2018년 1학기 IT 집중교육 1

Homework 2

\* 코드 및 분석 내용을 copy할 경우 0점 처리

\* 제출기한: 3/18(일) 11:59 p.m.

\* ipynb파일을 아주BB에 제출

**[1] 데이터 분석과제**

\* ipynb 파일을 제출하는 경우 주석을 이용하여 문제에 대해 간단하게 설명하고 코드에 대한 설명을 포함해야 한다.

\* 출처: https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29

\* sklearn 패키지 이용 가능

1. **데이터 분석(1):** COIL2000 데이터는 보험가입자들의 개인정보와 가입된 보험에 대한 정보를 포함하는 데이터이다(데이터의 자세한 설명은 TicDataDescr.txt 참조). 이번 과제의 목표는 이 데이터의 label(Attribute 86, "CARAVAN:Number of mobile home policies")을 예측하는 logistic regression 모델을 만들고, 높은 예측도를 나타내도록 이 모델을 최적화하는 것이다. ticdata2000.csv 파일을 이용하여 logistic regression 모델을 학습시킨 뒤, test accuracy를 구해본다.
   1. 제공된 ticdata2000.csv 파일을 읽고 5개의 feature를 뽑아 EDA를 수행한다(이 과정에서 기술 통계량, 시각화(Plot)가 수행되어야 한다. 그리고 5개의 feature는 임의로 선택해도 되고 통계 수행을 통한 사전지식을 이용하여 선택해도 된다).
   2. 모든 feature에 대해 correlation coefficient array를 구하고 이를 heatmap으로 그려본다.
   3. B에서 구한 correlation coefficient를 바탕으로 multiple logistic regression 모델에서 사용할 feature를 정하고 regression 모델 학습 후 training accuracy를 출력한다(test data를 제공하지 않으니 validation data를 사용하여 regression 모델을 최적화시킨다).
   4. 제공된 submission\_format.csv과 같은 형식으로 ticeval2000.csv(test data)의 결과 label(Attribute 86, "CARAVAN:Number of mobile home policies")을 예측하는 파일을 homework2\_result.csv로 제출한다.
2. **데이터 분석(2)** 
   1. 수업 시간에 배운 feature engineering 기법을 이용하여 데이터 분석(1)에 대해 feature selection/engineering 진행 (최종적으로 몇 개의 feature를 왜 선정했는지 보고서에 기술)
   2. 2-A에서 찾은 feature들을 이용하여 logistic regression 모델을 학습시킨다. 이 때 최적의 regularization term을 추가하고 regularization strength를 조절하여 regression 모델 최적화
3. **보고서 작성**
   1. 데이터 분석(1)에서 feature selection에 사용한 알고리즘을 간단히 설명하고 몇 개의 feature를 왜 사용했는지 기술
   2. 데이터 분석(2)에서 regularization term을 찾는 과정에 대해 간단히 설명하고 training accuracy와 test accuracy를 기술한다.
   3. 위에서 분석한 내용을 토대로 데이터 분석 결과의 해석 진행

**[2] 최적화 과제**

\* 최종 test accuracy를 기준으로 상대평가 점수 부여.

1. 2주간 배운 모든 내용을 활용하여 Homework 1의 데이터 분석과제에서 사용한 Autouniv 데이터의 classification accuracy를 최적화한다 (수업 때 다루지 않은 알고리즘 사용가능).
2. 최적화하기 위해 사용한 알고리즘, 모델 설계 방법들을 보고서에 기술한다.