MACHINE LEARNING - I

UNIT # 4

1

TODAY'S AGENDA

- Recap of the previous lecture(s)
- Cross-Validation (Cont'd)
- Shrinkage Method: Ridge and Lasso
- Elastic Net
- Regression Tree

2

RECAP OF THE PREVIOUS LECTURES

- Regression Problem:
  - Y as a function of X1, X2, …., Xp
- Ordinary Least Square
  - Y = b0 + b1X1 + … + bpXp + e
- Evaluation metric(s) to assess fitted model quality
  - Residence Standard Error or Root Mean Squared Error (Same thing)
  - R2

RECAP (CONT'D)

- Polynomial Models
  - Y = b0 + b1X1 + b2X1^2 + …… + e (how many coefficients?)
- Interaction Models
  - Y = b0 + b1X1 + b2X1^2 + b3X2 + b4X2^2 + b5X1X2 + …… + e (how many coefficients?)
- Categorical Variables
  - Dummy vs one-hot encoding
- Model Assessment (Cont'd)
  - Train-Test split (Holdout or Validation method)
  - Cross Validation

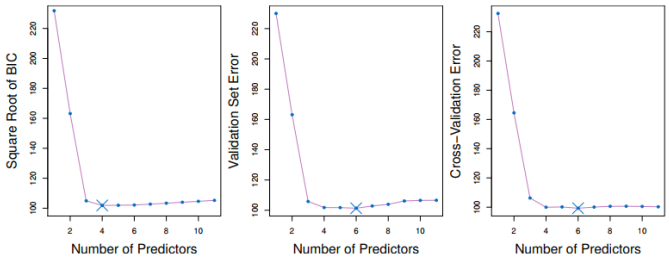
RECAP (CONT'D)

- Feature Selection Filter:
  - Correlation Filter
  - Variance Filter
  - Missing Value Filter
- Feature Selection Wrapper:
  - Best Subset (Practically Impossible)
  - Forward Selection
  - Backward Selection
- Model Assessment (Cont'd)
  - R2-Adj
  - AIC
  - BIC
- Shrinkage Method
  - Ridge
  - Lasso

VALIDATION AND CROSS VALIDATION

- In the past, performing cross-validation was computationally prohibitive for many problems with large p and/or large n, and so AIC, BIC, C_p, and adjusted R2 were more attractive approaches for choosing among a set of models.
- However, nowadays with fast computers, the computations required to perform cross-validation are hardly ever an issue. Thus, cross validation is a very attractive approach for selecting from among a number of models under consideration.

CROSS VALIDATION (CONT'D)



- Three quantities are displayed for the best model containing d predictors, for d ranging from 1 to 11.
- The overall best model, based on each of these quantities, is shown as a blue cross.
- One rule of thumb is to first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

SHRINKAGE METHOD

- 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 their variance.
- The two best-known techniques for shrinking the regression coefficients towards zero are ridge regression and the lasso.

RIDGE REGRESSION

- The least squares fitting procedure estimates β0, β1, …, βp using the values that minimize

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

- Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. In particular, the ridge regression coefficient estimates β^R are the values that minimize

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

- where λ ≥ 0 is a tuning parameter, to be determined separately.

RIDGE REGRESSION (CONT'D)

- The second term, $\lambda \sum_j \beta_j^2$, called a shrinkage penalty, is small when β1, …, βp are close to zero, and so it has the effect of shrinking penalty the estimates of βj towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- When λ = 0, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- However, as $\lambda \rightarrow \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.

RIDGE REGRESSION (CONT'D)

- Note that the shrinkage penalty is applied to $\beta 1, \ldots, \beta p$, but not to the intercept $\beta 0$.
- We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $xi1 = xi2 = \ldots = xip = 0$.

RIDGE REGRESSION – IMPACT OF SCALE

- The OLS coefficient estimates are scale equivariant: multiplying Xj by a constant c simply leads to a scale scaling of the least squares coefficient estimates by a factor of 1/c.
- Thus, regardless of how the jth predictor is scaled, $X_j\beta_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant.
- For instance, consider the income variable, which is measured in dollars. One could reasonably have measured income in thousands of dollars, which would result in a reduction in the observed values of income by a factor of 1,000.
- Now due to the sum of squared coefficients term in the ridge regression formulation, such a change in scale will not simply cause the ridge regression coefficient estimate for income to change by a factor of 1,000.

RIDGE REGRESSION – IMPACT OF SCALE (CONT'D)

- In other words, $X_j\beta_{j,\lambda}$ will depend not only on the value of λ, but also on the scaling of the jth predictor.
- In fact, the value of $X_j\beta_{j,\lambda}$ may even depend on the scaling of the other predictors!
- Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula so that they are all on the same scale.

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

LASSO REGRESSION

- Unlike best subset, forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model.
- The penalty $\lambda \sum \beta_j^2$ will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).
- This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables p is quite large.
- The lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage.

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

LASSO REGRESSION (CONT'D)

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the lasso performs variable selection.
- As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.

LASSO VS RIDGE

- Neither ridge regression nor the lasso will universally dominate the other. In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.
- However, the number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

SELECTING THE TUNING PARAMETER

- Implementing ridge regression and the lasso requires a method for selecting a value for the tuning parameter λ.
- Cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error for each value of λ.
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

PYTHON DEMO OF RIDGE AND LASSO WITH CV

ELASTIC NET

- An L2 penalty minimizes the size of all coefficients, although it prevents any coefficients from being removed from the model.
- l2_penalty = sum j=0 to p beta_j^2
- An L1 penalty minimizes the size of all coefficients and allows some coefficients to be minimized to the value zero, which removes the predictor from the model.
- l1_penalty = sum j=0 to p abs(beta_j)
- Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training.

ELASTIC NET (CONT'D)

- a hyperparameter "alpha" is provided to assign how much weight is given to each of the L1 and L2 penalties. Alpha is a value between 0 and 1 and is used to weight the contribution of the L1 penalty and one minus the alpha value is used to weight the L2 penalty.
- elastic_net_penalty = (alpha * l1_penalty) + ((1 – alpha) * l2_penalty)
- For example, an alpha of 0.5 would provide a 50 percent contribution of each penalty to the loss function. An alpha value of 0 gives all weight to the L2 penalty and a value of 1 gives all weight to the L1 penalty.
- elastic_net_loss = loss + (lambda * elastic_net_penalty)

REGRESSION TREE

- Regression Tree (and Classification Tree) is a flexible data-driven method.
- It is transparent and easy to interpret.
- Trees are based on separating records into subgroups by creating splits on predictors. These splits create logical rules.
- As with other data-driven methods, trees require large amounts of data.

21

EXAMPLE 1

- https://sefiks.com/2018/08/28/a-step-by-step-regression-decision-tree-example/

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

22

EXAMPLE 1 (CONT'D)