2018 가을학기 기계학습 기말과제

1. 개요

- 본인이 원하는 데이터를 이용하여 Classfication 혹은 Regression Model 설계
- 데이터를 가공하고, 수정하여도 무관하다.
- 가능한 많은 모델을 시도하여 볼 것(최소 4 가지 이상의 모델, 딥러닝 라이브러리를 사용한 모델이 최소한 1 가지 이상은 포함되어야함). 수업시간에 다루지 않은 모델 또한 사용 가능.
- 자신이 선택한 Dataset 에 대한 최대한의 모든 분석을 해보도록 한다.

2. Dataset 제한 조건

Public Data Repository 에서 데이터를 자유롭게 선택하거나, Machine learning cup 대회 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 10 개 이상의 feature(attribute)
- 최소 5000 개 이상의 instance
- 금지 : KDD Cup 1999 Dataset(작년에 과제로 사용된 Dataset)

참고용 Sample Dataset

- Adult Data set (https://archive.ics.uci.edu/ml/datasets/Adult) : 14 개의 feature, 약 5 만개의 instance
- Mushroom Data set (https://archive.ics.uci.edu/ml/datasets/Mushroom) : 22 개의 feature, 약 8 천개의 instance
- Video Game Sales (https://www.kaggle.com/gregorut/videogamesales) : 11 개의 feature, 약 1 만 6 천개의 instance

- Mobile App Store (7200 apps) (https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps) : 17 개의 feature, 약 7 천개의 instance

등등

3. Tools

기계학습 도구

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

딥러닝 도구

- keras(권장)
- Tensorflow, pytorch, theano 등의 python 기반 라이브러리 또는 caffe, DeepLearning4J 등의 타언어 라이브러리 또한 사용 가능

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
    - 분석에 시도한 trial 및 error 에 대한 설명(초기 모델에서 기대 이하의 성능이 나올 경우, 모델의 설정에 다양한 변화를 주고 validation 을 통해 개선되는 점을 보여줄 것)
    - Data set 을 train-set, validation-set, test-set 으로 구분하여 사용하며, 최종 모델 평가는 test-set 을 통해 평가
    - 실험 설계 및 평가 분석에 overfitting 의 관점이 사용되어야 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(markdown 으로 작성) 및 프로그램 코드를 동시에 제출 가능.
- 기타 환경에서 작업 시, 프로그램 코드 및 보고서를 압축하여 제출.(보고서 양식은 따로 없음)

6. 과제 마감

- 2018 년 12 월 23 일(일) 오후 11 시 59 분까지 블랙보드에 제출

7. Help Session

- 라이시움 407 호에서 기계학습 분석 튜토리얼 열람 가능, 사진 촬영은 금지 (기본 열람 가능 시간 : 12 월 1 일/8 일/15 일 오후 2 시~6 시, 불가능할 경우 조교에게 메일을 보내서 별도로 시간을 잡을 것)
- 기말고사 직전 수업시간에 help session 을 통해 질의응답을 갖는 시간을 가질 예정
- 데이터 셋 선정에 어려움을 겪고 있거나, 과제 진행에 궁금한 점이 있거나, 진행 상황에 대해 상담이나 논의가 필요한 경우 등등, 도움이 필요한 경우 조교 이은헌(booksky@korea.ac.kr)에게 메일을 보내서 해결할 수 있도록 하기 바람.