Lecture 3

Model Assessment and Selection for Prediction

Presented January 17, 2018
Statistics 211 - Statistical Methods II

Dan Gillen
Department of Statistics
University of California, Irvine
Model complexity

<table>
<thead>
<tr>
<th>In regression analyses, we can base model selection on a pre-specified set of predictor variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>variable selection which includes/excludes a particular variable ('best' subsets regression)</td>
</tr>
<tr>
<td>shrinkage methods which include all predictors but controls the size of the coefficients (one form of this is called ridge regression...more later!)</td>
</tr>
<tr>
<td>Each approach employs a measure of ‘complexity’</td>
</tr>
<tr>
<td>number of covariates</td>
</tr>
<tr>
<td>amount of control on the size of a coefficient</td>
</tr>
<tr>
<td>Generically we will refer to this measure as a tuning parameter</td>
</tr>
</tbody>
</table>
Determining a specific value for the tuning parameter is part of the model selection process.

For best subsets regression the tuning parameter is fairly easy to conceptualize, mainly because we can think in terms of the interpretation of predictors and their associated coefficients.

Other classes of restricted estimators also have associated measures of complexity.

- polynomial transformations
- piecewise polynomials
- natural cubic splines
- smoothing splines
Again, in each case we can still embed the choice of tuning parameter into the model selection process.

- in particular, we can view the determination of the level of complexity of our model as a model selection problem.

The selection process requires a means of assessing any given model:

- test or generalization error
- error observed in an independent sample

Our goal is to develop tools for the joint tasks of model assessment and selection.
Generalization performance

We can formalize model assessment via a loss function and use expected prediction error, EPE, as a criterion for choosing a model.

- choose $f(\cdot)$ which minimizes EPE

$$f^*(\cdot) = \arg\min_{f(\cdot)} E[L(Y, f(X))]$$

Two examples of commonly considered loss functions are:

1. Squared error ($L_2$) loss: $E(Y - f(X))^2$
2. Absolute ($L_1$) loss: $E|Y - f(X)|$
**Generalization performance**

- $L_2$ loss is commonly used for many reasons, and in this case we have $f^*(\cdot) = E[Y|X=x]$, the conditional expectation or regression function.

- In this case there are many ways we can estimate $E[Y|X=x]$, and we would like a framework that can be used to assess, and order, competing choices.
Generalization performance

For a specified outcome variable $Y$ and vector of predictor variables $X$, suppose we have a prediction model $\hat{f}(X)$, the form of which has been determined on the basis of a training sample.

We measure errors between $Y$ and $\hat{f}(X)$ by specifying a loss function $L(Y, \hat{f}(X))$.

The test or generalization error is the expected prediction error over an independent test sample.

$$EPE = E_{X,Y} \left[ L(Y, \hat{f}(X)) \right]$$

the expectation is taken over the joint distribution of $X$ and $Y$

the average error, were the prediction model to be applied to an independent sample from the population
Generalization performance

▶ If we knew the true joint distribution of $(X, Y)$, we could evaluate this expression directly
  ▶ feasible in a simulation study where we know the truth

▶ However, in real life situations we won’t know this joint distribution and so, for a given $\hat{f}(X)$, we need to estimate EPE

▶ A tempting choice could be the training error

$$\text{err} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i))$$
Generalization performance

▶ Unfortunately the training error is not a good estimate of test error
   ▶ the problem is that the estimate \( \hat{y}_i = \hat{f}(x_i) \) uses \( y_i \)
   ▶ the solution is specifically chosen because it does well in predicting the training data

▶ More specifically, the training error consistently decreases with model complexity
   ▶ an extreme case is including a parameter for every observation (a saturated model), so that \( \hat{f}(x_i) = y_i \) and there is zero training error!

▶ A model with zero training error can be viewed as an overfit to the training data and will typically generalize poorly
   ▶ high sampling variability
Model assessment and selection

- We’ve already identified two separate goals we might have in mind: model selection and model assessment

- Model selection deals with estimating the performance of competing models in order to choose the best one
  - estimate the test error distribution across these models
  - choose the model which corresponds to the minimum

- Model assessment deals with evaluating the generalization error when applying the final model to new data
  - the final model is still chosen on the basis of the training data
  - seek an honest assessment of generalization error
Model assessment and selection

In a data-rich situation, we could approach these goals jointly by splitting the data into three parts:

[ Training data ] [ Validation data ] [ Test data ]

• Training data: *fit the models*
  • obtain point estimates for any given model under consideration
  • repeated use across models

• Validation data: *choose between models*
  • estimate the prediction error for model selection
  • repeated use across models

• Test data: *estimate generalization error of the final model*
  • one-time use, at the end of the analysis
Model assessment and selection

Typically, we are not in a position to split the data into three parts.

A compromise might be to split the data into two parts:

\[
\begin{bmatrix}
\text{Training data} \\
\text{Test data}
\end{bmatrix}
\]

and approximate the validation step:
- analytically: \(C_p\), AIC and BIC
- efficiency sample re-use: cross-validation and the bootstrap

Even still, it may not be that splitting into two parts is feasible:
- consider whether or not these methods can be used to obtain reasonable assessments of generalization error.
The bias-variance decomposition

**Squared error loss**

- For a continuous outcome, suppose the data arise from the model

  \[ Y = f(X) + \epsilon \]

  - where \( E[\epsilon] = 0 \) and \( \text{Var}[\epsilon] = \sigma^2 \)

- Under \( L_2 \) loss, the expected prediction error for an estimate \( \hat{f}(\cdot) \) at \( X = x_0 \) can be decomposed as

  \[
  \text{EPE}(x_0) = \sigma^2 + \left\{ E[\hat{f}(x_0)] - f(x_0) \right\}^2 + \text{Var}[\hat{f}(x_0)]
  \]

  - irreducible error + bias\(^2\) + variance
The bias-variance decomposition

Squared error loss

- This decomposition is specific to the $L_2$ loss but can be evaluated for any given estimator.
- For linear regression we have

$$EPE(x_0) = \sigma^2 + \left\{ f(x_0) - E[\hat{f}(x_0)] \right\}^2 + \| h(x_0) \|^2 \sigma^2$$

- where $h(x_0) = x_0 (X^T X)^{-1} X^T$
Assessing EPE

Earlier, we noted that the training error

\[
err = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i))
\]

would not typically be a good estimate of EPE.

In particular, we would expect err to be somewhat lower than the true EPE,
- that is, the estimate would be overly optimistic.

Part of the discrepancy is due to where the evaluation points occur.
- EPE refers to expected error on an independent sample.
- referred to as *extra-sample* error.
### Assessing EPE

Methods that directly estimate the extra-sample error include cross-validation and the bootstrap:
- both involve the clever use and re-use of the training data

Towards an analytic treatment of understanding the nature of the optimism associated with using the training data to evaluate generalization error, we can consider the *in-sample* error:

\[
\text{Err} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_y \left[ \mathbb{E}_{Y_i^{\text{new}}} \left[ L(Y_i^{\text{new}}, \hat{f}(x_i)) \right] \right]
\]

The notation $Y_i^{\text{new}}$ indicates that we observe $n$ new outcome values *at each of the training points* $x_i, i = 1, \ldots, n$. 
### Assessing EPE

- Each of the $n$ components of the in-sample error averages over the randomness in two distributions:
  - the randomness in the observed outcomes in the training data, $y$
  - the randomness in the ‘new’ outcome observation, $Y_{i}^{\text{new}}$

- The **optimism** is defined as the expected difference between the in-sample error and the training error:

  $$\text{op} \equiv \text{Err} - E_y[\text{err}]$$

  - expectation is taken with respect to the sampling distribution based on the training data, $y$
Assessing EPE

For squared error loss, a little algebra leads to

$$\text{op} = \frac{2}{n} \sum_{i=1}^{n} \text{Cov}[\hat{y}_i, y_i]$$
### Assessing EPE

- This definition leads to the relation

\[
\text{Err} = \mathbb{E}_y [\text{err}] + \frac{2}{n} \sum_{i=1}^{n} \text{Cov}[\hat{y}_i, y_i]
\]

- So, the extent to which err is optimistic, as an estimator of Err, depends on how strongly \( y_i \) influences its own prediction.
Assessing EPE

- The expression simplifies if \( \hat{y}_i \) is linear in the \( y \)'s

\[
\hat{y}_i = \sum_{j=1}^{n} \pi_j y_j
\]

so that

\[
op = \frac{2}{n} \sum_{i=1}^{n} E_y [(\hat{y}_i - E_y[\hat{y}_i])(y_i - E_y[y_i])]
\]

\[
= \frac{2}{n} \sum_{i=1}^{n} \pi_i \text{Var}[y_i]
\]
Assessing EPE

For example, under the additive error model

\[ Y = f(X) + \epsilon \]

with \( E[\epsilon] = 0 \) and \( \text{Var}[\epsilon] = \sigma^2 \), we obtain

\[ \text{Err} = \mathbb{E}_y [\text{err}] + \frac{2}{n} \rho \sigma^2 \]

\( \rho \) is the number of parameters fit in the regression.
While decision theory tells us that EPE is a natural criterion for model selection, the in-sample error can still be useful.

- Having an analytic treatment makes the approach convenient.
- Can be effective if we focus on relative differences in error between model options, rather than the absolute error itself.

From the previous relation, the general form of an estimator for Err is

\[ \hat{\text{Err}} = \text{err} + \hat{\text{op}} \]

where \( \hat{\text{op}} \) is an estimate of the optimism.
Estimation of in-sample error

Mallow’s $C_p$

- For the linear model, squared error loss leads to Mallow’s $C_p$ statistic:

$$C_p = \text{err} + \frac{2}{n}p\sigma^2$$

$$= \frac{1}{n} \{ \text{RSS} + 2p\hat{\sigma}^2 \}$$

- The estimate $\hat{\sigma}^2$ is typically taken from a low-bias model
  - the most complex model under consideration

- The $C_p$ statistic penalizes the residual sum of squares by a factor proportional to the number of parameters being estimated
  - the more complex the model, the greater $p$ will be and the greater the penalty
### Estimation of in-sample error

#### Akaike information criterion; AIC

- The Akaike information criterion is a more general estimate of $Err$ when a log-likelihood function is used as the loss function
  - for a model parameterized by $\theta$, we take
    
    $$L(Y, f_\theta(X)) = -2 \log \Pr_\theta(Y|X)$$

  - sometimes referred to as *cross-entropy* loss or *deviance*
  - multiplying by -2 and taking the log makes the loss for the Normal distribution match the squared error loss

- We use this loss function all the time as a means for choosing the ‘best’ model from our training data
  - minimizing the observed loss is maximum likelihood estimation
Estimation of in-sample error

Akaike information criterion; AIC

AIC relies on the following relationship

\[-2E_Y \left[ \log \Pr_{\hat{\theta}}(Y|X) \right] \approx - \frac{2}{n} E_Y[\text{loglike}] + \frac{2p}{n} \]

- this relationship holds asymptotically as \( n \rightarrow \infty \)
- \( \hat{\theta} \) is the maximum likelihood estimate
- ‘loglike’ is the maximized log-likelihood

\[
\text{loglike} = \sum_{i=1}^{n} \log \Pr_{\hat{\theta}}(y_i|X)
\]
Estimation of in-sample error

Akaike information criterion; AIC

- For any general purpose likelihood AIC is defined as

\[ \text{AIC} = -\frac{2}{n} \text{loglike} + 2 \frac{p}{n} \]

- for the Normal model, with \( \hat{\sigma}^2 \) known, this is equivalent to \( C_p \)

- The penalty imposed by AIC is similar to that imposed by \( C_p \)
  - proportional to the number of parameters being estimated

- In more general settings, when the estimator is linear

\[ \hat{y} = Ly \]

we can replace \( p \) with the effective degrees of freedom
\( \text{df} = \text{tr}(L) \) (eg. penalized regression methods)
Estimation of in-sample error

Bayesian information criterion; BIC

- Suppose we have a set of $M$ candidate models, \{M_1, \ldots, M_M\}, each of which have a prior probability associated with it being the true model

\[
Pr(M_m), \quad m = 1, \ldots, M
\]

- $\sum_{m=1}^{M} Pr(M_m) = 1$

- Given training data,

\[
Data = \{(x_i, y_i); i = 1, \ldots, n\},
\]

we wish to choose the ‘best’ model

- in the sense of choosing the model with the highest posterior probability
Estimation of in-sample error

Bayesian information criterion; BIC

- Let $\theta_m$ denote the parameter(s) corresponding to model $\mathcal{M}_m$, $m = 1, \ldots, M$

- Assuming we have a prior distribution $\Pr(\theta_m|\mathcal{M}_m)$ for the parameters, the posterior probability of model $\mathcal{M}_m$ is

$$
\Pr(\mathcal{M}_m|\text{Data}) \propto \Pr(\mathcal{M}_m) \times \Pr(\text{Data}|\mathcal{M}_m) \\
\propto \Pr(\mathcal{M}_m) \\
\propto \int \Pr(\text{Data}|\mathcal{M}_m, \theta_m) \Pr(\theta_m|\mathcal{M}_m) d\theta_m
$$
Estimation of in-sample error

Bayesian information criterion; BIC

- Comparison of two models can be performed via the posterior odds

\[
\frac{\Pr(M_m | \text{Data})}{\Pr(M_l | \text{Data})} = \frac{\Pr(M_m)}{\Pr(M_l)} \times \frac{\Pr(\text{Data} | M_m)}{\Pr(\text{Data} | M_l)}
\]

- the right-most quantity is called the *Bayes factor*
Estimation of in-sample error

Bayesian information criterion; BIC

- Typically, we assume equal prior weight among the models, so that $\Pr(M_m)$ is constant, for $m = 1, \ldots, M$

- For $\Pr(\text{Data}| M_m)$, we can apply a Laplace approximation to the integral above to give

\[
\Pr(\text{Data}| M_m) \approx \log \Pr(\text{Data}| M_m, \hat{\theta}_m) - (\log n) \frac{p_m}{2}
\]

- $\hat{\theta}_m$ is the maximum likelihood estimate
Estimation of in-sample error

Bayesian information criterion; BIC

- If we take our loss function to be $-2 \log \Pr(\text{Data} | \mathcal{M}_m, \hat{\theta}_m)$, this leads to a general form for the *Bayesian information criterion*:

\[
\text{BIC} = -2 \log \text{like} + (\log n)p
\]

- as with the posterior odds, model comparison is performed on a relative basis