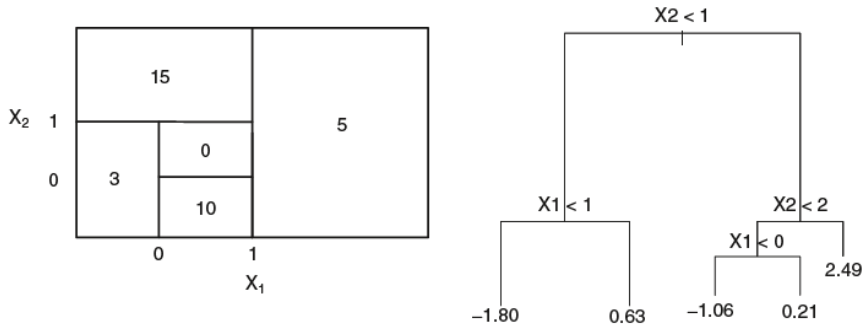
Data Mining HW 3

Due: 2022.05.23 24:00

Exercises for Tree



1. 위의 그림을 보고, 아래의 물음에 답하여라.

   (a) 왼쪽 그림을 보고, 오른쪽과 같은 tree를 그려라. 왼쪽 그림의 박스안의 숫자는 각 영역 내의 Y의 평균이다.

   (b) 오른쪽 그림을 보고 왼쪽과 유사한 다이어그램을 생성하여라. 예측 변수 공간을 올바른 영역으로 나누고, 각 영역의 평균을 표시해야 함.

2. Tree 강의노트 p.22의 데이터를 이용하여, 불순도 측도로 Gini 지수를 사용했을 때, 첫번째 분리 규칙을 찾아라.

3. 'Carseats.csv' 데이터를 이용하여 Sales(매출액)을 예측하려고 한다. 다음 물음에 답하여라.

   (a) 데이터를 training data (60%)와 testing data(40%)로 나누어라.

   (b) training data를 이용하여 회귀트리를 적합하여라. 그림을 그리고, 결과를 설명하여라. tets MSE는 얼마인가?

   (c) tree의 complexity를 고려하여 가지치기를 시행하여라. 가지치기 시행으로 test MSE는 향상되었는가?

   (d) Bagging을 시행하여라. tets MSE는 얼마인가?

   (e) Random Forest를 시행하여라. tets MSE는 얼마인가? 변수중요도를 수하여라. $m = 1, 10$을 포함하여 m값을 변화시키면서 tets MSE를 비교하고, 이 때 m의 효과를 설명하여라.

   - 변수설명 : ISLR2 패키지의 Carseats 데이터 설명 참고

4. ‘OJ.csv’ 데이터를 이용하여, 다음 물음에 답하여라.

   (a) 800개의 관측값으로 구성된 training data를 만들고, 나머지를 포함하는 testing data를 구성하여라.

   (b) training data를 이용하여 반응변수를 Purchase로 하는 tree를 적합하여라. 결과를 설명하여라. training error는 얼마인가? terminal node의 갯수는 몇개인가?

   (c) tree의 그림을 그리고 결과를 설명하여라.

   (d) testing data를 이용하여 예측을 하고, confusion matrix를 생성하여라. test error은 얼마인가?

   (e) cptable 및 그림을 이용하여 최적의 tree를 구하여라.

   (f) 가지치기 전/후 tree의 training error를 비교하여라.

   (g) 가지치기 전/후 tree의 test error를 비교하여라.

   - 변수설명 : ISLR2 패키지의 OJ 데이터 설명 참고