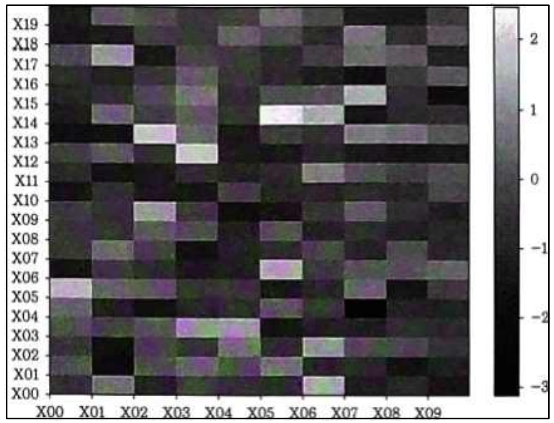
빅데이터 분석기사 기출 문제(5회 2022년 10월 1일)
4과목 빅데이터 결과해석

61. 다음 중 ROC 곡선 축을 구성하는 지표로 바르게 짝지어진 것은?
① 정확도, 정밀도
② 정밀도, 특이도
③ 특이도, 민감도
④ 재현율, f1-score
ROC 곡선의 x축은 거짓 긍정률(1-특이도)이고, y축은 참 긍정률(재현율, 민감도)이다.
ROC(Receiver Operating Characteristic Curve)
● ROC 곡선은 가로축(x)을 혼동행렬의 거짓 긍정률(FP Rate)로 두고, 세로축(y)을 참 긍정률(TP Rate, 재현율(Recall), 민감도(Sensitivity))로 두어 시각화한 그래프이다.
● ROC 곡선은 가능한 모든 임곗값(threshold)에 대한 거짓 긍정률(FP Rate)과 참 긍정률(TP Rate)의 비율을 표현한다.
● 그래프가 왼쪽 꼭대기에 가까울수록 분류 성능이 우수하다고 할 수 있다

62. 다음 중 교차 검증에 대한 설명으로 옳지 않은 것은?
① 훈련 데이터, 검증 데이터, 테스트 데이터의 비율은 보통 2 : 3 : 5의 비율로 구성된다.
② 홀드아웃 교차 검증은 데이터를 무작위로 7:3 또는 8:2 비율로, 학습 데이터와 검증 데이터로 나누는 방법이다.
③ 교차 검증은 과적합을 방지하기 위해 사용된다.
④ 데이터의 수가 적은 경우에 사용될 수 있다.
교차 검증(Cross Validation)에서 훈련 데이터는 검증, 테스트 데이터보다 많은 비율을 차지한다. 예를 들면 훈련 데이터 6, 검증 데이터 2, 테스트 데이터 2와 같이 검증할 수 있다.
교차 검증
● 교차 검증은 예측 모델의 정확도를 높이기 위해 데이터를 훈련, 평가 데이터로 나누어 여러 차례 검증하는 방법이다.
● 교차 검증의 목적은 과적합을 피하고, 매개변수를 튜닝하여 일반적인 모델을 만들고 더욱 신뢰성 있는 모델 평가를 하기 위해서이다.
● 데이터를 분할하여 일부는 분석 모형 학습에 사용하고, 나머지는 모델의 검증에 사용하는 방법을 여러 차례 반복 수행하고, 이를 통해 분석 모형이 새로운 데이터에 대해 일반화된 성능을 보일 수 있다.
홀드아웃 교차 검증
● 데이터를 무작위로 7:3 또는 8:2 비율로, 학습 데이터와 검증 데이터로 나누는 방법이다.
● 가장 보편적으로 랜덤 추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 60~80%이고, 테스트 데이터가 20~40%이다.

63. 다음 중 최종 모델을 평가하는 기준으로 옳지 않은 것은?
① 평가 지표
② 업무 관계자의 의견
③ 시스템 구현 가능성
④ 표본의 충분성
표본의 충분성은 분석 모델 개발 단계에서 고려되는 평가 기준이다.
최종 모형 선정 기준
● 분석 모형 개발 단계에서 구성한 여러 개의 분석 모델을 대상으로 실제 업무에 적용할 수 있는 최종 모형을 선정한다.
● 최종 모형 선정 절차는 최종 모형 평가 기준 선정, 최종 모형 분석 결과 검토, 알고리즘별 결과 비교 순이다.
① 최종 모형 평가 기준 선정
분석 모형 개발 후 분석 알고리즘 수행 결과를 검토하여 최종 모형을 선정한다. 정확도, 정밀도, 재현율 등의 평가 지표를 활용한다.
② 최종 모형 분석 결과 검토
최종 모형 선정 시 업무관계자(데이터 분석가, 데이터 처리자, 고객 등)의 리뷰를 종합하여 최적의 분석 모형을 선정한다.
③ 알고리즘별 결과 비교
분석 알고리즘에 따라 매개변수를 변경하여 결과를 비교하고, 이를 바탕으로 최종 모형을 선정한다.

64. 다음과 같은 차트가 나타내는 시각화 기법은?



① 트리맵 ② 카토그램

③ 히트맵 ④ 산점도 행렬

히트맵(Heat Map)

● 색상으로 표현할 수 있는 다양한 정보를 일정한 이미지 위에 열분포 형태로 표현한 그래프이다.

● 각 칸별 색상은 데이터 값의 크기를 나타내고, 색상이 짙을수록 데이터 값이 큰 것을 의미한다.

트리맵(Tree Map)

● 여러 계층 구조(트리 구조)의 데이터를 표현하는 그래프이다.

● 서로 다른 크기의 사각형을 이용하여 데이터의 비율을 나타내고, 사각형을 겹쳐 배치함으로써 데이터의 대분류와 하위분류를 구분한다.

카토그램(Catogram)

● 특정한 데이터 수치의 변화에 따라서 지도의 면적이 왜곡되는 그래프이다.

● 주로 의석 수나 선거인단 수, 인구 등의 데이터를 표현한다.

산점도 행렬(Scatter Plot Matrix)

● 산점도를 여러 개의 변수 조합별로 행렬의 형태로 표현한 그래프이다.

65. 다음 중 분류 모형 평가에 대한 설명으로 옳지 않은 것은?

① F1-Score는 정밀도와 재현율의 조화평균 값이다.

② 혼동행렬에서 모델이 참으로 예측한 수치는 TP+FP이다.

③ ROC Curve로 혼동행렬을 구할 수 있다.

④ AUC 값이 1에 가까울수록 모델의 분류 성능이 좋다.

혼동행렬의 요소인 재현율, 특이도가 ROC 곡선의 축을 이루지만, ROC 곡선 자체로는 혼동행렬을 구할 수 없다.

● F-Measure(F1-Score) : 0~1 사이의 범위를 가짐

● 공식 : $2 * \frac{Precision * Recall}{Precision + Recall}$

● AUC(Area Under the ROC Curve)는 진단의 정확도를 측정할 때 사용하는 것으로 ROC 곡선 아래의 면적을 모형의 평가 지표로 삼는다.

● AUC 값은 항상 0.5~1의 값을 가지며, 1에 가까울수록 좋은 모형이라고 평가한다.

66. 다음 중 분석 모형 평가 지표에 대한 수식으로 옳지 않은 것은?

① MAE : $\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$

② MSE : $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)$

③ MAPE : $\frac{100}{n}\times\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}}{y_i}\right|$

④ RMSE : $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

평균제곱근오차(MSE)의 수식은 $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 이다.

회귀모형 평가 지표

1. 평균절대오차(MAE)

● 모델의 실젯값과 예측값 사이에 절댓값을 취하여 평균한 값

● 수식 : $\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$

2. 평균제곱오차(MSE)

● 모델의 실젯값과 예측값 차이를 제곱하여 평균한 값

● 수식 : $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

3. 평균제곱근오차(RMSE)

● 평균제곱오차(MSE)에 제곱근을 씌운 값

● MSE는 값이 커지는 경향이 있으므로 제곱근을 씌운 RMSE를 실무에서 일반적으로 사용한다.

● 수식 : $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

4. 평균절대백분율오차(MAPE)

● 평균절대오차(MAE)를 퍼센트로 변환한 값

● 다른 변수 사이의 오차를 비교할 수 있다.

● 수식 : $\frac{100}{n} \times \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}}{y_i} \right|$

67. 다음 중 정규성 검정 기법으로 옳지 않은 것은?
① 샤피로-윌크 검정
② Q-Q plot
③ 콜모고로프-스미르노프 검정
④ 카이제곱 검정

정상성(정규성)
● 잔차항이 평균 0인 정규분포 형태를 이뤄야 한다.
● 샤피로-윌크 검정, 콜모고로프-스미르노프 검정을 통해 통계량 확인이 가능하다.
● Q-Q Plot에서 잔차가 오른쪽으로 치우친 직선 형태의 경우 정규성을 띈다고 할 수 있다.

샤피로-윌크 검정(Shapiro-Wilk Test)
● 데이터가 정규분포를 따르는지 확인하는 방법이다.
● R언어에서 sharpiro.test() 함수를 사용하여 검정하며, p-value가 0.05보다 작은 경우 귀무가설(H_0)을 기각하고, 대립가설(H_1)을 채택한다.
● 다만 언어의 sharpiro.test() 함수를 사용하는 경우 데이터의 수는 3개에서 5,000개 이하로만 사용 가능하다.

콜모고로프-스미르노프 적합성 검정(K-S 검정)
● 데이터의 누적 분포 함수와 비교하고자 하는 분포의 누적 분포 함수 간에 최대 거리를 통계량으로 사용하는 가설 검정 방법이다.
● R언어에서 ks.test() 함수를 사용하여 검정하며, p-value가 0.05보다 작은 경우 귀무가설(H_0)을 기각하고, 대립가설(H_1)을 채택한다.

Q-Q plot
● 그래프를 통해 정규성 가정을 시각적으로 검정하는 방법이다.
● 대각선 참조선을 따라서 데이터가 분포할 경우 정규성 가정을 만족한다고 할 수 있다.
● Q-Q plot 해석은 주관적일 수 있기 때문에 보조 수단으로 사용하는 것이 좋다.

카이제곱 검정
어떤 그룹이 서로 독립적인 아닌지 확인하는 방법으로 카이제곱 검정 유형으로 독립성 검정, 적합성 검정, 동질성 검정이 있다.

68. 다음 중 데이터 시각화에 대한 설명으로 옳지 않은 것은?
① 데이터 구조화 단계에서는 시각화를 위한 요건을 정의하고, 사용자에 따른 시나리오를 작성한다.
② 데이터 시각표현 단계에서는 데이터 모델링을 수행한다.
③ 데이터 시각화 단계에서는 여러 변수를 비교하여 분석 정보의 시각화를 구현한다.
④ 데이터 시각표현 단계에서는 데이터가 목적과 의도에 맞게 시각적으로 잘 표현되었는지 확인한다.

데이터 시각화 단계에서는 모델링 작업을 수행하지 않고, 이미 모델링된 분석 결과를 바탕으로 데이터를 시각화하여 표현하는 단계이다.

데이터 시각화 절차(프로세스)
구조화 -> 시각화(기본틀) -> 시각 표현(차트 완성)

69. 다음 중 비교 시각화 기법이 아닌 것은?
① 히스토그램 ② 스타차트
③ 플로팅 바 차트 ④ 체르노프페이스

히스토그램은 관계시각화 유형에 속한다.

시간 시각화 유형 : 막대그래프, 누적막대그래프, 점그래프, 선그래프, 영역차트, 계단식 그래프, 추세선
공간 시각화 유형 : 등치지역도, 등치선도, 도트맵, 버블맵, 카토그램
분포 시각화 유형 : 파이차트, 도넛차트, 트리맵, 누적영영그래프
관계 시각화 유형 : 산점도, 산점도 행렬, 버블 차트, 히스토그램, 네트워크 그래프
비교 시각화 유형 : 플로팅 바 차트, 히트맵, 체르노프페이스, 스타차트, 평행좌표 그래프

70. 다음 중 관계 시각화 기법이 아닌 것은?
① 산점도 행렬 ② 누적막대그래프
③ 네트워크그래프 ④ 버블차트

71. 민감도가 0.6, 정밀도가 0.4인 경우 F1-Score는 얼마인가?
① 0.24 ② 0.48
③ 0.5 ④ 0.6

민감도(Sensitivity, 재현율(Recall), 참 긍정률(TP Rate) : 실제 긍정 범주 중 긍정의 비율
정밀도(Precision) : 예측 긍정 범주 중 긍정의 비율

F-Measure(F1-Score) : $2 * \frac{정밀도(Precision) * Recall(재현율)}{정밀도(Precision) + Recall(재현율)} = 2 * \frac{0.4 * 0.6}{0.4 + 0.6}$

$= 2 * \frac{0.24}{1} = 0.48$

이 된다.

72. 다음 앙상블 모형에 대한 설명 중 옳은 것을 모두 고른 것은?

가. 랜덤 포레스트가 대표적인 앙상블 모델이다. 나. 배깅은 부트스트랩 샘플을 사용한다. 다. 부스팅은 정답에 더 높은 가중치를 적용하여 모델의 성능을 높이는 방법이다.

① 가 ② 가, 나, 다

③ 나, 다 ④ 가, 나

부스팅은 오답에 더 높은 가중치를 적용하여 모델의 성능을 높이는 방법이다.

부스팅(Boosting)

● 예측력이 약한 모형들을 결합하여 예측력이 강한 모형을 만드는 알고리즘으로 분류가 잘못된 데이터에 가중치를 적용하여 표본을 추출하는 기법이다.

● 대용량 데이터 분석에 유리하고, 높은 계산 복잡도를 가진다.

● 알고리즘 : AdaBoost, GBM, XGBoost

배깅(Bagging)

● 부트스트랩 샘플링으로 추출한 여러 개의 표본에 각각 모형을 병렬적으로 학습하고, 추출된 결과를 집계하는 기법이다.

● 배깅은 사이즈가 작거나 결측값이 있는 경우에 사용하기가 유리하고, 선능 향상에 효과적인 특징이 있다.

랜덤 포레스트(Random Forest)

● 의사결정나무 기반 앙상블 알고리즘으로 모든 속성(Feature)들에서 임의로 일부를 선택하고, 그 중에서 정보 획득량이 가장 높은 것을 기준으로 데이터를 분할한다.

● 분류기를 여러 개 사용할수록 성능이 좋아지고, 예측편향을 줄이고, 과대적합을 피할 수 있으며, 이상치에 영향을 적게 받는다.

73. 다음 중 신경망 모델에서 발생하는 기울기 소실 문제에 대한 설명으로 옳은 것은?

① 오차 역전파 과정에서 기울기가 감소하여 가중치가 업데이트되지 않은 현상을 말한다.

② 은닉층의 활성화 함수로 시그모이드 함수를 사용하면 문제가 완화된다.

③ 그래디언트 클리핑을 하면 문제가 완화된다.

④ 신경망 학습 과정에서 기울기가 점차 커지다가 발산하는 현상이다.

기울기 소실 문제를 해결하기 위해 은닉층의 활성화 함수로 ReLU함수, Leaky ReLU 함수 등을 사용한다. ③, ④는 모두 기울기 폭주 문제에 대한 설명이다.

기울기 소실(Gradient Vanishing) : 오차 역전파(Backprogation) 과정에서 입력층으로 갈수록 기울기가 점점 작아지는 현상을 의미한다.

시그모이드 함수(Sigmoid Function)

● 로지스틱 회귀 함수에 로짓 변환을 한 형태이다.

● 기울기 소실 문제의 원인이 된다.

기울기 클리핑은 기울기 폭주를 막기 위해 일정 임곗값을 넘지 못하도록 기울기 값을 자르는 방법이다.

74. 의사결정나무에 대한 설명으로 옳지 않은 것은?

① 가지분할은 의사결정나무에서 나무의 가지를 생성하는 과정이다.

② 의사결정나무의 해석이 어려운 이유는 계산 결과가 의사결정나무에 직접적으로 나타나지 않기 때문이다.

③ 의사결정나무는 전체 자료를 몇몇의 소집단으로 분류하거나 예측하는 방법이다.

④ 연속적으로 발생하는 의사결정 문제를 시각화해서 의사결정이 이루어지는 시점과 성과 파악이 쉽다.

의사결정나무(Decision Tree)

● 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내는 모형으로 그 모양이 나무와 비슷하여 의사결정나무라고 부른다.

● 의사결정나무의 해석이 용이한 이유는 계산 결과가 의사결정나무에 직접적으로 나타나기 때문이다.

● 가지 분할(Split)은 나무의 가지를 생성하는 과정이고, 가지치기(Prunning)는 생성된 가지를 잘라내어 모형을 단순화 시키는 과정이다.

● 의사결정나무는 뿌리 마디, 자식 마디(2개 이상의 마디들), 부모 마디, 끝마디, 중간 마디, 가지, 깊이로 구성이 된다.

75. 다음 중 재현율 공식으로 옳은 것은?

① $\frac{TP}{(TP+FN)}$ ② $\frac{FP}{(FP+TN)}$

③ $\frac{TP}{(TP+FP)}$ ④ $\frac{FN}{(FN+TN)}$

분류 모형 평가 지표

1. 정확도(Accuracy, 정 분류율)

● 전체 범주 중 정확히 예측한 비율

● 수식 : $\frac{TP+TN}{TP+TN+FP+FN}$

2. 오차 비율(Error Rate)

● 전체 범주 중 잘못 예측한 비율

● 수식 : $\frac{FP+FN}{TP+TN+FP+FN}$

3. 참 긍정률(TP Rate) = 재현율(Recall) = 민감도(Sensitivity)

● 실제 '긍정' 범주 중 '긍정'의 비율

● 수식 : $\frac{TP}{(TP+FN)}$

4. 특이도(Specificity)

● 실제 '부정' 범주 중 '부정'의 비율

● 수식 : $\frac{TN}{(TN+FP)}$

5. 거짓 긍정률(FP Rate)

● 실제 '부정' 범주 중 '긍정'의 비율

● 수식 : $\frac{FP}{(TN+FP)}$

6. 정밀도(Precision)

● 예측 '긍정' 범주 중 '긍정'의 비율

● 수식 : $\frac{TP}{(TP+FP)}$

7. F-Measure(F1-Score)

● 수식 : $2*\frac{\text{정밀도(Precision)}*Recall(\text{재현율})}{\text{정밀도(Precision)}+Recall(\text{재현율})}$

76. 다음 중 인포그래픽에 대한 설명으로 옳지 않은 것은?

① 인포그래픽은 정보 제공자가 전달하고자 하는 주요한 정보를 하나의 그래픽으로 표현하여 보는 사람들이 쉽고 빠르게 정보를 이해할 수 있도록 만든 시각화 방법이다.

② 인포그래픽은 정보를 SNS 상에 쉽고 빠르게 전달할 수 있다.

③ 전문 분야에 대한 데이터를 전달하기 위해서는 전문적 용어를 위주로 사용하여 표현한다.

④ 복잡한 데이터를 쉽게 이해할 수 있도록 그래픽과 텍스트를 적절하게 조합하여 표현한다.

인포그래픽(Inforgraphics)

● 인포그래픽은 정보(Information)와 그래픽(Graphics)의 합성어로 정보 제공자가 전달하고자 하는 주요한 정보를 하나의 그래픽으로 표현하여 보는 사람들이 쉽고 빠르게 정보를 이해할 수 있도록 만든 시각화 방법이다.

● 인포그래픽의 목적은 정보형 메시지와 설득형 메시지를 담는 것이다.

● 정보형 메시지는 전달하는 데이터에 정보가 담겨야 하는 것을 의미하고, 설득형 메시지는 정보 제공자가 주장하고자 하는 내용을 담는 것을 의미한다.

● 인포그래픽의 종류는 지도형, 도표형, 타임라인형, 스토리텔링형, 만화형, 비교분석형이 있다.

인포그래픽의 종류

1. 지도형

● 특정 국가 혹은 지역의 지도를 바탕으로 정보를 표현하는 방식

● 예) 지역별 투표율

2. 도표형

● 다양한 도표와 그래프를 사용하여 정보를 표현하는 방식

● 예) 신규 서비스 사용자 가입률과 기업 인지도 변화

3. 타임라인형

● 특정 주제와 관련된 히스토리를 타임라인(시간 순서) 형식으로 표현하는 방식

● 예) 기업의 발전 과정

4. 스토리텔링형

● 특정 사건 혹은 주제에 대한 정보를 이야기를 들려주듯이 표현하는 방식

● 예) 추석 차례상 차리는 방법

5. 만화형

● 캐릭터 또는 애니메이션 작업을 통해 정보를 표현하는 방식

● 예) 올바른 손씻기 방법

6. 비교 분석형

● 두 가지 이상의 비교 집단에 대한 정보를 비교하여 표현하는 방식

● 예) 가전 업체 A, B 기업의 노트북, TV 비교

77. 다음 중 특정 기준에 따라 회귀계수에 벌점을 부여하여 모형의 복잡도를 낮추는 분석 기법은?
① 랜덤 포레스트
② 벌점화 회귀
③ 로지스틱 회귀
④ 다항선형 회귀
특정 기준에 따라 회귀계수에 벌점을 부여하여 모형의 복잡도를 낮추는 분석 기법은 벌점화 회귀이다.
벌점화된 선택 기준
● 모형의 복잡도에 패널티(벌점)를 적용하는 방법으로 AIC 방법과 BIC 방법이 있다.
● 페널티 적용 대상 모델에 AIC 와 BIC를 계산하여 그 값이 최소가 되는 모델을 선택한다.
① AIC(Akaike Information Criterion)
● 실제 데이터의 분포와 모형이 예측하는 분포 간의 차이를 나타내는 방법이다.
● AIC 값이 낮다 는 것은 모형의 적합도가 높은 것을 의미한다.
② BIC(Bayesian Information Criterion)
● AIC는 표본이 커질수로 정확도가 낮아지는 단점이 있다. 이러한 단점을 보완하기 위한 방법이 BIC 이다.
● 표본의 크기와 상관없이 벌점이 일정한 AIC와 달리 BIC는 표본의 크기가 커질수록 복한 모형을 더욱 강하게 제한할 수 있다.

78. 다음 중 과대적합에 대한 설명으로 옳지 않은 것은?
① 비선형 모형은 선형 모형보다 과대적합 발생 가능성이 낮다.
② 과대적합은 모형이 과도하게 복잡해진 상태이다.
③ 과대적합 모형은 분산이 크다.
④ 과대적합 모형은 일반화 성능이 낮다.
과대적합(Over-Fitting)
● 학습 모델을 지나치게 복잡하게 학습하여 학습 데이터셋에서 모델 성능이 높지만, 새로운 데이터가 주어진 경우 정확도가 낮아지는 현상이다.
● 과대적합은 선형 모형보다 비선형 모형에서 발생 가능성이 높다.
● 과대적합 모형은 분산이 크다.
● 과대적합 모형은 일반화 성능이 낮다.

79. 다음 중 데이터 시각화 절차에 해당하는 요소가 아닌 것은?
① 구조화
② 시각화
③ 시각표현
④ 데이터 보충
데이터 시각화 절차(프로세스)
구조화 -> 시각화(기본틀) -> 시각 표현(차트 완성)

80. 다음 중 설명력이 가장 좋은 ROC 곡선은?
① AUC : -0.95
② AUC : 0.88
③ AUC : 0.77
④ AUC : 0.5
AUC(Area Under the ROC Curve)는 진단의 정확도를 측정할 때 사용하는 것으로 ROC 곡선 아래의 면적을 모형의 평가 지표로 삼는다.
AUC 값은 항상 0.5~1의 값을 가지며, 1에 가까울수록 좋은 모형이라고 평가한다.
따라서 위의 보기 중에서 1에 가장 가까운 값은 AUC가 0.88인 것이다.