다중 회귀분석

Key words

#다중선형회귀 #가변수

#다중공선성 #더미변수

#분산팽창계수 #dmatrices

#VIF #variance_inflation_factor

01 다중 회귀분석 개요

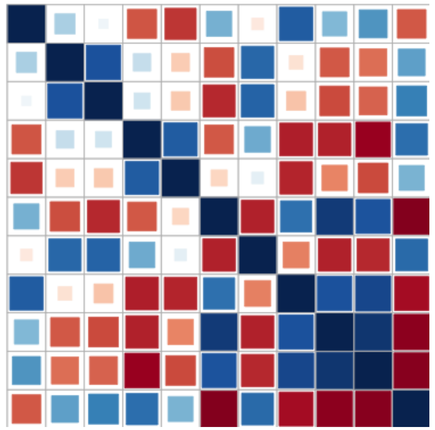
다중 회귀분석 특징

- 연속형 종속변수와 두 개 이상의 독립변수 간 선형관계 및 설명력을 확인하는 기법
- 필요 시 모델 성능 향상을 위한 파생변수 생성 및 성능 비교 필요
- 명목형 변수가 독립변수인 경우 가변수 변환 후 모델 적합

01 다중 회귀분석 개요

다중 공선성 문제

- 독립변수 간 강한 상관관계가 나타나는 문제
- 상관계수를 확인하여 그 값이 높은 것을 사전에 제거
- 회귀 모델 생성 이후 분산 팽창 계수(VIF) 확인(10 이상)하여 관련 변수 처리



<상관계수 행렬도 예시>

02 주요 함수 및 메서드 소개

patsy - dmatrices()

- 수식을 기반으로 데이터 행렬을 생성하는 pasty의 함수
- 분산 팽창 계수 확인을 위해 입력 데이터를 전처리 할 때 필요한 함수
- return_type 인자에 “dataframe” 으로 설정 시 후처리 용이

statsmodels - variance_inflation_factor()

- 분산 팽창 계수를 연산하기 위한 statsmodels 함수
- 분산 팽창 계수 연산을 위해 반복문 또는 list comprehension 사용

Q1

Price를 종속변수로 하고 나머지 수치형 변수를 독립변수로 했을 때 다중 공선성의 문제가 있다고 판단되는 변수의 개수는?

※ diamonds.csv 파일 사용

1. 5개
2. 4개
3. 3개
4. 2개

Q1

Price를 종속변수로 하고 나머지 수치형 변수를 독립변수로 했을 때 다중 공선성의 문제가 있다고 판단되는 변수의 개수는?

※ diamonds.csv 파일 사용

1. 5개
2. 4개 ✓
3. 3개
4. 2개

정답

2

본 문제에서 variance_inflation_factor() 함수로 산출한 VIF값 중
Intercept는 절편이기 때문에 10이 넘더라도 무시한다.
그리하여 다중 공선성의 문제가 있는 변수는 총 4개 이다.

Q2

price를 종속변수로 하고 carat과 depth를 독립변수로 하여 생성한 선형 회귀 모델을 사용하여 알아본 carat이 1이고 depth가 60, table이 55인 다이아몬드의 가격은 얼마인가?

※ diamonds.csv 파일 사용

1. 2818
2. 3219
3. 4912
4. 5681

Q2

price를 종속변수로 하고 carat과 depth를 독립변수로 하여 생성한 선형 회귀 모델을 사용하여 알아본 carat이 1이고 depth가 60, table이 55인 다이아몬드의 가격은 얼마인가?

※ diamonds.csv 파일 사용

1. 2818
2. 3219
3. 4912
4. 5681 ✓

정답

4

문제에서 제시한 조건으로 신규 데이터프레임을 생성한다.
해당 객체를 모델 객체의 predict() 메서드 입력값으로 설정 후
연산결과를 확인하면 5681 확인이 가능하다.

Q3

price를 종속변수로 하고 carat, color, depth를 독립변수로 하여 생성한 선형 회귀 모델을 사용하여 알아본 carat이 1이고 depth가 50, color가 E인 다이아몬드의 가격은 얼마인가?

※ diamonds.csv 파일 사용
※ 가변수 생성 시 마지막 변수 하나를 제거

1. 6885
2. 2142
3. 2840
4. 4351

Q3

price를 종속변수로 하고 carat, color, depth를 독립변수로 하여 생성한 선형 회귀 모델을 사용하여 알아본 carat이 1이고 depth가 50, color가 E인 다이아몬드의 가격은 얼마인가?

※ diamonds.csv 파일 사용
※ 가변수 생성 시 마지막 변수 하나를 제거

- [x] 1 6885
- [ ] 2 2142
- [ ] 3 2840
- [ ] 4 4351

정답

1

모델 생성 후 예측값을 입력할 때 기존 학습 모델의 더미변수 생성 규칙을 통일해주어야 한다. 즉, predict() 메서드에 입력되는 데이터프레임은 변수가 3개가 아닌 8개이다.

#다중 선형회귀분석

#다중 공선성

#patsy – dmatrices()

#statsmodels
– variance_inflation_factor()

분류: 로지스틱 회귀분석

Key words

#경계값
#threshold
#승산비
#OR

#Logit
#LogisticRegression
#roc_auc_curve
#predict_proba

#accuracy_score
#f1_score
#precision_score
#recall_score

01 로지스틱 회귀분석 개요

로지스틱 회귀분석 특징

- 이항 로지스틱 회귀 분석은 종속변수가 0과 1이며 베르누이 분포를 따를 경우 사용
- 모델의 산출 값은 각 데이터가 1이 될 확률이며 이진 분류를 위해서 경계값(threshold)이 필요
- 모델 평가를 위해 각종 분류 관련 지표 및 AUC 활용

01 로지스틱 회귀분석 개요

승산비(OR, Odds Ratio)

- 특정 독립변수를 제외한 나머지 값을 고정하고
  해당 독립변수가 1 증가 시 변화하는 승산(odds)의 비

$$\frac{odds(x_1, \dots, x_i + 1, \dots\ x_n)}{odds(x_1, \dots\ , x_i, \dots\ x_n)} = \frac{e^{\beta_1 + \beta_1 x_1, \dots + \beta_i(x_i+1) + \cdots + \beta_n x_n}}{e^{\beta_1 + \beta_1 x_1, \dots + \beta_i x_i + \cdots + \beta_n x_n}} = e^{\beta_i}$$

02 주요 함수 및 메서드 소개

statsmodels - Logit()

- 로지스틱 회귀분석을 실시하는 statsmodels의 함수
- endog, exog 인자에 각각 종속변수와 독립변수를 할당
- 산출 모델 객체의 params 어트리뷰트에 모델의 계수 저장
- 산출 모델 객체의 predict() 메서드로 예측값을 생산하며 이는 종속변수가 1이 될 확률값

02 주요 함수 및 메서드 소개

sklearn - LogisticRegression()

- 로지스틱 회귀분석을 실시하는 sklearn의 함수
- fit_intercept, solver 인자로 절편 적합 여부 및 최적화 알고리즘 설정 가능
- random_state 인자에 자연수를 할당하여 결과 고정 가능
- fit() 메서드에 독립변수 및 종속변수 할당
- 산출 모델 객체의 coef_ 어트리뷰트에 모델의 계수 저장
- 산출 모델 객체의 predict_proba() 메서드로 예측값을 생산하며 두 번째 열이 종속변수가 1 이 될 확률값

02 주요 함수 및 메서드 소개

sklearn - roc_auc_score()

- AUC(Area Under Curve)를 산출하는 sklearn의 함수
- y_true, y_score 인자에 각각 종속변수와 예측 확률값 할당

02 주요 함수 및 메서드 소개

sklearn - accuracy_score()

- 분류모델의 정확도를 산출하는 sklearn의 함수
- y_pred와 y_true에 각각 예측 분류 결과와 실제 값을 할당

sklearn - f1_score()

- 분류모델의 f1 값을 산출하는 sklearn의 함수
- y_pred와 y_true에 각각 예측 분류 결과와 실제 값을 할당

02 주요 함수 및 메서드 소개

sklearn - precision_score()

- 분류모델의 정밀도(precision)를 산출하는 sklearn의 함수
- y_pred와 y_true에 각각 예측 분류 결과와 실제 값을 할당

sklearn - recall_score()

- 분류모델의 재현율(recall)를 산출하는 sklearn의 함수
- y_pred와 y_true에 각각 예측 분류 결과와 실제 값을 할당

Q1

독립변수를 혈압, 혈당, BMI, 인슐린으로 하고 종속변수를 당뇨 여부로 할 경우 분류 정확도는 얼마인가?

※ diabetes.csv 파일 사용
※ statsmodels 함수 사용
※ 데이터는 학습:평가 = 8:2 로 분리 후 계산
※ Seed는 123

1. 0.73
2. 0.72
3. 0.71
4. 0.70

Q1

독립변수를 혈압, 혈당, BMI, 인슐린으로 하고 종속변수를 당뇨 여부로 할 경우 분류 정확도는 얼마인가?

※ diabetes.csv 파일 사용
※ statsmodels 함수 사용
※ 데이터는 학습:평가 = 8:2 로 분리 후 계산
※ Seed는 123

1. 0.73
2. 0.72
3. 0.71
4. 0.70

정답 4

학습/평가 데이터 분리 후 Logit() 함수로 학습을 실시한다.
산출되는 예측 확률값을 0.5 기준으로 나누어 분류하고
평가 데이터의 종속변수와 함께 accuracy_score() 함수로
분류 정확도를 확인한다.

Q2 독립변수를 혈당, BMI, 나이로 하고 종속변수를 당뇨 여부로 할 경우 나이의 승산비는 얼마인가?

※ diabetes.csv 파일 사용
※ statsmodels 함수 사용

1. 0.02
2. 1.03
3. 1.05
4. 0.99

Q2 독립변수를 혈당, BMI, 나이로 하고 종속변수를 당뇨 여부로 할 경우 나이의 승산비는 얼마인가?

※ diabetes.csv 파일 사용
※ statsmodels 함수 사용

1. 0.02
2. 1.03
3. 1.05
4. ✓ 0.99

정답 4

학습한 모델 객체의 params 어트리뷰트로 각 변수의 계수를 확보한 후, 해당 값과 np.exp() 함수로 승산비를 계산할 수 있다.

Q3

독립변수를 혈당, BMI, 나이로 하고 종속변수를 당뇨 여부로 할 경우 모델의 AUC는 얼마인가?

※ diabetes.csv 파일 사용
※ statsmodels 함수 사용

1. 0.56
2. 0.55
3. 0.54
4. 0.53

Q3 독립변수를 혈당, BMI, 나이로 하고 종속변수를 당뇨 여부로 할 경우 모델의 AUC는 얼마인가?

※ diabetes.csv 파일 사용
※ statsmodels 함수 사용

1. 0.56
2. 0.55
3. 0.54
4. 0.53

정답

3

데이터 분할 관련 언급이 없기 때문에 학습과 평가 데이터 세트를 동일하게 하여 예측 확률값을 산출한다. 그리고 AUC를 계산하기 위해 roc_auc_score() 함수를 사용할 수 있다.

#로지스틱 회귀분석

#statsmodels – Logit()

#scipy
- LogiticRegression()
- roc_auc_Score()

분류: 나이브 베이즈

Key words

#분류모델 #사전확률 #GaussianNB

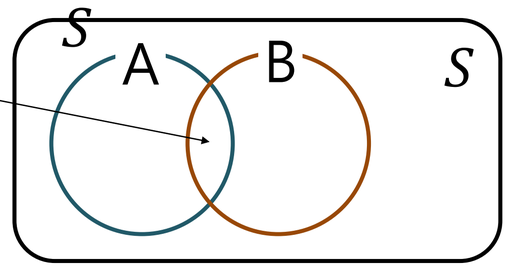
#나이브베이즈 #사후확률 #predict_proba

01 나이브 베이즈 개요

나이브 베이즈 분류기 특징

- 사전 확률 및 추가 정보를 기반으로 사후 확률을 추론하는 통계적 방법인 베이즈 추정 기반 분류
- 종속변수 각 범주의 등장 빈도인 사전확률(prior) 설정이 중요
- 각 데이터의 사전 확률을 기반으로 사후확률(posterior)을 계산

$$\mathrm{P}(A|B) = \frac{P(A \cap B)}{P(B)}$$



S: 표본공간

02 주요 함수 및 메서드 소개

sklearn - GaussianNB()

- 나이브베이즈 분류 모델을 위한 sklearn의 함수
- 독립변수와 종속변수는 GaussianNB() 함수의 메서드인 fit() 함수에 할당
- 모델 객체의 predict_proba() 메서드로 예측 확률값을 생산
- 이진 분류의 경우 출력된 예측 확률값의 두 번째 열이 1이 될 확률

Q1

BMI가 0 초과인 데이터만 사용하여 나이브 베이즈 분류를 실시하고자 한다. Outcome을 종속변수로 하고 나머지 변수를 독립변수로 할 때 종속변수의 사전확률은?

※ diabates.csv 파일 사용
※ Outcome 1을 대상으로 사전확률을 계산한다.

1. 0.8
2. 0.64
3. 0.35
4. 0.41

Q1

BMI가 0 초과인 데이터만 사용하여 나이브 베이즈 분류를 실시하고자 한다. Outcome을 종속변수로 하고 나머지 변수를 독립변수로 할 때 종속변수의 사전확률은?

※ diabates.csv 파일 사용
※ Outcome 1을 대상으로 사전확률을 계산한다.

1. 0.8
2. 0.64
3. 0.35
4. 0.41

정답

3

종속변수의 사전확률은 입력되는 종속변수의 각 항목의 비율이다. 해당 비율을 구할 때 value_counts() 메서드를 사용하면 수월하다.

Q2 혈당, 혈압, 나이를 독립변수로 하고 당뇨 발병 여부를 종속변수로 했을 때 그 정확도는 얼마인가?

※ diabates.csv 파일 사용
※ Outcome 1을 대상으로 사전확률을 계산한다.

1. 78%
2. 76%
3. 74%
4. 72%

Q2 혈당, 혈압, 나이를 독립변수로 하고 당뇨 발병 여부를 종속변수로 했을 때 그 정확도는 얼마인가?

※ diabates.csv 파일 사용
※ Outcome 1을 대상으로 사전확률을 계산한다.

1. 78%
2. 76%
3. 74%
4. 72%

정답

2

별도의 언급이 없기 때문에 학습/평가 데이터를 동일하게 사용하며 출력된 확률값을 나누는 문턱값 또한 0.5로 지정한다. 정확도를 산출하기 위해 sklearn의 accuracy_score() 함수를 사용한다.

Q3 임신여부, 연령대, BMI, 혈당을 독립변수로 하고 당뇨 발병 여부를 종속변수로 했을 때 나이브 베이즈와 로지스틱 회귀 분석을 실시하고 둘 중 정확도가 높은 모델의 정확도는?

※ diabates.csv 파일 사용
※ BMI가 0 초과인 것을 사용하며 학습/평가 데이터 세트를 8:2로 분할, Seed는 123
※ 연령대는 Age가 21인 경우 20으로, 39일 경우 30으로 계산한다.
※ sklearn의 로지스틱 회귀 함수를 사용하며, 임계값(threshold)은 0.5로 한다.

1. 74%
2. 76%
3. 78%
4. 83%

Q3 임신여부, 연령대, BMI, 혈당을 독립변수로 하고 당뇨 발병 여부를 종속변수로 했을 때 나이브 베이즈와 로지스틱 회귀 분석을 실시하고 둘 중 정확도가 높은 모델의 정확도는?

※ diabates.csv 파일 사용
※ BMI가 0 초과인 것을 사용하며 학습/평가 데이터 세트를 8:2로 분할, Seed는 123
※ 연령대는 Age가 21인 경우 20으로, 39일 경우 30으로 계산한다.
※ sklearn의 로지스틱 회귀 함수를 사용하며, 임계값(threshold)은 0.5로 한다.

1. 74%
2. 76%
3. 78%
4. 83%

정답 4

임신여부와 연령대는 기존 데이터에서 제공하지 않는 변수로 별도 생성 후 모델에 반영해야 한다. 모델 객체의 predict_proba() 메서드를 활용하여 예측 확률값을 생성해야 정확도를 확인 할 수 있다.

#나이브 베이즈

#사전확률 중요

#sklearn – GaussianNB()

KNN(K-Nearest Neighbor)

Key words

#분류모델 #KNeighborsClassifier

#회귀모델 #KNeighborsRegressor

01 KNN 개요

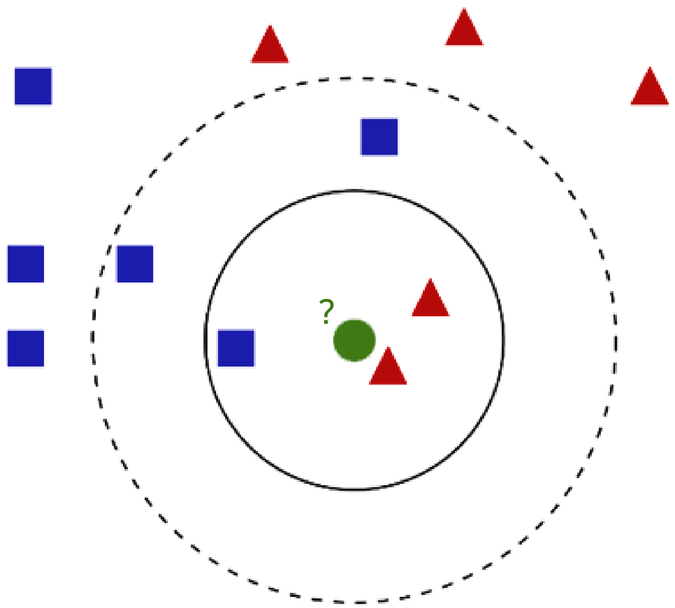
KNN 분류(Classification)

- 새로운 값은 기존의 데이터를 기준으로 가장 가까운
  k개의 최근접 값을 기준으로 분류됨
- k는 동률의 문제 때문에 짝수는 되도록이면 피하는 것이 좋음
- k가 1에 가까울수록 과적합, k가 클수록 과소적합이 되기 때문에
  적절한 k값 선정 필요

01 KNN 개요

KNN 회귀(Regression)

- 기본 개념은 분류모델과 같으며 k개의 인접한 자료의 (가중)평균으로 예측



02 주요 함수 및 메서드 소개

sklearn - KNeighborsClassifier()

- KNN 분류 모델을 학습하기 위한 sklearn의 함수
- n_neighbors 인자에 학습 시 고려할 이웃 데이터의 개수를 지정
- n_neighbors가 1에 가까울수록 과적합되며 커질수록 과소적합되는 경향 존재
- KNeighborsClassifier() 함수의 fit() 메서드에 독립변수와 종속변수 할당

02 주요 함수 및 메서드 소개

sklearn - KNeighborsRegressor()

- KNN 회귀 모델을 학습하기 위한 sklearn의 함수
- n_neighbors 인자에 학습 시 고려할 이웃 데이터의 개수를 지정
- n_neighbors가 1에 가까울수록 과적합되며 커질수록 과소적합되는 경향 존재
- KNeighborsClassifier() 함수의 fit() 메서드에 독립변수와 종속변수 할당

Q1 당뇨 발생 여부를 예측하기 위해 임신 횟수, 혈당, 혈압을 사용할 경우 그 정확도는 얼마인가?

※ diabates.csv 파일 사용
※ 데이터를 학습7, 평가3의 비율로 분할하시오. (Seed는 123)
※ 설정은 모두 기본값으로 하시오.

1. 0.73
2. 0.64
3. 0.84
4. 0.67

Q1 당뇨 발생 여부를 예측하기 위해 임신 횟수, 혈당, 혈압을 사용할 경우 그 정확도는 얼마인가?

※ diabates.csv 파일 사용
※ 데이터를 학습7, 평가3의 비율로 분할하시오. (Seed는 123)
※ 설정은 모두 기본값으로 하시오.

- [x] 0.73
- 2 0.64
- 3 0.84
- 4 0.67

정답

1

KNeighborClassifier() 함수를 그대로 사용할 경우 5개의 이웃 데이터만 사용하여 학습이 된다. 학습한 모델 객체를 기반으로 예측값을 생산하고 accuracy_score() 함수로 정확도를 산출한다.

Q2

종속변수를 당뇨 발병 여부로 하고 임신여부, 혈당, 혈압, 인슐린, 체질량지수를 독립변수로 하여 정확도를 확인했을 때 그 k 값과 정확도가 올바르게 연결되지 않은 것은?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오. (Seed는 123)
※ k를 제외한 설정은 모두 기본값으로 하시오.

1. k = 3, Acc. = 0.71
2. k = 5, Acc. = 0.73
3. k = 10, Acc. = 0.78
4. k = 20, Acc. = 0.79

Q2

종속변수를 당뇨 발병 여부로 하고 임신여부, 혈당, 혈압, 인슐린, 체질량지수를 독립변수로 하여 정확도를 확인했을 때 그 k 값과 정확도가 올바르게 연결되지 않은 것은?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오. (Seed는 123)
※ k를 제외한 설정은 모두 기본값으로 하시오.

1. k = 3, Acc. = 0.71
2. k = 5, Acc. = 0.73
3. k = 10, Acc. = 0.78
4. k = 20, Acc. = 0.79

정답

4

KNeighborsClassifier() 함수의 이웃 개수 설정을 변수로 두고 반복문을 작성하는 경우 보다 간결하고 정확하게 결과값을 산출할 수 있다.

Q3

종속변수를 체질량지수로 하고 임신여부, 혈당, 혈압, 인슐린을 독립변수로 하여 예측값을 확인했을 때 그 k 값과 RMSE가 올바르게 연결되지 않은 것은?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오. (Seed는 123)
※ k를 제외한 설정은 모두 기본값으로 하시오.

1. k = 3, RMSE = 8.455
2. k = 5, RMSE = 8.675
3. k = 10, RMSE = 8.404
4. k = 20, RMSE = 8.510

Q3

종속변수를 체질량지수로 하고 임신여부, 혈당, 혈압, 인슐린을 독립변수로 하여 예측값을 확인했을 때 그 k 값과 RMSE가 올바르게 연결되지 않은 것은?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오. (Seed는 123)
※ k를 제외한 설정은 모두 기본값으로 하시오.

1. k = 3, RMSE = 8.455
2. k = 5, RMSE = 8.675
3. k = 10, RMSE = 8.404
4. k = 20, RMSE = 8.510

정답

3

각 모델에 따른 RMSE 산출은 반복문으로 처리하는 것이 좋으며,
RMSE 계산을 위해서는 mean_squared_error() 함수와
** 연산자 활용을 권장한다.

#KNN 분류

#KNN 회귀

#sklearn
– KNeighborsClassfier()
– KNeighborsRegressor()

의사결정나무 모델
: 분류 및 회귀나무

Key words

#분류나무 #DecisionTreeClassifier

#회귀나무 #DecisionTreeRegressor

01 의사결정나무 모델 개요

분류 나무(Classification Tree)

- 종속변수가 **명목형**인 경우 사용하는 의사결정나무 모델
- 각 노드 분류 알고리즘은 이진 분류 시 **지니지수(Gini index) 기반의 CART** 사용
- 과적합 방지 및 모델 단순화를 위해 Depth, Impurity 등 관련 설정 활용

01 의사결정나무 모델 개요

회귀 나무(Regression Tree)

- 종속변수가 **연속형**인 경우 사용하는 의사결정나무 모델
- 각 노드 분류에는 **평균과 표준편차**를 활용
- 과적합 방지 및 모델 단순화를 위해 Depth, Impurity 등 관련 설정 활용

02 주요 함수 및 메서드 소개

sklearn - DecisionTreeClassifier()

- 의사결정나무의 **분류나무**를 수행할 때 사용하는 sklearn의 함수
- max_depth와 random_state로 모델의 성장과 결과 고정 설정 가능
- DecisionTreeClassifier() 함수의 fit() 메서드에 독립변수와 종속변수를 할당

sklearn - DecisionTreeRegressor()

- 의사결정나무의 **회귀나무**를 수행할 때 사용하는 sklearn의 함수
- max_depth와 random_state로 모델의 성장과 결과 고정 설정 가능
- DecisionTreeRegressor() 함수의 fit() 메서드에 독립변수와 종속변수를 할당

Q1

당뇨병 발병 여부를 예측하기 위하여 의사결정나무를 사용하고자 한다. 이 때 혈당, 혈압, 임신 횟수를 기반으로 예측을 했을 때 예측 정확도는?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오.
※ Seed는 123

1 60%

2 61%

3 62%

4 63%

Q1

당뇨병 발병 여부를 예측하기 위하여 의사결정나무를 사용하고자 한다. 이 때 혈당, 혈압, 임신 횟수를 기반으로 예측을 했을 때 예측 정확도는?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오.
※ Seed는 123

1. 60%
2. 61%
3. 62%
4. 63% ✓

정답

4

데이터 분리와 모델 학습 시 Seed를 123으로 고정시켜야 올바른 결과를 산출할 수 있다. 추가로 정확도 계산은 accuracy_score() 함수에 평가 세트의 종속변수와 모델 예측값을 입력하여 계산한다.

Q2 환자의 BMI를 예측하기 위하여 회귀나무를 사용하고자 한다. 이 때 혈당, 혈압, 피부 두께를 독립변수로 했을 경우 RMSE는 얼마인가?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오.
※ Seed는 123

1. 9.9
2. 9.8
3. 9.7
4. 9.6

Q2 환자의 BMI를 예측하기 위하여 회귀나무를 사용하고자 한다. 이 때 혈당, 혈압, 피부 두께를 독립변수로 했을 경우 RMSE는 얼마인가?

※ diabates.csv 파일 사용
※ 데이터를 학습8, 평가2의 비율로 분할하시오.
※ Seed는 123

- [x] 9.9
- [ ] 2 9.8
- [ ] 3 9.7
- [ ] 4 9.6

정답

1

실수로 DecisionTreeClassifier() 함수를 사용하지 않도록 한다.
그리고 mean_squared_error() 함수와 ** 연산자를 활용하여
RMSE를 연산할 수 있다.

Q3

분류나무의 파라미터를 바꿔가면서 성능 평가를 하려고 한다. 당뇨 발병 여부를 종속변수로 하고 혈당, 혈압, 임신 횟수, BMI, 나이를 독립변수로 하고 Depth를 3에서 6까지 변화시킬 때 그 결과로 틀린 것은?

※ diabates.csv 파일 사용
※ 데이터를 학습7, 평가3의 비율로 분할하시오.
※ Seed는 345

1. Depth 3, 정확도: 0.77
2. Depth 4, 정확도: 0.76
3. Depth 5, 정확도: 0.76
4. Depth 6, 정확도: 0.78

Q3

분류나무의 파라미터를 바꿔가면서 성능 평가를 하려고 한다. 당뇨 발병 여부를 종속변수로 하고 혈당, 혈압, 임신 횟수, BMI, 나이를 독립변수로 하고 Depth를 3에서 6까지 변화시킬 때 그 결과로 틀린 것은?

※ diabates.csv 파일 사용
※ 데이터를 학습7, 평가3의 비율로 분할하시오.
※ Seed는 345

1. Depth 3, 정확도: 0.77
2. Depth 4, 정확도: 0.76
3. Depth 5, 정확도: 0.76
4. Depth 6, 정확도: 0.78

정답

4

Seed를 345가 아닌 123으로 설정하지 않도록 주의한다.
모델 학습은 for 반복문을 활용하는 것이 보다 간결하고 정확하다.

#의사결정나무

#분류나무 →명목형 종속변수

#회귀나무 →연속형 종속변수

#결과고정/튜닝/과적합 방지

추천: 연관성 분석

Key words

#추천시스템 #mlxtend

#연관규칙 #apriori

#장바구니분석 #association_rules

01 연관성 분석 개요

연관성 분석(Association Rule) 특징

- 상품 또는 서비스간의 관계 속에서 유용한 규칙을 찾을 때 사용
- 유통 분야에서 주로 활용되며 **장바구니 분석(Market Basket Analysis)**이라는 별칭 존재
- 비즈니스적으로 중요한 요소를 고려하기 어렵고, 연산량이 많음

01 연관성 분석 개요

주요 평가 지표

- 지지도(Support): 상품 X와 Y를 동시에 구매한 비율, **규칙의 중요성**
- 신뢰도(Confidence): 상품 X를 구매 시 Y를 구매한 비율(조건부 확률), **규칙의 신뢰성**
- 향상도(Lift): 상품 X 구매 시 임의 상품 구입 대비 Y 를 포함하는 경우의 비중, **규칙의 상관성**

향상도 해석

- Lift > 1: 품목 간 양의 상관 관계(보완재)
- Lift = 1: 품목 간 상호 독립 관계
- Lift < 1: 품목 간 음의 상관 관계(대체재)

02 데이터 소개

제품 구매 데이터 - association_rules_mart.csv

- 익명화된 고객의 제품 구매 데이터 4만건

	Date	ID	Item
0	2014-01-01	1249in804	citrus fruit
1	2014-01-01	1249in804	coffee
2	2014-01-01	1381ht273	curd
3	2014-01-01	1381ht273	soda
4	2014-01-01	1440kn258	other vegetables
5	2014-01-01	1440kn258	yogurt

03 주요 함수 및 메서드 소개

mlxtend - apriori()

- 구매 아이템 빈도를 계산하는 mlxtend의 함수
- 입력 데이터 세트는 구매 아이템 기반으로 더미변수화(OHE) 되어있어야 함
- min_support 와 max_len 인자로 최소 지지도와 아이템 조합 최대값을 설정
- use_colnames 인자를 True로 하여 분석을 하는 것을 권장

03 주요 함수 및 메서드 소개

mlxtend - association_rules()

- 구매 아이템 빈도를 활용하여 연관규칙을 계산하는 mlxtend의 함수
- metric에 필터링 기준 지표를 설정하고 min_threshold에 그 경계값을 지정

Q1

최소 지지도와 신뢰도를 0.005로 설정하고 연관성 분석을 실시했을 때 지지도가 0.1 이상인 규칙은 몇 개 인가?

※ association_rules_mart.csv 파일 사용
※ 사전 중복 제거 실시

1. 0
2. 26
3. 2310
4. 74998

Q1

최소 지지도와 신뢰도를 0.005로 설정하고 연관성 분석을 실시했을 때 지지도가 0.1 이상인 규칙은 몇 개 인가?

※ association_rules_mart.csv 파일 사용
※ 사전 중복 제거 실시

1. 0
2. 26
3. 2310
4. 74998

정답

2

apriori() 함수와 association_rules() 함수에서 지지도와 신뢰도를 0.005를 기준으로 연관규칙을 산출한 후 지지도 0.1 기준으로 필터링 한 다음 row 개수를 확인해야 한다.

Q2

최소 지지도와 신뢰도를 0.005로 설정하고 연관성 분석을 실시했을 때 지지도가 0.01 이상인 규칙 중 향상도가 가장 높은 규칙과 관련 없는 품목은?

※ association_rules_mart.csv 파일 사용
※ 사전 중복 제거 실시
※ max_len = 3으로 설정

1. meat
2. egg
3. milk
4. beer

Q2

최소 지지도와 신뢰도를 0.005로 설정하고 연관성 분석을 실시했을 때 지지도가 0.01 이상인 규칙 중 향상도가 가장 높은 규칙과 관련 없는 품목은?

※ association_rules_mart.csv 파일 사용
※ 사전 중복 제거 실시
※ max_len = 3으로 설정

1. meat
2. egg
3. milk
4. beer ✓

정답 4

연관규칙 결과물에서 향상도를 기준으로 sort_values() 메서드를 사용하여 내림차순 정렬을 한 후 관련 아이템을 확인하면 된다.

Q3

판매 실적 상위 30위 품목만 사용하여 최소 지지도와 신뢰도를 0.005로 설정한 연관성 분석 결과를 보았을 때 지지도가 3% 이상인 규칙 중 가장 높은 향상도는 얼마인가?

※ association_rules_mart.csv 파일 사용
※ 판매 실적은 개수로 하며 1행당 1개로 취급

1. 0.034
2. 0.179
3. 0.160
4. 0.308

Q3

판매 실적 상위 30위 품목만 사용하여 최소 지지도와 신뢰도를 0.005로 설정한 연관성 분석 결과를 보았을 때 지지도가 3% 이상인 규칙 중 가장 높은 향상도는 얼마인가?

※ association_rules_mart.csv 파일 사용
※ 판매 실적은 개수로 하며 1행당 1개로 취급

1. 0.034
2. 0.179
3. ✓ 0.160
4. 0.308

정답

3

value_counts() 와 sort_values() 메서드를 활용하여
상위 30위 품목을 산출한다. 그 후 해당 객체를 기반으로
merge() 또는 isin() 메서드를 활용하여 본 데이터를 필터링 한다.

#연관성 분석

#pivot 테이블로 데이터 전처리

#mlxtend – apriori()
– association_rules()

주성분 분석(PCA)

Key words

#주성분분석 #누적분산비

#PCA #cumsum

#분산비

01 주성분 분석(PCA) 개요

주성분 분석(PCA) 특징

- 특정 데이터의 주성분(Principal Component)를 찾는 방법
- 대표적인 차원 축소 기법
- 입력 변수 개수와 각 주성분의 설명 비를 고려하여 주성분 개수 결정

주성분(Principal Component)

- 입력 변수를 기반으로 최대의 분산을 가지는 새로운 변수
- 각 주성분은 직교하기 때문에 상관계수가 0에 가까움

02 주요 함수 및 메서드 소개

sklearn - PCA()

- 주성분 분석을 시행하기 위한 sklearn의 함수
- n_component 인자에 산출할 주성분 개수 입력
- PCA() 함수로 생성한 객체의 fit_transform() 메서드로 주성분 연산
- PCA() 함수로 생성한 객체의 explained_variance로
  각 주성분의 분산 파악 가능

02 주요 함수 및 메서드 소개

pandas - cumsum()

- 숫자 원소가 있는 시리즈 객체의 누적합을 계산하기 위한 pandas의 메서드
- 주성분의 분산 또는 분산비를 활용하여 누적 분산 또는 누적 분산비 계산 용이

Q1

x, y, z 변수와 해당 변수를 기반으로 3개의 주성분을 생성했을 때 기존 변수의 상관계수와 주성분의 상관계수의 최대값은 각각 얼마인가?

※ diamonds.csv 파일 사용
※ Seed는 123으로 한다.

1. 0.974, 0.970
2. 0.970, 0.952
3. 0.1, 0.970
4. 0.975, 0

Q1

x, y, z 변수와 해당 변수를 기반으로 3개의 주성분을 생성했을 때 기존 변수의 상관계수와 주성분의 상관계수의 최대값은 각각 얼마인가?

※ diamonds.csv 파일 사용
※ Seed는 123으로 한다.

1. 0.974, 0.970
2. 0.970, 0.952
3. 0.1, 0.970
4. 0.975, 0

정답

4

PCA() 함수에 n_components에 3을 할당하여 3개의 주성분을 산출한다. 산출물을 데이터프레임으로 변환 후 corr() 메서드로 각 상관계수를 알아본다.

Q2

x, y, z, table, depth 변수를 사용하여 주성분 분석을 실시하였을 때 누적 분산이 99%가 최초로 넘는 주성분은 몇 번째 주성분인가?

※ diamonds.csv 파일 사용
※ Seed는 123으로 한다.

1. 2
2. 3
3. 4
4. 5

Q2

x, y, z, table, depth 변수를 사용하여 주성분 분석을 실시하였을 때 누적 분산이 99%가 최초로 넘는 주성분은 몇 번째 주성분인가?

※ diamonds.csv 파일 사용
※ Seed는 123으로 한다.

1. 2
2. 3
3. 4
4. 5

정답

2

제시된 다섯 개의 변수로 주성분 계산 후 주성분 학습용 객체에서 explained_variance_ratio_ 어트리뷰트를 기반으로 누적 분산비를 계산할 수 있다. 이 때 cumsum() 메서드를 활용하면 수월하게 계산할 수 있다.

Q3

가격을 예측하기 위해 기존 변수와 주성분 변수의 성능을 비교하고자 한다. 독립변수를 carat과 x를 둔 회귀모델 1번과 carat과 주성분 변수 한 개를 사용한 회귀모델 2번 중 성능이 더 좋은 모델의 RMSE는?

※ diamonds.csv 파일 사용
※ 데이터 세트는 학습8, 평가2로 분할한다.
※ 주성분은 x, y, z 변수로 생성한 첫 번째 주성분을 사용
※ Seed는 123으로 한다.

1. 1번, 1526
2. 1번, 1528
3. 2번, 1526
4. 2번, 1529

Q3

가격을 예측하기 위해 기존 변수와 주성분 변수의 성능을 비교하고자 한다. 독립변수를 carat과 x를 둔 회귀모델 1번과 carat과 주성분 변수 한 개를 사용한 회귀모델 2번 중 성능이 더 좋은 모델의 RMSE는?

※ diamonds.csv 파일 사용
※ 데이터 세트는 학습8, 평가2로 분할한다.
※ 주성분은 x, y, z 변수로 생성한 첫 번째 주성분을 사용
※ Seed는 123으로 한다.

- [x] 1번, 1526
- [ ] 2 1번, 1528
- [ ] 3 2번, 1526
- [ ] 4 2번, 1529

정답

1

각각의 모델을 별도로 만들어 계산하며 RMSE 계산 시 mean_squared_error() 함수를 활용한다. RMSE는 오차 평가 지표로 낮을수록 좋기 때문에 높은 값을 선택하지 않도록 한다.

#주성분 분석

#sklearn – PCA()

#fit_transform() 메서드로 주성분 연산

#explained_variance로
각 주성분 분산 파악

실전 종합 문제 1

Key words

#기술통계량

#가설검정

#군집분석

01 시나리오

서울시 소재의 DS 운수에서는 정기적으로
회사 경영진과 버스기사 간 연봉과 근무 처우 관련하여 협상을 한다.
올해부터 데이터 분석을 본격적으로 도입하기로 하였다.
그리하여 각자의 의견을 조율함에 있어 데이터 분석을 기반으로 보다
객관적인 의사결정을 하고자 한다.

02 데이터 설명

서울시 2019년 승차 정보 - Seoul_Bus_2019.csv

변수명	설명
Year_Month	연도와 월
Line_ID	노선 식별자
Line_No	노선 번호
Line_Name	노선 이름
Station_ID	정류소 식별자
Station_Name	정류소 이름
H01 ~ H24	각 시간대별 승차 인원

02 데이터 설명

서울시 버스 정보 - Seoul_Bus_info.csv

변수명	설명
Bus_no	버스 노선 번호
type	구분

03 문제

1번

지선, 간선 버스의 경우 노선당 년 수익이 10억 이하인 경우
배차간격 조정이나 노선 변경 등 수익구조 개선이 필요하다고 한다.
승객 1명당 기대 수익이 천 원이라고 했을 때 몇 개의 버스 노선이 대상인가? (정답 예시: 1)

03 문제

2번

간선 버스노선 버스기사들은 간선 버스노선이 지선 버스노선 대비 정류장 개수가 많아
버스 기사 확충 또는 배차 간격 조정을 사측에 요구하고 있다.
경영지원팀은 요구를 수용하기 위해 지선과 간선 노선의 정류장 개수를 파악하고
이 차이가 통계적으로 유의한지 확인해보려 한다.
간선 버스노선의 평균 정류장 개수와 지선 버스노선의 평균 정류장 개수의 평균을
적절한 검정을 통해 그 차이를 비교하고 그 검정통계량의 절대값을 소수점 둘 째 자리까지
반올림하여 기술하시오. (정답 예시: 1.23)

03 문제

3번

출퇴근 시간 배차간격 조정을 위해 우선적으로 각 정류소별 시간대별 출근 시간 승차 패턴을 파악하고자 한다. 지선 버스 노선을 대상으로 각 정류장별 승차인원을 계층적 군집분석을 활용하여 6개의 군집으로 분할하였을 때 출근시간대에 승차 인원이 가장 많은 정류소 군집의 번호는 몇 번인가? (정답 예시: 2)

※ 출근 시간대는 오전 7시 부터 9시 까지로 정의
※ 군집분석 시 자료는 Min-Max 정규화 실시
※ 거리 계산은 유클리디안 거리로 하고 유사도는 Ward.D 방법을 사용

실전 종합 문제 2

Key words

#분산분석

#회귀분석

01 시나리오

DS 금융의 올해 목표는 고객의 금융 서비스 이용 패턴 기반의 신규 서비스 런칭과 보다 객관적이고 투명한 고객 신용거래를 지원하기 위해 본격적으로 데이터 분석 기법을 도입하기로 하였다.
이를 위하여 1만명의 고객 데이터를 샘플로 하여 파일럿 프로젝트를 실시하기로 하였다.

02 데이터 설명

은행 고객 데이터 - financial_info_10k_persons.csv

변수명	설명
ID	고유 번호
is_attrited	이탈 여부(이탈: 1)
Age	나이
Gender	성별
Dependent_cnt	부양가족 수
Edu_level	교육 수준
Marital_status	결혼 상태
Income	수입
Card	카드 등급
Period_m	가입 기간(월)

02 데이터 설명

은행 고객 데이터 - financial_info_10k_persons.csv

변수명	설명
Total_rel_cnt	서비스 이용 횟수
Inactive_last_12m	최근 12개월동안 금융 거래가 없었던 기간(월)
Contact_cnt_last_12m	최근 12개월동안 영업점 방문 횟수
Credit_limit	신용 한도
Total_trans_amt	누적 송금액
Total_trans_cnt	누적 송금 횟수

03 문제

1번

고객의 총 송금액이 교육 수준, 혼인 여부에 따라서 어떤 특징을 보이는지 분산분석을 통해 알아보고자 한다. 총 송금액을 종속변수로 했을 때 독립변수간 교호작용 여부를 알아보고 해당 p-value를 반올림하여 소수점 둘 째 자리까지 기술하시오. (정답 예시: 0.12)

03 문제

2번

고객의 신용 한도는 다양한 정보를 기반으로 결정된다.
고객의 금융활동이 누적됨에 따라 신용 한도는 바뀌기도 하는데 이를 비교하고자 한다.
고객의 신용한도를 종속변수를 공통으로 하고 부양가족, 나이, 학력, 성별, 결혼여부를
1번 회귀 모델. 1번 모델에 가입기간과 누적 송금 횟수를 독립변수에 더한 회귀 모델의 결정계수
차이의 절대값을 반올림하여 소수점 셋 째 자리까지 기술하시오. (정답 예시: 0.123)

03 문제

3번

신규 고객이 개인정보를 입력할 경우 예상 신용 한도를 보여주려고 한다.
부양가족과 수입이 없는 29세 고졸(High School) 미혼의 남성의 경우 예상 신용 한도를
정수 부분만 기술하시오. (정답 예시: 1)

실전 종합 문제 3

Key words

#연관성 분석

01 시나리오

DS 마트의 경영부서는 기존의 매출 데이터를 기반으로 마케팅 팀에 전달한
집중 홍보 상품 목록 선정과 매장 매대 진열 및 소비자 동선 수정을 이번 달 목표로 잡았다.
이를 위해 매출 데이터, 상품 정보, 고객정보를 전산팀으로부터 인계 받아 분석을 준비하였다.
정제된 로그 데이터를 활용하여 다음의 분석을 실시하시오.

02 데이터 설명

마트 매출 데이터 - association_rules_mart.csv

변수명	설명
Date	판매일
ID	고객 식별자
Item	판매 품목

02 데이터 설명

고객 데이터 - association_rules_customers.csv

변수명	설명
ID	고객 식별자
Gender	성별(남자: M, 여자: F)
Age	나이

02 데이터 설명

품목 데이터 - association_rules_products.csv

변수명	설명
product	제품명
price	가격

03 문제

1번

2014년에는 매출이 발생했으나 2015년에는 매출이 발생하지 않은 품목은
총 몇 개 인가?(정답 예시: 1)

03 문제

2번

전해 12월의 매출은 차년도 매출과 꽤 연관이 깊다고 한다.
2014년도 12월 매출 상위 3개 품목을 확인하고 해당 품목의 2015년 매출 비중을 반올림하여 소수점 셋 째 자리까지 기술하시오. (정답 예시: 0.123)

03 문제

3번

남성과 여성의 상품 구매 성향이 다르다는 가정을 확인하기 위해서 2015년 데이터를 기반으로 연관규칙 분석을 실시하고 지지도가 0.05 이상인 규칙 중 향상도가 가장 높은 조건의 결과절(consequent) 품목을 남녀 차례대로 기술하시오.
(정답 예시: water, sugar)

※ 단, 구매 품목이 1건인 회원의 정보는 제외한다.
※ 최초 지지도와 신뢰도 설정은 0.005로 한다.