비즈니스를 위한 데이터 마이닝

2020년도 가을학기 기말고사

시험 시간: 14시 00분 ~ 15시 40분

- 기말고사 문제 및 데이터는 과제란의 기말고사에 들어 있습니다.
- 다음 페이지부터 있는 문제를 풀어, Jupyter Notebook 또는 다양한 형태의 파일로 기말고사 과제에 제출하세요.
- 문제를 풀 때 필요한 가정이 있다면, 어떤 가정하에 문제를 풀었는지 기술하시오.

creditcard_sampled.csv 는 신용카드 부정거래 탐지 데이터이며, 각 컬럼은 아래와 같다.

변수	설명
Time	첫번째 거래로부터 소요된 시간(단위: 초)
V1, …, V28	개인 정보 문제로 인해 실제 변수가 아니고 차원 축소를 통해 변형된 변수
Amount	거래 금액
Class	1: 부정 거래, 0: 정상 거래

1. (30점) 신용카드 부정거래 분류 모형 생성을 위해 차원 축소 후, 의사결정나무를 통해 학습하고자 한다.

(1) Class를 목표변수로 하고, 나머지 모든 변수를 입력변수로 활용한다. entropy를 불순도 지표로 사용하여 의사결정나무를 생성하고 5겹 교차 검증을 통해 성과를 측정하려고 한다. 성과지표는 f1을 활용하며, max_depth 값을 2부터 15까지 변화시켜가며 측정하고 결과를 시각화하라.

(2) Time, Amount, Class를 제외한 V1, …, V28 변수를 차원축소하고자 한다. 95% 분산이 보존되게끔 차원의 수를 축소하여라.

(3) (2)에서 차원 축소된 결과를 데이터프레임으로 변환하고 Time, Amount 컬럼을 추가하라. Class를 목표변수로 하고, 차원 축소된 결과와 Time, Amount 변수를 입력변수로 활용한다. entropy를 불순도 지표로 사용하여 의사결정나무를 생성하고 5겹 교차 검증을 통해 성과를 측정하려고 한다. 성과지표는 f1을 활용하며, max_depth 값을 2부터 15까지 변화시켜가며 측정하고 결과를 시각화하라.

(4) 차원 축소 전 최고 성과와 차원 축소 후 최고 성과를 적고, 어느 값이 큰 지 판단하라.

Carseats.csv 는 지역별 카시트 판매량 데이터이며, 각 컬럼은 다음과 같다.

변수	설명
Sales	지역 판매량 (단위: $1,000)
CompPrice	경쟁업체 제품 가격
Income	지역 소득수준 (단위: $1,000)
Advertising	지역 광고비 (단위: $1,000)
Population	지역 인구 수 (단위: 1,000명)
Price	자사 제품 가격
ShelveLoc	매대 위치 (범주형: Bad, Medium, Good)
Age	지역 주민 평균 나이
Education	지역 교육 수준
Urban	매장 위치가 도시인 지 여부 (Yes/No)
US	매장 위치가 미국인 지 여부 (Yes/No)

2. (30점) 아래와 같이 군집화를 수행하여 비슷한 성격을 띄는 지역의 군집을 찾고자 한다.

(1) ShelveLoc 변수는 Bad는 0, Medium은 0.5, Good은 1로 변환하라. Urban과 US는 군집화에 사용하지 않는다.

(2) 모든 변수를 MinMaxScaler를 사용하여 표준화하라.

(3) 가우시안 혼합모델을 사용하여 군집화를 하고자 한다. n_init = 10으로 하고, 군집의 수는 2부터 20까지 늘려가면서 군집화를 수행한 후 군집 수에 따른 BIC, AIC 값을 시각화하라. BIC, AIC 값이 최소가 되는 군집의 수나 엘보우 방안으로 최적의 군집의 수를 찾아라.

(4) (3)에서 찾아진 최적의 군집 수를 활용하여 가우시안 혼합 모델 군집화를 수행한다. 각 군집에 속한 데이터의 수와 각 군집의 변수별 평균 값을 출력하고 군집 특성에 대한 설명을 기술하라.

reviewContent_sampled.csv는 식당에 대한 리뷰 데이터이다.

user_id : 사용자 아이디

prod_id : 식당 아이디

date : 리뷰 작성 날짜

review : 리뷰 내용

3. (30점) LDA를 통해 리뷰에 포함된 토픽을 찾아보자.

(1) review를 가지고 LDA를 수행한다. 토픽 수는 3으로 하고 나머지는 다음과 같이 인수를 입력한다. 단어문서 매트릭스는 CountVectorizer를 사용하며, stemming이나 표제어 추출은 시행하지 않는다.

```
lda_model = LatentDirichletAllocation(
    n_components=NUM_TOPICS,
    max_iter=10,
    learning_method='online',
    random_state=100,
    batch_size=128,
    evaluate_every = -1,
    n_jobs = -1,
)

vectorizer = CountVectorizer(
    analyzer='word',
    min_df=5,
    stop_words='english',
    lowercase=True,
    token_pattern='[a-zA-Z0-9]{2,}'
)
```

(2) (1)에서 수행한 LDA 결과를 pyLDAvis 라이브러리를 활용하여 시각화하고, 각 토픽에 적당한 토픽명을 붙여라.

4. 이 수업에서 개선되어야 하는 사항은 무엇인가? (작성 시 10점, 미작성 시 0점)