Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

Introduction to Variable Selection
And Regularization

GTx

Lessons

A. Introduction
B. Best Subset Selection
C. Stepwise Selection
D. Choosing the Optimal Model
E. Ridge Regression
F. Lasso (Least Absolute Shrinkage and Selection Operator) Regression

GTx

Variable Selection and Regularization

- Thanks to advances in Information Technology, we are able to capture large volumes of data.
- We now have several domains where each observation of a subject/object has hundreds, if not thousands, of attributes.
- For example, in business applications, merging databases from different sources results in a plethora of attributes for an observation.
- A critical problem is to determine which attributes are most relevant for decision makers.

GTx

Variable Selection and Regularization

- In this module, we will study approaches that help with variable selection and regularization.
- We can then use our models for prediction (which is a very common application of analytics) .

GTx

MSE for training and test data from a statistical learning method

- $SSE = \sum_i (y_i - \hat{y}_i)^2$, so **Mean Squared Error**, ***MSE = SSE/n***
- $\hat{y}_i$ is also written as $\hat{f}(x_i)$, the prediction that $\hat{f}$ (the estimated model) gives for the *i*th observation, where *f* is the relationship between *y* and *x*.
- While OLS tries to minimize SSE on the training data, we are very often interested in how the estimated model performs on the test data.
- We want to find a method (or model) that gives the lowest test MSE as opposed to training MSE.
- There is no guarantee that a model with the best training MSE will also have the best test MSE.

GTx

Variance and Bias of a model

- **Variance** of a model refers to the amount by which $\hat{f}$ (the estimated model) would change if we estimated it using a different training data set
- If a method has high variance, then small changes in the training data could yield large changes in $\hat{f}$. Generally, more flexible models have higher variance since they follow the training data closely
- **Bias** is the error that is introduced by approximating a complicated real-life problem with a simpler model (for e.g., using a linear model to model a true non linear relationship)
- In fitting models for prediction, one has to consider the trade-off between **Variance** and **Bias**

GTx

Variance and Bias

- At any point x_0 in the test data, we can compute the Expected test MSE which is (proof not shown in this course) as the sum of
  - Variance of $\hat{f}(x_o))$
  - Squared Bias of $\hat{f}(x_o))$
  - Irreducible variance of the error terms, ε
- $E(y_o - \hat{f}(x_o))^2 = \text{Var}(\hat{f}(x_o))) + |\text{Bias}\ (\hat{f}(x_o)))|^2 + \text{Var}(\varepsilon)$
- The left hand side refers to the expected test MSE, and is equal to the average test MSE we would get if we obtain several $\hat{f}$ using numerous training sets and testing each at a test data observation x_0.
- The overall expected test MSE is calculated by computing the average value of MSE over all test points
- The challenge is to find a method for which both variance and squared bias are low

GTx

Prediction Accuracy

- In linear regression, the model

$$\boldsymbol{Y = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p + e}$$

  is typically fitted with least squares and we get estimates of the βs
- We want to develop alternative ways of fitting to get better prediction accuracy and interpretability of the model.
- If the relationship between *Y* and *Xs* is linear, then least square estimates have low bias.
- Also, if *n* (# of observations) is > *p* (# of predictors), then OLS will generally have low variance and it will perform well on test data.
- If *n* is not much larger than p, then there can be variability in the least squares fit, resulting in overfitting and poor predictions on test data.
- If *p* > *n*, then least squares method can't be used.

GTx

Regression…

- If we can constrain or shrink the estimated coefficients, we can often reduce variance while not increasing bias significantly.
- This approach can improve accuracy of predictions on test data.
- In a lot of scenarios, many of the variables used in multiple regression are not relevant. So, if we can remove these variables, we can get a model that is better for interpretation.
- Approaches to remove non-relevant variables are also called **feature selection** or **variable selection**.

GTx

Alternatives to Using Least Squares

- **Subset Selection.** This involves finding a subset of the *p* predictor (or explanatory) variables that are best connected (or related) to the *Y* variable. We then use this subset of variables least squares regression to fit a model.

- **Shrinkage (Or Regularization)**. We fit a model with all *p* predictor variables but then shrink the coefficients to nearly 0 compared to least square values. Variance is reduced. If Lasso is used, then some coefficients move to 0, hence leading to variable selection.

- **Dimension Reduction**. This involves finding *M* (with $M < p$) linear combinations of the predictors. Then fit linear regression on these *M* new variables. (We won't study this topic in this course).

GTx

Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

Best Subset Selection

GTx

Best Subset Selection

- If there are p predictors, one could fit a separate OLS regression for each one of the possible combinations of the p predictors.
- Note that with p predictors, there are 2^p separate regressions possible
  - If $p = 2$, there are 4 regressions
  - If $p = 10$, there are 1024 regressions
  - If $p = 20$, there are 1048576 regressions
  - If $p = 30$, more than a billion regressions, …
- Also note that with p predictors, there are $\binom{p}{s}$ models that have exactly s predictors. The idea is, that for each value of s, pick the "best" (smallest SSE) among the regression models that have s predictors. Call this model M_s
- Since s ranges from 0 to p, we have $p + 1$ models, labelled M_0 to M_p
- We want to select a model that has a low error rate on the test data. Therefore, pick the best among all these $p + 1$ models but using cross-validation prediction error, adjusted R^2, or C_p, AIC, or BIC

GTx

Best Subset Selection

- Subset selection is also applicable to other models such as logistic regression.
- Instead of minimizing SSE, we use **deviance**, as the measure to be minimized

deviance = -2 * maximized log-likelihood

- Note that we want to minimize deviance as smaller values imply a better fit
- A big limitation of subset selection is computational limitations. With p = 30, there are over a billion possibilities!
- Subset selection is not practical with p over 30! Instead, we have to resort to alternative efficient methods.

GTx

Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

Stepwise Selection

GTx

Stepwise Selection

- As we saw in the previous lesson, best subset selection cannot be practically applied to problems with large p (say over 30).
- It also could have overfitting problems (on the training data) if *p* is very large. This means high variance of the coefficient estimates.
- Stepwise methods – Forward and Backward – are alternative approaches that involve a considerably smaller number of models.

GTx

Forward Stepwise Selection

- The steps start at $k = 0$ with M_0 the model with no explanatory variables and then go sequentially to $k = p - 1$
- For each value of *k*
  - Consider all of the $p - k$ models that can add one predictor each to M_k
  - Select the best of these $p - k$ models (having the best SSE or highest R^2 for example) and call the new model M_{k+1}
- We then select the best model from $M_0, \ldots, M_p$ that has a low error rate on the test data using cross-validation prediction error, adjusted R^2, or C_p, AIC, or BIC

*** Adapted from Algorithm 6.2 in ISLR

GTx

Forward Stepwise Selection

- The number of models that are considered is far fewer than for the best subset selection approach.
- We have the null model and for each of the *k* steps we consider (*p* – *k*) models.
- This yields a total of $1 + \sum_{k=0}^{p-1}(p - k)$ models, which is equal to *1 + p(p + 1)/2* models being considered.
- This is far less than the 2^p models being considered in best subset selection.
- For forward stepwise selection
  - If p = 2, there are 4 regressions to be considered
  - If p = 10, there are 56 regressions
  - If p = 20, there are 211 regressions
  - If p = 30, there are 466 regressions
  - If p = 40, there are 821 regressions, etc.
- Also, once a variable is part of M_k, it can't be removed from M_{k+1} onwards
- However, there is no guarantee that forward stepwise selection will find the "best" model.

GTx

Backward Stepwise Selection

- This is also an efficient alternative to best subset selection.
- It starts with the model that has all *p* predictors and then removes the least useful predictor one at a time.
- Let M_p be the full model with all *p* predictors (explanatory) variables.
- Set *k = p, p* -1,…,1
  - Consider all of the *k* models (with k -1 predictors) that have one fewer predictor than in M_k
  - Select the best of these *k* models (having the best SSE or highest R^2 for example) and call the new model M_{k-1}
- We select the best model from $M_0, \ldots, M_p$ that has a low error rate on the test data using cross-validation prediction error, Cp, adjusted R^2, AIC, or BIC

*** Adapted from Algorithm 6.3 in ISLR

GTx

Backward Stepwise Selection

- The number of models that are considered is equal to $1 + p(p + 1)/2$ models (which is the same as forward stepwise selection)
- When p is very large, best subset selection isn't computationally viable. Hence, backward stepwise selection can be applied.
- This does require that the number of observations, n, is larger than the number of variables, p.
- If $n < p$, forward stepwise selection can still be useful to some extent
- There is no guarantee that backward stepwise selection will find the "best" model.

GTx

Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

Choosing the Optimal Model

GTx

Optimal Model Selection

- Best Subset, Forward Selection, and Backward Selection each result in a set of models with a subset of the *p* predictor variables.
- We are interested in finding a model that has a low test error!
- Note that obtaining smallest SSE or R^2 for a model on the training data set is not helpful for prediction purposes.
- There are two ways to select a good model as far as test error is concerned:
  1. Make adjustments to the training error to deal with bias from overfitting
  2. Directly estimate test error with validation set or cross-validation methods (discussed in another module)
- Note that the training set Mean Squared Error MSE = SSE/*n*
- This is typically an underestimate of the test set MSE.

GTx

Adjustments to Training Error Using C_p, AIC, BIC, or adjusted R^2

- One approach to adjust training error is C_p
- If a fitted least squares model has *d* predictors, the C_p estimate for the test MSE is

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

  where $\hat{\sigma}^2$ is an estimate of the variance of the error for a model using all predictors.
- So $2d\hat{\sigma}^2$ is a penalty added to the training SSE.
- For models with a low test error C_p is usually small.
- So given a set of models to consider, we choose the model which has the lowest C_p value.

GTx

C_p, AIC, BIC, Adjusted R^2

- When models are fitted with maximum likelihood, then the **Akaike information criterion** (**AIC**) is used, where

  AIC $\approx \frac{1}{n\hat{\sigma}^2}(\text{SSE} + 2d\hat{\sigma}^2)$, which is similar to C_p
- **Bayesian information criterion** (**BIC**) is from a Bayesian perspective but is similar to C_p and AIC,

  BIC$= \frac{1}{n\hat{\sigma}^2}(\text{SSE} + \log(n)d\hat{\sigma}^2)$
- BIC usually places a heavier penalty than AIC for models with large p since log(n) > 2 when n > 7
- So given a set of models to consider, we choose the model which has the lowest AIC or BIC value.
- Adjusted R^2 has been discussed in the module on linear regression.

GTx

Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

Ridge Regression

GTx

Ridge Regression

- The idea behind Ridge Regression and LASSO is to fit a model will all *p* predictors using a method that constrains or regularizes the estimated coefficients or shrinks them towards zero. Doing so reduces the variance.
- In OLS, we try to find estimates of the coefficients β_i minimize

$$\boldsymbol{SSE} = \sum_{\boldsymbol{i}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}} =$$

$$\sum_{\boldsymbol{i}} (\boldsymbol{y_i} - (\boldsymbol{\beta_0} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{\beta_j x_{ij}}))^{\boldsymbol{2}}$$

Ridge Regression

- In Ridge Regression we try to find estimate of βs that minimize

  $\sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2$ = SSE + $\lambda \sum_{j=1}^{p} \beta_j^2$
- $\lambda \geq 0$ is a **tuning parameter**, which has to be determined separately (by cross-validation)
- The second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, is called a **shrinkage penalty**
- This penalty is small when $\beta_1, \ldots, \beta_p$ are close to 0
- If $\lambda = 0$, the ridge regression produces least square estimates
- As λ becomes very large, the penalty impact grows and the ridge regression coefficients go to zero
- Note that each value of λ in ridge regression will get a different set of $\beta_1, \ldots, \beta_p$ coefficient estimates

Ridge Regression

- Note that the penalty does not apply to the intercept β_0
- It is best to apply Ridge Regression after standardizing the predictor variables (i.e., dividing each predictor variable by its estimated standard deviation) so that all βs are on the same scale
- The Bias-Variance tradeoff comes into play as you vary λ
- The test **mean squared error** (**MSE**), is a function of the variance plus the squared bias.
- When $\lambda = 0$, the variance is high but there is no bias
- As λ increases, the "shrinkage" of the coefficients leads to a reduction in variance of the predictions, but with some slight bias increase.

GTx

Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

Lasso Regression

GTx

Lasso

- Each of Subset Selection, Forward Stepwise, and Backward Stepwise Selection will generally output a model with a subset of the predictor variables.
- However, in Ridge Regression while some βs may shrink towards zero, none of them are actually set to 0.
- This could be a handicap when it comes to **model interpretability** when there are a large number of predictor variables.
- The **Lasso** is an alternative to Ridge Regression that does not have this problem.

GTx

Lasso

- The Lasso coefficients minimize

  $\sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}))^2 + \lambda\sum_{j=1}^{p}|\beta_j|$ = SSE + $\lambda\sum_{j=1}^{p}|\beta_j|$
- This is very similar to the Ridge Regression with the change in the penalty function from β_j^2 to $|\beta_j|$
- $\lambda \geq 0$ is a **tuning parameter**, which has to be determined separately (by cross-validation) and has similar effects as ridge regression.
- The second term, $\lambda\sum_{j=1}^{p}|\beta_j|$, is called a **shrinkage penalty**
- If $\lambda = 0$, the lasso produces least square estimates.
- As λ becomes very large, the penalty impact grows and the lasso regression coefficients are set to zero.
- Each value of λ in lasso will get a different set of $\beta_1, \ldots, \beta_p$ coefficient estimates.

GTx

Comparing Ridge Regression and Lasso

- Lasso has a very useful advantage as it produces simpler models which are more interpretable.
- Note that the true relationship between *Y* and the *X* variables will determine which method has the minimum MSE.
- Neither one will dominate the other over all data sets. It will require techniques such as cross-validation to figure out which method is more useful for a data set.
- If only say a few out of a large set of *p* predictors are related to the *Y* variable, lasso will do better than Ridge Regression.

GTx

Selecting the λ Tuning Parameter

- Generally choose a range of values for λ
- Compute the cross-validation (cv) error for each value of λ
- Then pick the value of λ for which the cv error is minimum.
- Then refit the model with this value of λ and using all the observations.

GTx

Related R Packages

- The **leaps** package and its regsubsets() function for best subset selection in linear regression.
- regsubsets() can also be used to do Forward Stepwise Selection as well as Backward Stepwise Selection.
- The **glmnet** package is used to perform Ridge Regression and the Lasso.
- Section 6.5 of ISLR demonstrates some examples that use these two packages.

GTx

Recap of Lessons on Variable Selection and Regularization

A. Introduction
B. Best Subset Selection
C. Stepwise Selection
D. Choosing the Optimal Model
E. Ridge Regression
F. Lasso (Least Absolute Shrinkage and Selection Operator) Regression

GTx

Data Analytics in Business

Variable Selection & Regularization

Sridhar Narasimhan
Professor
Scheller College of Business

End