Bagging, random forest and boosting

Lecture

In the previous lecture, you are introduced to CART, a tree-based data mining method. In the lecture, we discussed the algorithm of CART and applied CART to the OJ purchase example in the tutorial. Trees are popular because they are user friendly, however CARTs are known for the poor predictive power and can be unstable. Therefore, many methods have since been proposed to improve the predictive power of CARTS. In MA3405, we will explore bagging, random forest and boosting

Textbook Reading

8 Tree-Based Methods 303

8.2 Bagging, Random Forests, Boosting 316
8.2.1 Bagging 316
8.2.2 Random Forests 319
8.2.3 Boosting 321

Tutorial

Exercise Let's talk

In a group of two, discuss the followings

- What is the black box machine learning method?
- What is wisdom of crowds?
- Explain the bagging algorithm (including how predictions are made)
- When is the bagging unreliable? Why?
- Explain the algorithm of random forest (including how predictions are made)
- under what condition, is the bagging the same as the random forest?
- Explain Adaboost algorithm
- Explain the algorithm of boost regression tree
- How does bagging, random forest and boosting derive variable importance rankings?

Chapter 8.4, Exercise 2 The boosting using depth-one trees (or stumps) leads to an additive model; i.e. a model of the form

$$f(X) = \sum_{j=1}^{p} f_j(X)$$

Explain why. (You can begin with Algorithm 8.2)

Chapter 8.4, Exercise 8 Predict Car seat sales.

The `Carseats` is a simulated car seat sale at 400 store locations. Predict Carseats sales using the available predictors.

1. Split the data set into a training set and a test set.
2. Fit a regression tree to the training set. Plot the tree, and interpret the results. What test error rate do you obtain?
3. Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?
4. 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
5. 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
6. 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.
7. Produce a plot with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis.
8. Which variables appear to be the most important predictors in the boosted model?
9. Which of the three methods is best suited this data?