Statistics 311: Modern Statistical Methods
Final Project
Emily Duncan
41762105

Emily Duncan
41762105

Stat 311
Final Project

5 lines on each, no formulas for Bootstrap and Cross-Validation:

Bootstrap

- Used to quantify the uncertainty associated with a given estimator
- Estimates the standard errors of the coefficients from a linear regression fit.
- Able to use a computer to emulate the process of obtaining new sample sets, we obtain distinct data sets by repeatedly sampling observations from the original data set.

Cross-Validation

- Validation Set Approach: model fit on training set ⇒predicts the responses for observations in the held-outset, gives an estimate of test error rate
- LOOCV: single observation, unbiased estimate for test error
- K-fold CV: randomly divides the set of observations in k-folds (1^{st} fold is validation)

Ancient Selection Methods, 5 lines each + code:

Forwards Stepwise Selection

- Variables once entered may be dropped if they're no longer significant as other variables are added.
- Variable is significant if increases the R-squared value or decrease the BIC value

Backwards Stepwise Selection

- Start by fitting a model with all variables, then the least significant is dropped
- Continue by successively refitting reduced models and dropping variables until all remaining variables are statistically significant.

Chapter 6: Shrinkage Methods

Ridge Regression

- **Type of response variable**: quantitative as we are making adjustments to a linear model of fit.
- **Type of predictors:** all p predictors
- **What it does:**
  - Shrinkage method similar to least squares method, coefficients are shrunken towards zero relative to the least squares estimates
  - Used to prevent overfitting, we want to shrink the estimated associated of each variable with the response
- **How it works:**
  - Uses λ as a tuning parameter
  - Like least squares, seeks coefficient estimates that fit the data well (by making the RSS small)
  - The tuning parameter controls the shrinkage of estimates.
    - When λ is 0, there is no shrinkage and ridge regression produces the least squares estimates
    - As λ increases the impact of the shrinking grows
  - For each different λ, ridge regression produces different sets of coefficients estimates.
  - We use cross-validation to find λ
- **Does it do inference:** No.
- **Prediction:** While the model fits data well, ridge regression does not perform variable selection which means we do not know which predictors are significant. The model has poor interpretability.
- **Pros of Ridge Regression:** Good fit of model (prediction accuracy), computationally not difficult like stepwise selection.
- **Cons of Ridge Regression:** No model interpretability (no variable selection), shrinking on the coefficients causes them to significantly underestimated which leads to an increase in bias.

The Lasso

- **Type of response variable:** quantitative
- **Type of predictors:** Starts with all p predictors but then forces some to 0.
- **What it does:**
  - Recent alternative to Ridge Regression
  - Shrinks coefficients, but the penalty can force some estimates to be exactly zero when the tuning parameter is large
- **How it works:**
  - Like Ridge Regression uses λ as the tuning parameter
  - When λ is sufficiently large, the lasso can give the null model where all coefficients are equal to 0.

    - Depending on the value of λ, the lasso can produce a model involving any number of predictors
    - We also use cross-validation to find λ.
- **Does it do inference:** So far no, but researchers are working on this.
- **Prediction:** The lasso performs variable selection since it is able to set coefficients of predictors to 0.
- **Pros of The Lasso:** Able to set the coefficients to exactly 0 which allows model interpretability, requires only one model fitting, computationally easier than subset selection
- **Cons of The Lasso:** Less accurate at predicting new data (since variables have been removed)
- **Comparison with Ridge Regression:**
    - The Lasso is able to perform variable selection unlike Ridge Regression
    - Ridge Regression can perform better when the response is a function of many predictors
    - Lasso generally has a simpler and more interpretable model
    - Ridge Regression has slightly lower variance (in most cases) and the minimum mean squared error is slightly smaller
    - Ridge Regression and the Lasso often have identical biases
    - **The Lasso and Ridge Regression often have very similar results.**

Chapter 8: Tree-Based Methods

<u>Trees</u>

- **Type of response variable:** qualitative or quantitative depending on regression trees or classification trees.
- **Type of predictors:** All p predictors, but the predictor space is stratified or segmented into a number of simple regions
- **What it does:**
    - Offers a different approach to other statistical learning methods, these tree-based methods are simple and useful for interpretation
- **How it works:**
    - *Regression trees:* illustrates a qualitative response
        - *How to build a regression tree:*
            - Divide the predictor space
            - For every observation that falls into a specific region R_k, we make the same prediction
                - Which is the mean of the response values for the training observations in R_k.
        - The predicted response for an observation is given by the mean response of the training observations that belong to the same terminal node (second step in building a regression tree summarized)

    - A smaller, pruned tree can lead to lower variance and a better interpretation, an unpruned tree overfits the data
    - Basically, we want to build a big tree and then prune it
        - We use cross-validation or validation set approach to estimate error to prune the tree.
    - *Classification trees:*
        - Predicts a quantitative response
        - Predict that each observation belongs to the most commonly occurring class of training observations in the region
        - Classification error rate is simply the fraction of the training observation in that region that do not belong to the most common class
        - Either Gini index or cross-entropy are used to evaluate the quality of a split
- **Does it do inference:** No.
- **Prediction:**
    - *For regression trees:* To see predictive quality we consider all predictors, then choose the predictor and cutpoint so that the resulting tree has the lowest RSS (so the ones that aren't pruned are significant)
    - *For classification trees:* Predicative quality is based on each observation belonging to the most commonly occurring class of training observations in the region.
- **Pros of Trees:** Can be easily displayed graphically and easily interpreted (from a picture), can handle qualitative predictors without dummy variables
- **Cons of Trees:** Generally, trees do not have the same level of predictive accuracy as other methods.

<u>Bagging</u>

- **Type of response variable:** quantitative or qualitative
- **Type of predictors:** all p predictors, until pruned.
- **What it does:**
    - General purpose procedure for reducing the variance of a method, it is used frequently with trees
    - Uses the bootstrap from Chapter 5 to improve trees.
- **How it works:**
    - *For Regression Trees:* We construct B regression trees using B bootstrapped training sets and averaging the resulting predictions
        - Averaging the predictions reduces the variance, because it is combining lots of trees into one single tree.
    - *For Classification Trees:* We record the class predicted by each of the B trees and take a majority vote.
        - Majority vote: The overall prediction is the most commonly occurring class among the B predictors
    - Both use out-of-bag error estimation instead of cross-validation.

        - Regression uses OOB MSE
        - Classification uses the classification error, OOB error is valid because the response for each observation is predicted by only using trees that were not fit using that observation (classification error above in Trees).
        - OOB is computationally easier than cross-validation for these large data sets.
- **Does it do inference:** No.
- **Prediction:** Can record total amount that the RSS decreased due to splits over a predictor (averaging over all B trees), here a large decrease indicates an important predictor. For classification trees, a similar approach with the decreased Gini index.
- **Pros of Bagging:** Improves prediction accuracy, reduces variance
- **Cons of Bagging:** Difficult to interpret bagged model

Random Forests

- **Type of response variable:** Quantitative or qualitative
- **Type of predictors:** m subset of all p predictors
- **What it does:**
    - Random forests force each split to consider m predictors instead of all p predictors.
    - Many splits will not consider a strong predictor and there will be less correlation between predictors.
- **How it works:**
    - Uses a random sample of m predictors (which is a subset of all p predictors) for split candidates.
    - The split is only allowed to use one of these m predictors, at each split there is a new m sample.
- **Does it do inference:** No.
- **Prediction:** Improves predictive accuracy by generating large number of bootstrapped trees based on these random samples and deciding the outcome by averaging in regression trees and majority vote in classification.
- **Pros of Random Forests:** Average of the trees less variable and more reliable, decorrelates the trees, less overfittitng.
- **Cons of Random Forests:** Difficult for interpretability in some cases,