2019.05.08

Data Mining

– Chapter8. Tree Based Models–

182STG21

임소현

1. Description

2. Implement

Lab : Linear Regression

- 트리 모델을 사용하기 위해 library tree 에서 tree() 함수를 사용한다. Lm 과 비슷한 방법으로 tree(y~x)로 표현한다. Summary(model)를 사용하면 terminal node 의 개수, 오분류율을 알려준다. plot(model)은 tree 모델을 시각화하여 알려주고 pretty=0 은 범주형 자료에 대해 각 범주를 포함하라는 옵션이다.

- cv.tree() func 은 최적의 tree 복잡도를 결정하기 위해 CV 를 실시한다. FUN=prune.misclass 는 classification error 를 원할 때 사용한다. 여기서 size 는 terminal nodes 의 수(size), 그에 따른 cv-error(dev), cost-complexity parameter 의 값(여기서는 k)을 출력한다.

- regression tree 에서는 deviance 가 tree 에 대한 sum of squared errors 이다. prune 을 할 때, classification tree 에서는 prune.misclass()이었지만 regression tree 에서는 prune.tree()을 사용한다.

- bagging 은 randomforest 의 특별한 경우다. 즉, bagging 은 m=k 인 경우다.따라서 randomForest()로 bagging, randomforest 모두를 시행할 수 있다. Importance 를 통해 변수별 중요도도 확인할 수 있다.

- randomForest()은 분류 문제에 대해서 root(p)개의 변수를, regression 문제에 대해서는 p/3 개의 변수를 사용한다.

- boosting model 을 적용하기 위해 gbm package 의 gbm()을 이용한다.

- modeling 후 중요한 변수에 대한 partial dependence plots 를 그릴 수 있다. 이 plot 은 다른 변수를 integrate out 한 후에, 반응변수에 대한 선택된 변수의 marginal effect 을 보여준다.

8. In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.
(a) Split the data set into a training set and a test set.
(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test error rate do you obtain?
(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?
(d) Use the bagging approach in order to analyze this data. What test error rate do you obtain? Use the importance() function to determine which variables are most important.
(e) Use random forests to analyze this data. What test error rate do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of *m*, the number of variables considered at each split, on the error rate obtained.

9. This problem involves the OJ data set which is part of the ISLR package.
(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
(b) Fit a tree to the training data, with Purchase as the response and the other variables except for Buy as predictors.

Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?
(c) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.
(d) Create a plot of the tree, and interpret the results.
(e) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
(f) Apply the cv.tree() function to the training set in order to determine the optimal tree size.
(g) Produce a plot with tree size on the *x*-axis and cross-validated classification error rate on the *y*-axis.
(h) Which tree size corresponds to the lowest cross-validated classification error rate?
(i) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.
(j) Compare the training error rates between the pruned and unpruned trees. Which is higher?
(k) Compare the test error rates between the pruned and unpruned trees. Which is higher?

10. We now use boosting to predict Salary in the Hitters data set.
(a) Remove the observations for whom the salary information is unknown, and then log-transform the salaries.
(b) Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.
(c) Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the *x*-axis and the corresponding training set MSE on the *y*-axis.
(d) Produce a plot with different shrinkage values on the *x*-axis and the corresponding test set MSE on the *y*-axis.
(e) Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapters 3 and 6.
(f) Which variables appear to be the most important predictors in the boosted model?
(g) Now apply bagging to the training set. What is the test set MSE for this approach?

3. Discussion

Appendix