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

**Homework 4**

Due April 23

**과제목표**

* **수업시간에 배운 classification 모델들을 실제 적용시켜본다.**
* **데이터셋에 따라서 다양한 모델을 적용해야 한다는 것을 실제로 확인한다.**
* **다양한 모델과 파라미터 선택을 통해서 성능을 높이고 이에 대해 이해한다.**

**제출**

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

**와인 데이터란** 와인의 등급을 분류하기 위한 목적으로 만들어진 데이터 집합. 해당 과제에 올려진 주피터 노트북의 데이터 로드 방식을 따라 과제를 진행.

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

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

**[1] 수업시간에 배운 알고리즘들의 파라미터를 최적화[10점]**

해당 데이터를 바탕으로 수업시간에 배운 분류 모델을 적용하고 최적의 파라미터를 찾아 본다. 주피터 노트북으로 해당 모델들을 적용시킨다. 노트북 제출시 주석을 달아 설명한다.

훈련과 검증을 위한 데이터 샘플 split 방식은 다음과 같다.

* 총 주어진 샘플 중 80% 를 training 데이터로 구성하여 훈련시킨다.
* 10% 를 validation 데이터로 구성하여 하이퍼파라미터를 최적화 시킨다.
* 최종적으로 남은 10%를 test 데이터로 구성하여 성능을 보고서와 주피터 노트북에 표시한다.

보고서 작성 요령

각각의 알고리즘에 대해서 정확도가 가장 높은 하이퍼파라미터를 찾는 과정에 대해 설명하고 해당 결과에 대해 EDA한 시각화 자료와 통계량 등을 이용하여 해석한다.

적용해야 할 알고리즘과 최소한의 하이퍼파라미터 탐색의 종류

1. Knn
   1. 수업시간에 배운 metric 들 중 최소 두 개 비교
   2. 이웃의 개수 k 를 최소 3개 비교
2. Decision Tree
   1. gini 와 entropy criterion의 비교
3. Ensemble (decision tree-based)
   1. bagging과 random forest간의 비교
4. SVM
   1. linear SVM과 kernel trick이 적용된 SVM간의 비교 (kernel 의 종류는 상관 없음)

**[2] Cross-validation 을 이용한 최적화된 알고리즘 탐색 [10]**

[1] 에서 정의한 validation set을 이용하여 모델 선택을 하는 것이 아닌, 수업시간에 배운 두가지 cross-validation 방법인 k-fold cv와 loocv를 이용하여 최적화된 하이퍼파라미터와 모델을 찾아낸다.

최적화 시킬 모델은 [1]을 기반으로 하며, 가장 좋은 성능을 가진 모델과 하이퍼파라미터를 기반으로 보고서를 작성한다.

최종적인 성능은 [1]과 동일한 10%의 test set에서 측정하고, 나머지 90%를 cross-validation을 이용하여 모델 최적화를 진행한다.

보고서 작성 요령

cross-validation을 적용하여 정확도가 가장 높은 하이퍼파라미터와 모델을 찾는 과정에 대해 설명하고 해당 결과에 대해 EDA한 시각화 자료와 통계량 등을 이용하여 해석한다.