Y R_{\text{gic}} = \mathbb{P}_Y R_{in}$)

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. Addendum: After a student's question, I wanted to add Kaufman, Rosset (2015) and Janson, Fithian, Hastie (2015) as references)

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

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

This leads to Mallows Cp/Stein's unbiased risk estimator (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)

Example: Least squares regression onto $\mathbb{X}$, with

- $\mathbb{V} Y_i = \sigma^2$
- $\mathrm{Cov}(Y_i, Y_{i'}) = 0$ for $i \neq i'$

Degrees of freedom: Stein's Rule

Suppose that $Y \sim N(\mu, \sigma^2)$ and $f : \mathbb{R} \to \mathbb{R}$ is absolutely continuous

$$\frac{1}{\sigma^2}\mathbb{P}(Y - \mu)f(Y) = \mathbb{P}f'(Y)$$

(Start with standard normal, use integration by parts and differentiate the density)

This extends to the multivariate scenario: $Y \sim N(\mu, \sigma^2 I)$, $f : \mathbb{R}^n \to \mathbb{R}^n$

$$\frac{1}{\sigma^2}\mathbb{P}(Y_i - \mu_i)f_i(Y) = \mathbb{P}\left[\frac{\partial f_i}{\partial Y_i}\bigg|_Y\right]$$

This gives us a powerful result

$$df = \mathbb{P}\left[\sum_{i=1}^{n} \frac{\partial f_i}{\partial Y_i}\bigg|_Y\right]$$

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 ρ.

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

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

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 ρ is from γ

(far indicates we need a notion of distance)

Kullback-Leibler

One central idea is Kullback-Leibler discrepancy

(This has many features of a distance, but is not a true distance as $KL(\rho, \gamma) \neq KL(\gamma, \rho)$)

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

(Here, $\mathbb{P}\ell_\gamma$ indicates $\ell_\gamma = -\log(\gamma(Y))$)

This gives us a sense of the loss incurred by using γ instead of ρ

Kullback-Leibler discrepancy

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

Example: In regression, we let $\gamma_\theta = N(X^\top \beta, \sigma^2)$ over all $\theta \in \mathbb{R}^p \times \mathbb{R}^+$

As $KL(\rho, \gamma_\theta) = -\mathbb{P}_\rho[\log(\gamma_\theta(Y))]$, we minimize this over θ

Again, $\mathbb{P}_\rho$ is unknown, so we can minimize $-\hat{\mathbb{P}} \log(\gamma_\theta(Y))$ instead

This is the maximum likelihood estimator

$$\hat{\theta}_{ML} = \arg\max_\theta \prod_i \gamma_\theta(Y_i)$$

Kullback-Leibler discrepancy

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

$$-\mathbb{P}_\rho[\log(\gamma_{\hat{\theta}_{ML}}(Y))]$$

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

This approximation[2] is exactly AIC:

$$\mathrm{AIC} = \log(\gamma_{\hat{\theta}_{ML}}(Y)) + p$$

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

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

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

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

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

[2] See "Multimodel Inference" Burnham, Anderson (2004)

Summary

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

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

For SURE, if $Y \sim N(\mu, \sigma^2 I)$, and $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$ is some procedure

$$\frac{1}{\sigma^2}\mathbb{P}(Y_i - \mu)\hat{\mu}_i = \mathbb{P}\frac{\partial \hat{\mu}_i}{\partial Y_i}$$

Alternatively:

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

Lastly,

$$\text{GCV} = n^{-1}\left(\frac{||Y - \mathbb{X}\beta||_2}{(1 - \text{df}/n)}\right)^2 \Leftrightarrow \log(\hat{R}) - 2\log(1 - \text{df}/n)$$

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

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,

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

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

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

More general cross-validation schemes: Examples

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

- 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. ("consistent cross validation", "modified cross-validation",...)

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)

Differing notions of risk

Compare

$$\underbrace{\mathbb{P}_Z \ell_{\hat{f}} = \mathbb{P}_{Z|\mathcal{D}} \ell_{\hat{f}}}_{\text{This is actually of interest}} \qquad \text{Versus} \qquad \underbrace{\mathbb{P}_{Z,\mathcal{D}} \ell_{\hat{f}}}_{\text{This is actually what gets estimated}}$$

Note: The X in Z doesn't in general equal any X_i

Hence, we look to a risk that is composed of predicting a new supervisor at every X_i:

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

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%)

Theoretical risk bounds

Suppose we have a set of procedures $\mathcal{F}$

We have some notation of how complex $\mathcal{F}$ is, called $\mathrm{complexity}$

Specify some $\eta \in [0, 1]$

Then, with probability at least $1 - \eta$ for any $f \in \mathcal{F}$

$$\mathbb{P}_Z \ell_f \leq \frac{\hat{\mathbb{P}}_{\mathcal{D}} \ell_f}{(1 - \sqrt{\delta})_+}$$

where

$$\delta = \frac{1}{n}\left(\mathrm{complexity}\left(1 + \log\left(\frac{n}{\mathrm{complexity}}\right)\right) - \log(\eta/4)\right)$$

(This is out of the Vladimir Vapnik "empirical risk minimization" line of research)