Class 5: Linear Model Selection and Regularization

MSA 8150: Machine Learning for Analytics

Alireza Aghasi

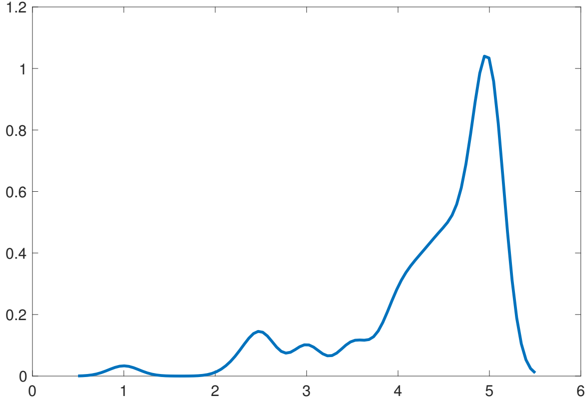
Institute for Insight, Georgia State University

Quiz 1 Results

mean: 4.3542

std: 0.7267

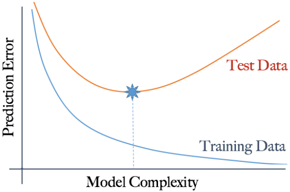
Histogram:



Brief Overview of Cross-Validation

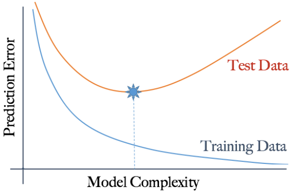
Why Cross Validation?

– As mentioned earlier, model selection based on the RSS or R^2 statistics can be misleading, since the training error is not a good representative of the actual test error



Why Cross Validation?

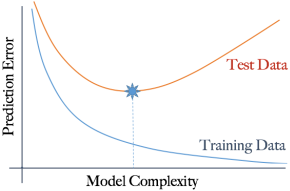
- As mentioned earlier, model selection based on the RSS or R^2 statistics can be misleading, since the training error is not a good representative of the actual test error



- Instead through a process of splitting the data into training and validations sets, we were able to use LOOCV or K-Fold CV as estimates of the test error

Why Cross Validation?

- As mentioned earlier, model selection based on the RSS or R^2 statistics can be misleading, since the training error is not a good representative of the actual test error



- Instead through a process of splitting the data into training and validations sets, we were able to use LOOCV or K-Fold CV as estimates of the test error
- We discussed why K-Fold CV is a more desirable estimate, computationally and statistically

Adjusting the Training Statistics for Test Error Approximation

Adjusting Techniques

– We introduce few other ways of adjusting the training error **to make it a better representative of the test error**

Adjusting Techniques

- We introduce few other ways of adjusting the training error **to make it a better representative of the test error**
- These adjustments are **not as reliable as Cross validation**, but they are easier to **calculate**

Adjusting Techniques

- We introduce few other ways of adjusting the training error **to make it a better representative of the test error**
- These adjustments are **not as reliable as Cross validation**, but they are easier to **calculate**
- These quantities were **more widely used before** the widespread use of computers for regression and machine learning

Adjusting Techniques

- We introduce few other ways of adjusting the training error **to make it a better representative of the test error**
- These adjustments are **not as reliable as Cross validation**, but they are easier to **calculate**
- These quantities were **more widely used before** the widespread use of computers for regression and machine learning
- Now that computers can help performing multiple fits computationally fast enough, often K-Fold CV is considered as the desirable test error approximation

List of Other Techniques

Methods to adjust the training error for the number of variables to estimate the test MSE:

- C_p statistic

List of Other Techniques

Methods to adjust the training error for the number of variables to estimate the test MSE:

- C_p statistic
- Akaike information criterion (AIC)

List of Other Techniques

Methods to adjust the training error for the number of variables to estimate the test MSE:

- C_p statistic
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

List of Other Techniques

Methods to adjust the training error for the number of variables to estimate the test MSE:

- C_p statistic
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Adjusted R^2

C_p Statistic

- For a fitted least squares model with d predictors

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

C_p Statistic

- For a fitted least squares model with d predictors

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- $\hat{\sigma}^2$ is an estimate of the noise variance

C_p Statistic

- For a fitted least squares model with d predictors

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)

C_p Statistic

- For a fitted least squares model with d predictors

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- It is an unbiased estimate of the **test MSE**

C_p Statistic

- For a fitted least squares model with d predictors

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- It is an unbiased estimate of the **test MSE**
- The smaller C_p, the better the model (we can pick models with the smallest C_p statistic)

C_p Statistic

- For a fitted least squares model with d predictors

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- It is an unbiased estimate of the **test MSE**
- The smaller C_p, the better the model (we can pick models with the smallest C_p statistic)
- Becomes a better estimate of the test error as the sample size, n, increases

AIC: Akaike Information Criterion

- Defined for a large class of models based on the maximum likelihood criterion

AIC: Akaike Information Criterion

- Defined for a large class of models based on the maximum likelihood criterion
- When we consider the noise ϵ be of i.i.d Gaussian, the MLE and MSE return identical results and in this case we have

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

which is a multiple of C_p (no preference over using one vs the other)

AIC: Akaike Information Criterion

- Defined for a large class of models based on the maximum likelihood criterion
- When we consider the noise ϵ be of i.i.d Gaussian, the MLE and MSE return identical results and in this case we have

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

  which is a multiple of C_p (no preference over using one vs the other)
- $\hat{\sigma}^2$ is an estimate of the noise variance

AIC: Akaike Information Criterion

- Defined for a large class of models based on the maximum likelihood criterion
- When we consider the noise ϵ be of i.i.d Gaussian, the MLE and MSE return identical results and in this case we have

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

  which is a multiple of C_p (no preference over using one vs the other)
- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)

AIC: Akaike Information Criterion

- Defined for a large class of models based on the maximum likelihood criterion
- When we consider the noise ϵ be of i.i.d Gaussian, the MLE and MSE return identical results and in this case we have

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

  which is a multiple of C_p (no preference over using one vs the other)
- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- The smaller AIC, the better the model (we can pick models with the smallest AIC statistic)

BIC: Bayesian Information Criterion

– Takes a Bayesian approach to estimate the test error

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \to \infty$) choosing the model with the highest posterior probability of being the best model

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \to \infty$) choosing the model with the highest posterior probability of being the best model
- In the case of least squares the formulation is

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

  which takes an *almost* similar form as the previous two statistics

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \to \infty$) choosing the model with the highest posterior probability of being the best model
- In the case of least squares the formulation is

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

  which takes an *almost* similar form as the previous two statistics
- $\hat{\sigma}^2$ is an estimate of the noise variance

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \to \infty$) choosing the model with the highest posterior probability of being the best model
- In the case of least squares the formulation is

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

  which takes an *almost* similar form as the previous two statistics
- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \to \infty$) choosing the model with the highest posterior probability of being the best model
- In the case of least squares the formulation is

  $$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

  which takes an *almost* similar form as the previous two statistics
- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- The smaller BIC, the better the model (we can pick models with the smallest AIC statistic)

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \rightarrow \infty$) choosing the model with the highest posterior probability of being the best model
- In the case of least squares the formulation is

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

  which takes an *almost* similar form as the previous two statistics
- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- The smaller BIC, the better the model (we can pick models with the smallest AIC statistic)
- When $n < 7$, BIC imposes a smaller penalty on the number of variables, but for $n > 7$ that $\log n > 2$ the penalty is larger

BIC: Bayesian Information Criterion

- Takes a Bayesian approach to estimate the test error
- Asymptotically ($n \to \infty$) choosing the model with the highest posterior probability of being the best model
- In the case of least squares the formulation is

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\hat{\sigma}^2)$$

  which takes an *almost* similar form as the previous two statistics
- $\hat{\sigma}^2$ is an estimate of the noise variance
- $\hat{\sigma}^2$ is normally estimated using all the predictors (full model)
- The smaller BIC, the better the model (we can pick models with the smallest AIC statistic)
- When $n < 7$, BIC imposes a smaller penalty on the number of variables, but for $n > 7$ that $\log n > 2$ the penalty is larger
- In other words in standard observation regimes where n is sufficiently large, BIC tends to pick smaller models than AIC or C_p

Adjusted R^2

- Presents a way of making the R^2 statistic dependent on the number of predictors

Adjusted R^2

- Presents a way of making the R^2 statistic dependent on the number of predictors
- Recall the R^2 statistic:

$$R^2 = 1 - \frac{RSS}{TSS}, \qquad \text{where} \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Adjusted R^2

- Presents a way of making the R^2 statistic dependent on the number of predictors
- Recall the R^2 statistic:

$$R^2 = 1 - \frac{RSS}{TSS}, \qquad \text{where} \quad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- The formulation for adjusted R^2 is

$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Adjusted R^2

- Presents a way of making the R^2 statistic dependent on the number of predictors
- Recall the R^2 statistic:

$$R^2 = 1 - \frac{RSS}{TSS}, \qquad \text{where} \quad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- The formulation for adjusted R^2 is

$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

- Unlike the other three statistics that being small indicates a better model, for adjusted R^2 we are interested in models that tend to generate values closer to 1

Adjusted R^2

- Presents a way of making the R^2 statistic dependent on the number of predictors
- Recall the R^2 statistic:

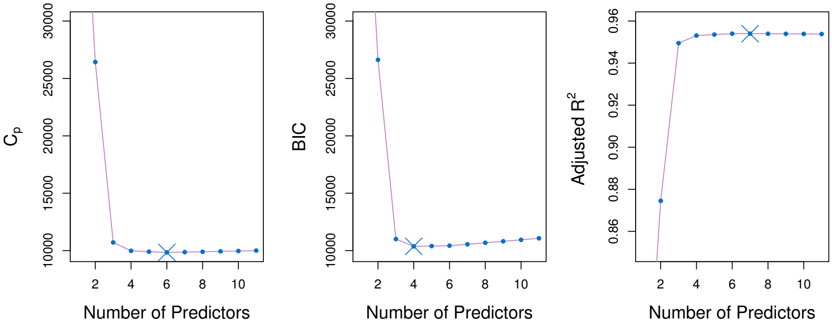
$$R^2 = 1 - \frac{RSS}{TSS}, \qquad \text{where} \quad TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- The formulation for adjusted R^2 is

$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

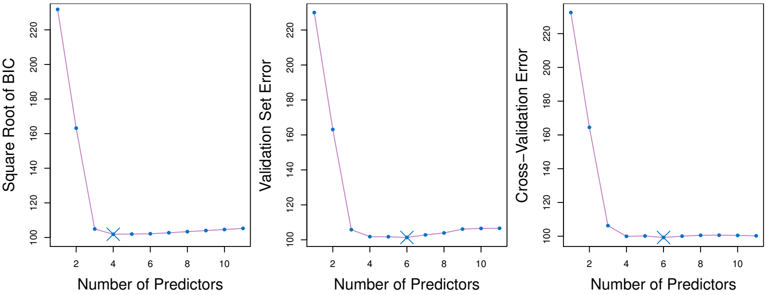
- Unlike the other three statistics that being small indicates a better model, for adjusted R^2 we are interested in models that tend to generate values closer to 1
- The use of C_p, AIC and BIC is more motivated in statistical learning theory than the adjusted R^2

Example Comparing the Performances



C_p, BIC, and adjusted R^2 for the best models of each size for the Credit data set

Comparison Against CV Techniques



- The results are not much different
- Note that nowadays CV methods are computationally fast to implement and regardless of the model can always be used as a reliable selection tool

How to Use These Statistics in Model Selection

- **Best subset selection** formal procedure (NP-hard and computationally not possible for large p)

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

How to Use These Statistics in Model Selection

- **Forward stepwise selection** (computationally tractable)

How to Use These Statistics in Model Selection

- **Forward stepwise selection** (computationally tractable)
- At each step the variable that gives the greatest additional improvement to the fit is added to the model

---

Algorithm 6.2 *Forward stepwise selection*

---

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

How to Use These Statistics in Model Selection

- **Forward stepwise selection** (computationally tractable)
- At each step the variable that gives the greatest additional improvement to the fit is added to the model

---

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- Forward selection can even be used when $n < p$

How to Use These Statistics in Model Selection

- **Backward stepwise selection** (computationally tractable)

How to Use These Statistics in Model Selection

- **Backward stepwise selection** (computationally tractable)
- Begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time

---

Algorithm 6.3 *Backward stepwise selection*

---

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

How to Use These Statistics in Model Selection

- **Backward stepwise selection** (computationally tractable)
- Begins with the full least squares model containing all p predictors, and then iteractively removes the least useful predictor, one-at-a-time

---

Algorithm 6.3 *Backward stepwise selection*

---

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- Backward selection requires $p < n$ (to allow the full model to be fit)

What are Shrinkage Methods and Why Useful?

You would probably hear **Ridge Regression** and **LASSO** quite often

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly **reduce the model variance**

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \cdots, \beta_p$ using the values that minimize

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$RSS_{Ridge} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

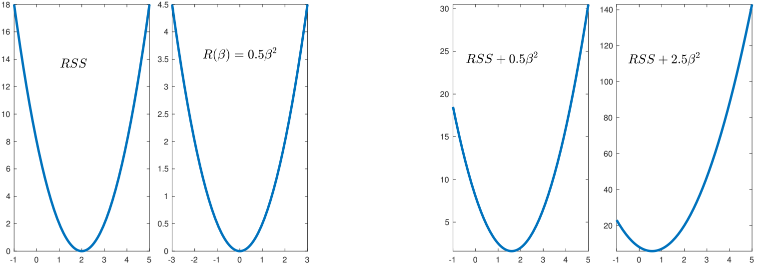
- Here, λ is a tuning parameter

Ridge Regression

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small
- However, the second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, called a shrinkage penalty, encourages solutions that are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates (trade off between bias and variance)
- Selecting a good value for λ is critical; often cross-validation is used for this

Effect of Increasing λ on the β

– The figure below shows how increasing the Ridge penalty pushes the minimizers of the mixed RSS objective to zero



Shrinkage Example

- Previously from the homework assignments you remember that the least squares solution to fit data point $(x_1, y_1), \cdots, (x_n, y_n)$ was obtained via the minimization:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta x_i)^2 \quad \therefore \quad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

- We can show that if we run the Ridge regression problem

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta x_i)^2 + \lambda \beta^2$$

  the new estimate becomes

$$\hat{\beta}^R = \frac{\sum_{i=1}^{n} x_i y_i}{\lambda + \sum_{i=1}^{n} x_i^2}$$

- Note how increasing λ pushes $\hat{\beta}^R$ towards zero

In Class Exercise

– For the simple regression problem of fitting $(x_1, y_1), \cdots, (x_n, y_n)$, to the model $y = \beta_0 + \beta_1 x$ show that the least-squares estimates for the Ridge regularized objective

$$\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 + \lambda(\beta_0^2 + \beta_1^2)$$

are

$$\hat{\beta}_1^R = \frac{\sum_{i=1}^n x_i y_i - \frac{n^2}{n+\lambda}\bar{x}\bar{y}}{\lambda + \sum_{i=1}^n x_i^2 + \frac{n^2}{n+\lambda}\bar{x}^2}, \quad \hat{\beta}_0^R = \frac{1}{n+\lambda}\left(\sum_{i=1}^n y_i - \hat{\beta}_1^R \sum_{i=1}^n x_i\right)$$

What Happens in Multiple Regression?

- In this case we previously had

$$RSS = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta)$$

  which led to

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

- In the case of regularized problem ($\|.\|$ denotes the L-2 norm)

$$(\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda \|\boldsymbol{\beta}\|^2$$

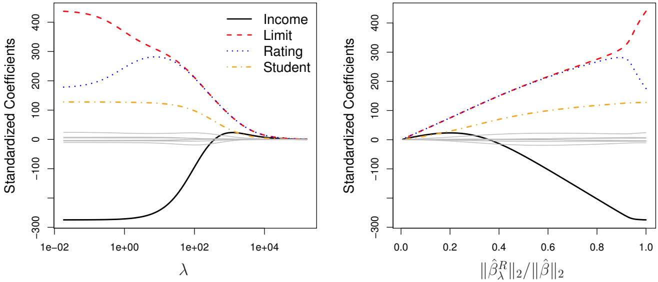
  we will have

$$\hat{\boldsymbol{\beta}}^R = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

  where $\boldsymbol{I}$ is the identity matrix

Credit Data Example

- Left: each curve corresponds to the ridge regression coefficient estimate for one of the ten variables, plotted as a function of λ
- The right-hand panel displays the same ridge coefficient estimates as the left-hand panel, but instead of displaying λ on the x-axis, we display $\|\hat{\beta}^R\|/\|\hat{\beta}\|$ (how much **shrinkage** happens by increasing λ)



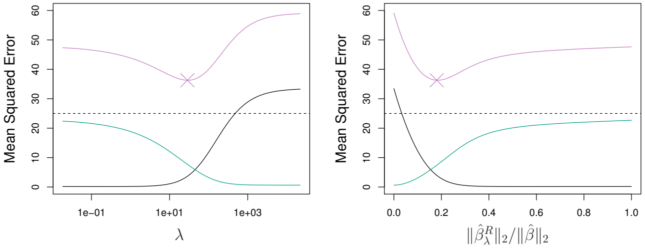
Scaling of the Predictors

- In the standard least-squares if we scale a feature value by c, the corresponding coefficient scales by c^{-1}
- However when we have the Ridge regularized objective, this is no more the case
- To see a consistent behavior, for the Ridge regularized problem we often work with standardized features:

$$\tilde{x}_{i,j} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Bias-Variance Trade-Off

- A toy example: squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R\|/\|\hat{\beta}\|$. The horizontal dashed lines indicate the minimum possible MSE (**the standard least squares,** $\lambda = 0$ **in nowhere close**). The purple crosses indicate smallest ridge regression model MSE values



- Remember (test error = bias + variance + noise variance)

Questions?

References

- https://www.alsharif.info/iom530, 2013.
- J. Friedman, T. Hastie, and R. Tibshirani.
  The elements of statistical learning.
  Springer series in statistics, 2nd edition, 2009.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/model_selection.pdf, 2013.
- G. James, D. Witten, T. Hastie, and R. Tibshirani.
  An introduction to statistical learning: with applications in R, volume 112.
  Springer, 2013.