

Multivariate Data Analysis Assignment #6

Decision Tree for Classification

Dataset: Heart Disease UCI

(<https://www.kaggle.com/ronitf/heart-disease-uci>)

해당 데이터셋은 총 303명 환자의 현재 의료 정보를 바탕으로 심장병을 진단하는 목적의 데이터셋이다. 총 14개의 변수로 구성되어 있으며 가장 마지막 column인 target이 1이면 실제 심장병 발병 환자이고 0이면 발병하지 않은 환자이다.

Dataset

Heart Disease UCI

<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

ronitf • updated 10 months ago (Version 1)

Data

Kernels (404)

Discussion (21)

Activity

Metadata

Download (3 KB)

New Kernel

1445

Your Dataset download has started. Show your appreciation with an upvote

1445

Reddit API Terms

classification, binary classification, health, biology

Description

Context

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Content

Attribute Information:

Data (3 KB)

Data Sources

heart.csv303 x 14

About this file

Data Set Information:

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply

Columns

age age in years

sex (1 = male; 0 = female)

cp chest pain type

trestbps resting blood pressure (in mm Hg on admission to the hospital)

chol serum cholestoral in mg/dl

fbs (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg resting electrocardiographic

전체 데이터셋을 임의로 200 명이 포함된 Training dataset 과 103 명 Validation dataset 으로 구분한 뒤 다음 각 물음에 답하시오. 분류 성능을 평가/비교할 때는 TPR, TNR, Precision, Accuracy, BCR, F1-Measure 를 모두 고려하여 의견을 서술하시오.

[Q1] 실습 시간에 사용한 "tree" package 를 사용하여 Classification Tree 를 학습한 결과물을 Plotting 하고 이에 대한 해석을 수행하시오. 또한 해당 Tree 를 pruning 을 수행하지 않은 상태에서 Validation dataset 에 대한 분류 성능을 평가하시오.

[Q2] 앞에서 생성한 Tree 에 대해서 적절한 Pruning 을 수행한 뒤 결과물을 Plotting 하고 이에 대한 해석을 수행하시오. Pruning 전과 후에 Split 에 사용된 변수는 어떤 변화가 있는가? Validation dataset 에 대한 분류 성능을 평가하고 [Q1]의 결과와 비교해보시오.

[Q3] "tree" package 이외에 R에서 Classification Tree 를 학습할 수 있는 package 를 최소 세 개 이상 찾아서 [Q1]과 [Q2]에서 사용한 데이터셋과 동일한 데이터셋으로 Classification Tree 를 학습하고 분류 성능을 평가해 보시오. 각 package 에 대해서 아래 사항들에 대해서 개별적으로 답하시오.

[Q3-1] 사용한 패키지의 이름

[Q3-2] 사용한 패키지가 Classification Tree 를 학습할 때 사용자가 지정할 수 있는 옵션의 종류와 의미

[Q3-3] 본인이 실제로 옵션을 변화시켜 가면서 학습한 Classification Tree 들의 차이점 및 패키지별로 최종적으로 선정한 Best model 에 대한 설명

[Q3-4] 각 패키지에서 제공하는 Tree plot 및 (가능할 경우) 다른 시각화 package 를 사용하여 도시한 Tree plot 들 간의 비교

[Q3-5] 각 패키지에서 제공한 Classification tree 들의 분류 성능 비교 및 논의