**Machine Learning & Data Mining, Spring 2020**

**Homework 6**

Due May 7

**과제목표**

* **수업시간에 배운 linear regression과 logistic regression 모델을 실제로 적용시켜 본다.**
* **Regression 모델에서 최적화하기 위한 variable selection 모델을 구현하고 적용한다.**

**제출**

* **요구한대로 작성한 보고서와 주피터 노트북(주석 필수)**
* **최종 점수는 비율적으로 수정해서 들어갈 예정**

**Linear regression용 데이터**

캘리포니아 주택가격 데이터셋을 사용, y값은 주택중앙값을 의미.  **homework 2 (assignment 1)** 과제에 올려진 주피터 노트북의 데이터 로드 방식을 따라 과제를 진행.

**Logistic regression용 데이터**

주어진 주피터 노트북에서 make\_classification 메서드로 생성된 데이터를 이용하여 진행, 데이터 생성 부분 코드의 경우 학생이 수정하면 안됨.

**scikit-learn 라이브러리를 바탕으로 과제를 진행한다**. scikit-learn 라이브러리에 대한 매뉴얼 홈페이지는 다음과 같다.

<https://scikit-learn.org/stable/>

1. **Coefficient interpretation (logistic regression)**
   1. 주어진 데이터의 모든 feature에 대해서 correlation coefficient array를 구하고 이 중 correlation coefficient 가 높은 5개의 feature를 이용해 feature간의 correlation을 **heatmap으로 그린다.** [5]
   2. 위에서 구한 correlation coefficient array를 바탕으로 multiple **logistic** regression 모델에서 사용할 feature를 정하고 90%의 데이터를 학습데이터로 사용하여 regression 모델 학습 후 10% 테스트 데이터에 대한 **test acc를 출력**한다. [5]
2. **Variable selection (linear regression)**
   1. Linear regression 모델의 variable 선택을 통한 모델을 비교하기 위해서 수업시간에 배운 Adjusted 을 기반으로 forward selection을 수행한 후 **test 를 출력하고** x축은 선택한 variable의 개수, y축은 테스트 데이터의 Adjusted 과 학습데이터의 Adjusted 를 각각 다른 색으로 **그래프를 이용해 출력한다**.[10]
      * forward selection 을 이용하여 variable을 4개 까지 선택한다. (for 문 이용)
      * sklearn의 r2\_score 함수에 기반하여 Adjusted 를 만든다.
3. **Regularized linear model (linear regression)**
   1. Regularization 방법 중 하나인 lasso regression을 이용하여 5-fold cross validation을 통해 최적 하이퍼파라미터 탐색 후, 최종 모델의 test 를 출력한다. [10]
   * Scikit learn의 라이브러리를 기반하여 수행한다.

\* 주어진 모든 문제에 대해서 전체의 10% 데이터를 test set으로 두어 이 데이터에서의 성능을 test accuracy/로 출력한

\* lasso in sklearn : linear\_model.Lasso