비즈니스를 위한 데이터 마이닝

2020년도 가을학기 중간고사

시험 시간: 14시 00분 ~ 15시 50분

- 중간고사 문제 및 데이터는 과제란의 중간고사에 들어 있습니다.
- 다음 페이지부터 있는 문제를 풀어, Jupyter Notebook 또는 다양한 형태의 파일로 중간고사 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 의사결정나무를 통해 입력 변수를 찾고, 랜덤 포레스트를 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) 가장 성과가 좋은 max_depth 값일 때 의사결정나무 모형의 입력변수 중요도를 시각화하라.

(3) (2)에서 보여진 중요도 값에서 중요도가 가장 높은 6개의 변수만 입력변수로 사용하여 랜덤 포레스트 모형을 생성한다. 랜덤 포레스트 모형에서 생성하는 의사결정나무의 숫자는 200으로하고, 사용하는 변수는 최대 3개, 사용하는 데이터는 최대 40%로 하며 한 부분집합에 같은 데이터가 중복으로 들어가는 것을 허용한다. 랜덤 포레스트 모형의 성과는 5겹 교차검증으로 측정하며 성과지표는 AUC를 사용하라. 5겹 교차 검증의 평균 AUC 값을 보고하라.

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) LinearSVR, SVR(kernel='rbf'), GradientBoostingRegressor(max_depth = 3, learning_rate = 0.5) 세 가지 알고리즘을 활용하여 각각 판매량 예측 모형을 만들려고 한다.

(1) 학습에 필요한 전처리 과정이 있다면 수행하고, 어떤 과정을 거쳤는 지를 기술하시오.

(2) 각 모형의 최적 모형을 하이퍼 파라미터 튜닝을 통해 찾아라. 위에 이미 주어진 값은 활용하여야 하며, 주어지지 않은 값들을 하이퍼파라미터 튜닝을 통해 최적값을 찾아야 한다. GradientBoostingRegressor는 최적 n_estimators 를 찾아야 한다.

(3) 하이퍼파라미터 튜닝을 통해 찾아진 최적의 파라미터를 활용하여 세 모형의 예측 성과를 산출하고 각 모형의 성과를 비교하라. 성과는 5겹 교차 검증을 통해 산출하고 성과지표는 RMSE를 사용한다.

breast-cancer.csv 는 양성과 악성 종양 여부를 알려주는 class 컬럼과 다른 속성 컬럼들을 포함하고 있다. 각 컬럼은 다음과 같다.

code: id number
clump: Clump Thickness (1 – 10)
cell_size: Uniformity of Cell Size (1 – 10)
cell_shape: Uniformity of Cell Shape (1 – 10)
adhesion: Marginal Adhesion (1 – 10)
single: Single Epithelial Cell Size (1 – 10)
nuclei: Bare Nuclei (1 – 10)
chromatin: Bland Chromatin (1 – 10)
nucleoli: Normal Nucleoli (1 – 10)
mitoses: Mitoses (1 – 10)
class: 2 for benign, 4 for malignant (2 양성 종양, 4 악성 종양)

3. (30점) 양성 종양과 악성 종양을 분류하는 간접 투표 모형을 만들고자 한다.

(1) 직접 투표 모형 생성에 활용되는 분류 모형은 GaussianNB, SVC, LinearSVC, DecisionTreeClassifier, LogisticRegression를 각각 Bagging 하여 만들자. 각 모형의 인수는 모두 default 값을 활용하여 200개의 모형을 만들자. 사용되는 최대 변수의 수는 4로 지정하고, 최대 데이터의 수는 100으로 설정하자. 한 데이터가 하나의 부분 데이터 중복하여 들어가는 것은 허락하지 않는다.

(2) (1)에서 만들어진 다섯 개의 BaggingClassifier를 활용하여 간접 투표 방식으로 결과를 분류하는 VotingClassifier 를 만들고 성과를 측정하라. 성과는 5겹 교차 검증을 활용하며, AUC 지표를 활용하여 측정한다.

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