STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 7 — September 22

Lecturer: Purnamrita Sarkar *Scribe: Brian Greco & Rohit Arora*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

7.1 Model Selection

7.1.1 F-Test

While the z-test, tests the significance of individual parameters, it does not simultaneously test the significance of parameters in our linear model. To do this we can use the F-test. The null hypothesis for the F-test is that a subset of parameters are zero and the alternative is that at least one of them is non-zero.

Following is the expression for the F-statistic for a linear model, where RSS is the residual sum of squares and k is the number of dropped features in the reduced model.

$$F_{k,n-p-1} \sim \frac{\left(RSS\left(p\right) - RSS\left(p+k\right)\right)/k}{RSS\left(p+k\right)/\left(n-p-k-1\right)}$$

If the F-statistic is too high then we will reject the null hypothesis. Which means that not all dropped features are redundant.

7.1.2 Model building advice

Remember you cannot use the data twice, once for doing model selection and then again for calculating prediction error. As we keep increasing the model complexity, the training error keeps decreasing and after some point the test error increases. This is the point beyond which all you are doing is fitting the noise in your training data, i.e. over-fitting. This is useless because what you really care about is having an estimation procedure that predicts unseen data well, i.e. *generalizes* well.

When selecting which model to use, you may split the data into three parts, training data, *validation* data, and test data. You can fit many different models on the training data, choose a model by determining which gives the smallest error on the validation data, and then test prediction performance on the test data.

Bottom line–your test data should be in a VAULT to be taken out only once for reporting the error!

7.1.3 Model selection criteria

p is the number of predictors in our model, $ln(\mathcal{L})$ is the log-likelihood of the of the data evaluated at the MLE.

1. Mallows $C_p = \hat{R}_{tr} + 2p\hat{\sigma}^2$. Here we replace the R_{tr} by the training error calculated over data.
2. AIC (Akaike Information Criterion) $= -2\ln(\mathcal{L}) + 2p$.
3. BIC (Bayes Information Criterion) $= -2\ln(\mathcal{L}) + log(n)p$. BIC is more conservative.

7.1.4 Similarity between AIC and Mallow's C_p

We know that our residuals are normal and iid, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Hence their likelihood and log-likelihood can be written is

$$\mathcal{L}\left(\epsilon|\hat{\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\hat{\sigma^2}}\epsilon^\top \epsilon\right\} = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\hat{\sigma^2}}\left(y - X\hat{\boldsymbol{\beta}}\right)^\top \left(y - X\hat{\boldsymbol{\beta}}\right)\right\}$$

$$ln\left(\mathcal{L}\right) = C - \frac{\left(y - X\hat{\boldsymbol{\beta}}\right)^\top \left(y - X\hat{\boldsymbol{\beta}}\right)}{2\hat{\sigma^2}} = C - \frac{\hat{R}_{tr}}{2\hat{\sigma^2}}$$

Now we can simplify the formula for AIC

$$AIC = -2\log\left(\mathcal{L}\right) + 2p \approx \frac{\hat{R}_{tr}}{\hat{\sigma^2}} + 2p = \hat{\sigma^2}\left(\hat{R}_{tr} + 2p\right)$$

7.1.5 The optimism of training error

We will show that the training error is always strictly smaller than the test error, known as the "optimism of training error."

Typically we would consider the following Gaussian model:

$$\boldsymbol{y}_{tr} \sim N(\boldsymbol{X}^{tr}\boldsymbol{\beta}, \sigma^2 I) \qquad \boldsymbol{y}_{test} \sim N(\boldsymbol{X}^{test}\boldsymbol{\beta}, \sigma^2 I).$$

Then we define

$$R_{tr} = E_{y^{tr}}\left[\sum_{i=1}^{n}(y_i^{tr} - (\boldsymbol{x}_i^{tr})^T\hat{\boldsymbol{\beta}})^2\right]$$

$$R_{test} = E_{y^{tr}}E_{y^{test}}\left[\sum_{i=1}^{n}(y_i^{test} - (\boldsymbol{x}_i^{test})^T\hat{\boldsymbol{\beta}})^2\right]$$

The test error really is $E_{y^{tr}}\left[E_{y^{test}}\left[\sum_{i=1}^{n}(y_i^{test}-(\boldsymbol{x}_i^{test})^T\hat{\boldsymbol{\beta}})^2|y^{tr}\right]\right]$. We will drop the conditional notation for simplicity. Note that $\hat{\boldsymbol{\beta}}$ depends on the training data and is in fact a linear function of $\boldsymbol{y}^{tr}$.

In order to show this we will use some new setup just because that helps the analysis. We will fix the independent variables at $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. We will observe two sets of response variables from these, one training set $y_1^{tr}, ..., y_n^{tr}$ on which we will calculate the estimate of $\boldsymbol{\beta}$ and a test set and $y_1^{test}, ..., y_n^{test}$, where will we check how well the $\hat{\boldsymbol{\beta}}$ performs in terms of the RSS. In this case the test error is coined "in sample error".

$$\begin{aligned}
R_{tr} - R_{test} &= \sum_{i=1}^{n} E_{y^{tr}}((y_i^{tr})^2 + (\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2 - 2y_i^{tr}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})) \\
&\quad - \sum_{i=1}^{n} E_{y^{tr}}E_{y^{test}}(y_i^{test^2} + (\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}})^2 - 2y_i^{test}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})) \\
&= -2\sum_{i=1}^{n} E\left[y_i(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}) - Ey_iE(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})\right] \\
&= -2\sum_{i=1}^{n} Cov(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta})
\end{aligned}$$

The third line is true because $E_{y^{tr}}E_{y^{test}}(y_i^{test^2}) = E_{y^{test}}(y_i^{test^2}) = E_{y^{tr}}(y_i^{tr^2})$ since y_i^{tr} and y_i^{test} are identically distributed. The fourth line uses the fact that $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{y}^{test}$ are independent! This is crucial here.

Why is the covariance positive? Well, your best linear fit is designed to have positive covariance with the data y_i! In fact, this is intuitively telling us how training error underestimates the test error by looking at the data twice. Furthermore observe that:

$$\begin{aligned}
\sum_{i=1}^{n} Cov(y_i, \hat{y}_i) &= E\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})(\hat{y}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \\
&= E\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y - \boldsymbol{x}_i^T\boldsymbol{\beta}) \\
&= E\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(y - \boldsymbol{X}\boldsymbol{\beta}) \\
&= E((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}))
\end{aligned}$$

The above is equivalent to $\sigma^2E(\boldsymbol{z}^TH\boldsymbol{z})$, where $\boldsymbol{z} \sim N(0, I)$ and $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Using the fact that $tr(H) = tr(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) = tr(I_{k\times k}) = k$, we arrive at $R_{tr} - R_{test} = -2\sigma^2k$, where k is the number of coavriates of our data.

So,

$$R_{test} = R_{tr} + 2\sigma^2 k = \text{Lack of fit + model complexity.}$$