TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Final Assignment 2020 CS7CS4/CSU44061 Machine Learning

Downloading Dataset

- Download the assignment dataset from `https://www.scss.tcd.ie/Doug.Leith/CSU44061/2020/final.php`. Important: You must fetch your own copy of the dataset, do not use the dataset downloaded by someone else.
- The data file consists of lines of json content. Each line contains data for one user review gathered from the Steam online gaming platform. To read in the data you can use:

```
import json_lines
X=[]; y=[]; z=[]
with open('reviews_0.jl', 'rb') as f:
    for item in json_lines.reader(f):
        X.append(item['text'])
        y.append(item['voted_up'])
        z.append(item['early_access'])
```

The 'text' field in the json gives the review text, the 'voted_up' field is true when the reviewer recommends the game and otherwise 'false', the 'early_access' field is true when the review is for a beta version of the game (i.e. before full release).

Assignment

1. Write a short report evaluating whether the review text can be used to (i) predict the review polarity (where a game has been "voted up" or not by the reviewer) and (ii) predict whether the review is for an "early access" version of a game or not. Hint: Select two or three machine learning approaches, apply them to the dataset and critically evaluate their classification performance. Remember its v important to clearly explain/justify any design choices that you make and any conclusions you arrive at. Include any code you use in an appendix. [75 marks]
2. (i) What are under-fitting and over-fitting, give an example of each. [5 marks]

   (ii) Give pseudo-code implementing k-fold cross-validation. [5 marks]

   (iii) Why does k-fold cross-validation provide a way to select a model hyperparameter so as to strike a balance between over/under-fitting. [5 marks]

   (iv) Discuss three pros and cons of a logistic regression classifier vs a kNN classifier. [5 marks]

   v) Give two examples of situations when a kNN classifier would give inaccurate predictions. Explain your reasoning. [5 marks]