

서포트벡터머신

서포트벡터머신(support vector machine, SVM)은 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, 이를 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적인 선형 분류 모델(non-probabilistic binary linear classifier)을 만들어 데이터가 사상(mapping)된 공간에서 경계로 표현 되는데 이 중 가장 큰 폭을 가진 경계를 찾는 방법입니다. SVM은 지도 학습(supervised learning) 모델이며, 주로 분류와 회귀분석을 위해서 사용합니다. 선형 분류와 비선형 분류 모두 가능합니다.

메뉴 호출하기

- 고급분석 > 분류분석 > 지도 학습 > 서포트벡터머신(SVM)



• 변수설정 탭

서포트벡터머신

변수설정 분석옵션 자료분할 출력옵션

데이터

전체변수

id
bweight
lowbw
gestwks
preterm
matage
hyp
sex

① 종속변수(필수)

> <

* 분류분석 진행 - 질적변수 선택
* 회귀분석 진행 - 양적변수 선택

설명변수

② 질적변수(선택-1개이상가능)

> <

③ 양적변수(선택-1개이상가능)

> <

④ ☐ 설명변수 표준화

도움말 재설정 확인 취소

메뉴 요소	설명
① 종속변수	종속변수에 해당하는 변수를 전체변수로부터 선택할 수 있습니다. 한 개의 변수가 필수적으로 선택되어야 하며 양적변수와 질적변수 모두 사용이 가능합니다. 종속변수에 결측치가 존재하는 관측치는 분석에서 제외됩니다.
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.
④ 설명변수 표준화	설명변수를 표준화 하여 분석에 사용합니다.

• 분석옵션 탭

서포트벡터머신

변수설정 분석옵션 자료분할 출력옵션

① 분석방법

☒ 분류분석(C-classification) ☐ 회귀분석(EPS-regression)

② 커널함수

☒ radial ☐ polynomial ☐ linear

차수 Gamma Cost

* 차수, gamma, cost는 0보다 큰 수로 입력
 * gamma와 cost는 실표로 구분하여 여러 값 입력가능. 입력된 값 중 최적모델 적합
 * gamma와 cost 값을 여러개 지정한 경우 튜닝 분석이 진행됨

도움말 재설정 확인 취소

메뉴 요소	설명
① 분석방법	<p>[변수설정] 탭에서 종속변수로 선택한 자료형에 따라 아래 기법 중 하나를 선택합니다.</p> <ul style="list-style-type: none"> C-classification (Default) : 종속변수로 질적변수를 택한 경우 선택합니다. EPS-regression : 종속변수로 양적변수를 택한 경우 선택합니다.
② 커널함수	<p>서포트벡터머신의 3가지 커널함수 중 하나를 선택합니다.</p> <ul style="list-style-type: none"> radial (Default) : 가우시안 커널을 사용합니다. polynomial : polynomial을 선택할 경우, '차수'가 활성화됩니다. linear : linear을 선택할 경우, 'Gamma'가 비활성화됩니다. 차수 : [커널함수]-'polynomial' 선택 시 활성화됩니다. 차수는 양의 정수만 입력 가능하며, Default는 3입니다. Gamma : [커널함수]-'radial' 또는 'polynomial' 선택 시 활성화됩니다. 0보다 큰 수를 입력해야 합니다. 감마를 여러 개 입력할 수 있으며 이 경우, 튜닝분석이 진행됩니다. Cost : Cost는 과적합을 막는 정도를 말합니다. 0보다 큰 수를 입력해야 분석이 가능합니다. 비용 값을 여러 개 입력할 수 있으며 이 경우 튜닝분석이 진행됩니다. Default는 1입니다.

• 자료분할 탭

서포트벡터머신

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
① 훈련 및 검증	<p>서포트벡터머신 모형 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

서포트벡터머신

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 모형을 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 모형을 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

서포트벡터머신 모형 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_SVM"이라는 변수명으로 저장됩니다.

- 출력옵션 탭

서포트벡터머신

변수설정 분석옵션 자료분할 **출력옵션**

저장

훈련자료

① ☐ 적합값

시험자료

② ☐ 예측값

③ ☐ 자료분할지표

도움말 재설정 **확인** 취소

메뉴 요소	설명
① 적합값	적합값을 괄호 안의 변수명으로 저장합니다. (Fitted_SVM_Train)
② 예측값	[자료분할] 탭에서 '비율에 따라 임의로 분할' 또는 '변수로 분할' 을 택할 경우 예측값이 활성화됩니다. 예측값을 괄호 안의 변수명으로 저장합니다. (Predicted_test_SVM)
③ 자료분할지표	각 관측값이 훈련 혹은 시험자료 중 어떤 자료로 사용되었는지 여부를 괄호 안의 변수명으로 저장합니다. (Partition_idx_SVM)