

의사결정나무

의사결정나무(decision tree)는 기계학습 중 하나로 특정 항목에 대한 의사 결정 규칙과 그 결과들을 나무 구조로 분류를 수행할 수 있는 통계분석 모듈입니다. 의사결정나무의 장점은 분석 과정이 직관적이고 이해하기 쉽고 질적 및 양적변수를 모두 사용할 수 있다는 점입니다. 계산 비용도 적어서 대규모의 데이터에서도 비교적 빠르게 연산이 가능합니다. 데이터마이닝에서 사용되는 결정 트리 분석법은 크게 분류와 회귀 두 종류로 나뉩니다. 분류 나무 (classification tree) 분석은 예측된 결과로 입력 데이터가 분류되는 클래스를 출력하고, 회귀 나무(regression tree) 분석은 예측된 결과로 특정 의미를 지니는 실수 값을 출력합니다. 의사결정나무 학습법은 지도 분류 학습에서 가장 유용하게 사용되고 있는 기법 중 하나입니다.

메뉴 호출하기

- 고급분석 > 분류분석 > 지도 학습 > 의사결정나무



• 변수설정 탭

의사결정나무

변수설정

분석옵션

자료분할

출력옵션

데이터

전체변수

id

gestwks

preterm

matage

sex

① 종속변수(필수)

>

<

hyp

설명변수

② 질적변수(선택-1개이상가능)

>

<

lowbw

③ 양적변수(선택-1개이상가능)

>

<

bweight

도움말

재설정

확인

취소

메뉴 요소	설명
① 종속변수	종속변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 한 개의 변수가 필수적으로 선택되어야 하며 질적변수 및 양적변수 모두 사용이 가능합니다.
② 질적변수	설명변수 중 질적변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 종속변수와 중복하여 선택할 수 없습니다. 질적변수와 양적변수 중 적어도 하나 이상의 변수를 선택해야 분석이 가능합니다.
③ 양적변수	설명변수 중 양적변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 종속변수와 중복하여 선택할 수 없습니다. 질적변수와 양적변수 중 적어도 하나 이상의 변수를 선택해야 분석이 가능합니다. 설명변수가 양적일 때는 낮은 예측능력을 보일 수 있습니다.

• 분석옵션 탭

의사결정나무

변수설정

분석옵션

자료분할

출력옵션

① 분석옵션

분석방법

☒ 분류 (Classification)

☐ 회귀 (Regression)

☐ 조건부 검정 (Conditional inference)

② ☐ 가지치기(Pruning)

도움말

재설정

확인

취소

메뉴 요소	설명
① 분석방법	<p>[변수설정] 탭의 '종속변수'에서 선택한 변수가 질적변수이면 '분류'를, 양적변수이면 '회귀'를 선택합니다.</p> <ul style="list-style-type: none"> 분류 (Classification) (Default) : 예측된 결과로 입력 데이터가 분류되는 클래스를 출력합니다. 회귀 (Regression) : 예측된 결과로 특정 의미를 지니는 실수 값을 출력합니다. 조건부 검정 (Conditional Inference) : 과적합을 피하기 위하여 여러 테스트에 대해 보정 분할 기준으로 비파라미터 (non-parametric) 테스트를 사용 하는 통계 방법입니다. 조건부 검정을 선택할 경우, 가지치기가 비활성화 됩니다.
② 가지치기(Pruning)	<p>[분석옵션]-[분석방법]에서 '분류'나 '회귀'를 선택할 경우 활성화됩니다. 분석결과에 과적합의 위험성이 존재할 경우 가지치기 과정을 거쳐서 의사결정나무 모형을 최적화 합니다.</p>

• 자료분할 탭

의사결정나무

변수설정 분석옵션 **자료분할** 출력옵션

변수목록

id
bweight
lowbw
gestwks
preterm
matage
hyp
sex

① 훈련 및 검증(필수)

• 분할검증

② • 모든 데이터를 훈련에 이용
• 비율에 따라 임의로 분할
• 변수로 분할

훈련(train) 자료 %
시험(test) 자료 %

분할변수(1-훈련, 2-시험)

③ • 교차검증

• Leave-one-out 교차검증
• K-fold 교차검증 K 10

④ 예측(선택)

분할변수(1-예측, 2-훈련 및 검증)

도움말 재설정 확인 취소

메뉴 요소	설명
① 훈련 및 검증	<p>의사결정나무모형 적합에 사용될 데이터를 훈련자료(training data)와 시험자료(test data)로 분할하는 방식으로 다음 2가지 옵션 중 1개를 선택할 수 있습니다.</p> <ul style="list-style-type: none"> 분할검증 (Default) : 훈련자료와 시험자료로 분할된 자료로 모형을 1회 검증하는 방법입니다. 교차검증 : 훈련자료와 시험자료를 변경해가며 여러 차례 반복 검증하는 방법입니다.
② 분할검증	<p>[분할검증]을 선택하는 경우 다음의 3가지 옵션이 활성화되어 이 중 1개를 선택할 수 있습니다.</p> <ul style="list-style-type: none"> 모든 데이터를 훈련에 이용 (Default) : 시험자료 없이 모든 개체를 모형 적합에 사용합니다. 비율에 따라 임의로 분할 : 훈련자료와 시험자료의 비율을 설정하여 임의로 분할하는 방식입니다. Default 값은 훈련자료 70%, 시험자료가 30% 입니다. 사용자는 훈련자료에 0~100을 입력할 수 있으며, 시험자료에는 100에서 입력한 값을 뺀 수치가 자동으로 입력됩니다. 임의로 분할된 개체들 중 훈련자료와 시험자료의 인덱스를 저장하려면 [출력옵션]-[저장]-[자료분할지표]를 선택합니다. 변수로 분할 : 훈련자료와 시험자료로 사용될 개체가 결정되어 있는 경우 이 옵션을 선택합니다. 이때, 훈련자료에 해당하는 개체는 1, 시험자료에 해당하는 개체는 2의 값을 갖는 인덱스 변수를 분할변수로 지정해주어야 합니다.

• 자료분할 탭

의사결정나무

변수설정 분석옵션 **자료분할** 출력옵션

변수목록

- id
- bweight
- lowbw
- gestwks
- preterm
- matage
- hyp
- sex

① 훈련 및 검증(필수)

☒ 분할검증

② ☒ 모든 데이터를 훈련에 이용

☐ 비율에 따라 임의로 분할

훈련(train) 자료 %

시험(test) 자료 %

☐ 변수로 분할

분할변수(1-훈련, 2-시험)

>

<

③ ☐ 교차검증

☐ Leave-one-out 교차검증

☒ K-fold 교차검증 K

④ 예측(선택)

분할변수(1-예측, 2-훈련 및 검증)

>

<

도움말 재설정 **확인** 취소

메뉴 요소

설명

③ 교차검증

[교차검증]을 선택하는 경우 다음의 2가지 옵션이 활성화되어 이 중 1개를 선택할 수 있습니다.

- Leave-one-out 교차검증 : 한 개체를 시험자료로 사용하고 나머지 개체를 모두 훈련자료로 하여 모델을 적합하는 방식으로 모든 개체에 대해 이 과정을 반복한 뒤, 전체 개체 수만큼의 모형으로부터 얻은 정확도의 평균을 모형의 최종 정확도로 계산합니다.
- K-fold 교차검증 : 전체 개체를 K개의 그룹으로 임의로 분할하여, 하나의 그룹을 시험자료로 사용하고 나머지 그룹을 모두 훈련자료로 하여 모델을 적합하는 방식으로 K개의 그룹에 대해 이 과정을 반복한 뒤, 그룹 수만큼의 모형으로부터 얻은 정확도의 평균을 모형의 최종 정확도로 계산합니다.

- K : [교차검증]-[K-fold 교차검증]을 선택할 경우 활성화됩니다. K-fold 교차검증에 사용할 K의 값을 입력합니다. 2 이상의 정수만 입력 가능하며, 전체 개체 수보다 더 큰 정수가 입력되는 경우 자동으로 Leave-one-out 교차검증을 실시합니다. Default는 10입니다.

④ 예측 > 분할변수

의사결정나무모형 적합에 사용될 훈련 및 검증 데이터와 해당 모형으로부터 예측값을 얻을 예측 데이터가 분할되어 있는 경우 사용됩니다. 훈련 및 검증에 사용되는 개체는 2, 예측에 사용되는 개체는 1의 값을 갖는 인덱스 변수를 분할변수로 지정해주어야 합니다. 예측분할변수를 지정하지 않아도 분석이 가능합니다. 예측분할변수가 지정된 경우, 예측에 해당하는 개체에 해당하는 예측값이 엑셀 시트에 "Predicted_pred_Tree"라는 변수명으로 저장됩니다.

• 출력옵션 탭

의사결정나무

변수설정

분석옵션

자료분할

출력옵션

출력

① 그래프유형

☒ 기본형
 ☐ 확장형

②

☐ 교차타당성 그래프 출력

저장

훈련자료

③

☐ 적합값

시험자료

④

☐ 예측값

⑤

☐ 자료분할지표

도움말

재설정

확인

취소

메뉴 요소	설명
① 그래프유형	<p>기본형과 확장형 중 하나를 선택합니다.</p> <ul style="list-style-type: none"> 기본형 (Default) : 기본 의사결정나무 그래프를 출력합니다. 분석옵션 탭에서 [분석방법]-'조건부 검정'을 선택한 경우 이 옵션이 비활성화됩니다. 확장형 : 막대도표가 추가된 의사결정 나무를 출력합니다.
② 교차타당성 그래프 출력	<p>분류분석을 진행할 경우 상대오차(relative error) 결과를 그래프로 출력해줍니다. 회귀분석을 진행할 경우 추가로 결정계수(r-square) 와 상대오차(relative error) 결과 그래프를 출력 해줍니다. 조건부 검정 분석에서는 그래프를 출력하지 않습니다.</p>
③ 적합값	<p>적합값을 괄호 안의 변수명으로 저장합니다. (Fitted_tree_Train)</p>
④ 예측값	<p>[자료분할] 탭에서 '비율에 따라 임의로 분할' 또는 '변수로 분할' 을 택할 경우 예측값이 활성화됩니다. 예측값을 괄호 안의 변수명으로 저장합니다. (Predicted_test_tree)</p>
⑤ 자료분할지표	<p>각 관측값이 훈련 혹은 시험자료 중 어떤 자료로 사용되었는지 여부를 괄호 안의 변수명으로 저장합니다. (Partition_idx_Tree)</p>