**2020 가을학기 기계학습 기말과제(수정)**

**1. 개요**

* 본인이 원하는 데이터를 이용하여 Classification(or Regression) Model 설계
* 필요한 경우 데이터를 가공하거나 수정하여도 무관하다.
* 가능한 많은 모델을 시도하여 볼 것(최소 4가지 이상의 classification or regression모델+1개 이상의 clustering 모델). 수업시간에 다루지 않은 모델 또한 사용 가능.
* 자신이 선택한 Dataset에 대한 **최대한의 모든 분석**을 해보도록 한다.

**2. Dataset 제한 조건**

Public Data Repository에서 데이터를 자유롭게 선택하거나, Machine learning competition 등에서 제공하는 데이터 등을 사용

**Public Dataset 참고 사이트**

* UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)
* KD Nugget (<https://www.kdnuggets.com/datasets/index.html>)
* Kaggle (<https://www.kaggle.com/>)
* Google Dataset Search (<https://toolbox.google.com/datasetsearch>) 등등

**제한 조건**

* 최소 10개 이상의 feature(attribute)
* 최소 5000개 이상의 instance

**참고용 Sample Dataset**

* Adult Data set (<https://archive.ics.uci.edu/ml/datasets/Adult>) : 14개의 feature, 약 5만개의 instance
* Mushroom Data set (<https://archive.ics.uci.edu/ml/datasets/Mushroom>) : 22개의 feature, 약 8천개의 instance
* Video Game Sales (<https://www.kaggle.com/gregorut/videogamesales>) : 11개의 feature, 약 1만 6천개의 instance
* Mobile App Store (7200 apps) (<https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps>) : 17개의 feature, 약 7천개의 instance

**3. Tools**

기계학습 도구

* python 기반의 Jupyter notebook 환경(권장) : scikit-learn 라이브러리(권장), 그 외 python에서 사용 가능한 기계학습 라이브러리 사용 가능, 본인이 직접 구현한 기계학습 코드 사용 가능
* GUI기반의 기계학습 도구(Weka, Orange, RapidMiner 등등) 사용 금지

**4. 실험 진행 및 보고서 내용**

* 실험 내용에 대한 전체 요약
* 선정한 데이터에 대한 설명 및 데이터 URL
* 실험 설계 및 방법(진행 내용은 최대한 구체적으로 작성바람.)
  + 자신이 선정한 모델 및 모델 선정에 대한 이유
  + 본인이 선택한 데이터에 대하여 전처리 등이 필요한 경우 normalization, regularization, discretization, missing value handling 등의 다양한 전처리를 적용해도 됨
  + Feature selection 기법을 1개 이상 적용해야함.
  + 분석에 시도한 trial 및 error에 대한 설명 (Default 모델만 사용하는 것이 아니라, **모델의 parameter 설정에 다양한 변화를 주는 과정**이 포함되어 있어야 함. 그리고 **validation을 통해 최적의 모델을 선정하는 과정**이 포함되어 있어야 함.)
  + 데이터셋에 대해 clustering을 적용시켜 군집을 생성해보고, 생성된 군집과 classification의 label(class)을 비교해보는 내용이 포함되어야 함.
* 실험 결과 분석
  + Data set을 train-set, validation-set, test-set으로 구분하여 사용하거나 또는 train-set, test-set으로 구분하고 validation 과정은cross-validation을 사용해 검증해야 함. **최종적인 모델 평가는 반드시 test-set을 통해 평가.**
  + 평가에 사용된 모델의 test set에 대한 평가 결과를 비교하고, 평가 결과에 차이가 있는 경우, 모델의 결과를 분석할 수 있는 경우 모델을 분석하고, 그렇지 않은 모델인 경우 데이터셋의 특성으로 유추해서 작성해야 함.
  + Feature selection 기법을 적용했을 때와 적용하지 않았을 때의 성능 차이를 비교하고, 차이에 대한 원인을 분석 또는 유추하여 작성하여야 함.
  + 실험 설계 및 평가 분석에 **overfitting의 관점**이 포함되어야 있어야 함.
* 결론

**5. 제출물**

보고서 및 프로그램 코드 제출

* 사용한 모델 및 코드는 모두 채점할 때 실행 가능해야함.
* jupyter Notebook 환경에서 작업 시 ipynb 파일에 보고서 작성(markdown으로 작성) 및 프로그램 코드를 동시에 제출 가능.
* 기타 환경에서 작업 시, 프로그램 코드 및 보고서를 압축하여 제출.(보고서 양식은 따로 없음)

**6. 과제 인터뷰**

* 80명의 모든 학생을 한 명씩 인터뷰하기에는 시간이 촉박하기 때문에, 제한 시간을 주고, 해당 시간 내 과제에 대한 질문에 대한 답변 동영상을 업로드하는 방식으로 진행할 예정.
* 추가적인 인터뷰가 필요하다고 판단되는 경우 추가 인터뷰 있을 수도 있음.

**7. 과제 마감**

* 2020년 12월 20일(일) 오후 11시 59분까지 블랙보드에 제출

**8. Help Session**

* 데이터 셋 선정에 어려움을 겪고 있거나, 과제 진행에 궁금한 점이 있거나, 진행 상황에 대해 상담이나 논의가 필요한 경우 등등, 도움이 필요한 경우 조교 이메일([koreaAIassistant@gmail.com](mailto:koreaAIassistant@gmail.com))에게 메일을 보내서 해결할 수 있도록 하기 바람.