MoA Prediction

한현수

1. Problem

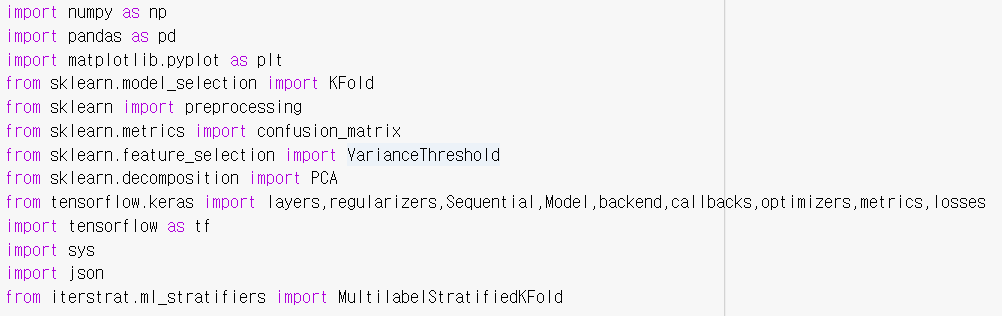
In this competition, you will have access to a unique dataset that combines gene expression and cell viability data. The data is based on a new technology that measures simultaneously (within the same samples) human cells’ responses to drugs in a pool of 100 different cell types (thus solving the problem of identifying ex-ante, which cell types are better suited for a given drug). In addition, you will have access to MoA annotations for more than 5,000 drugs in this dataset.

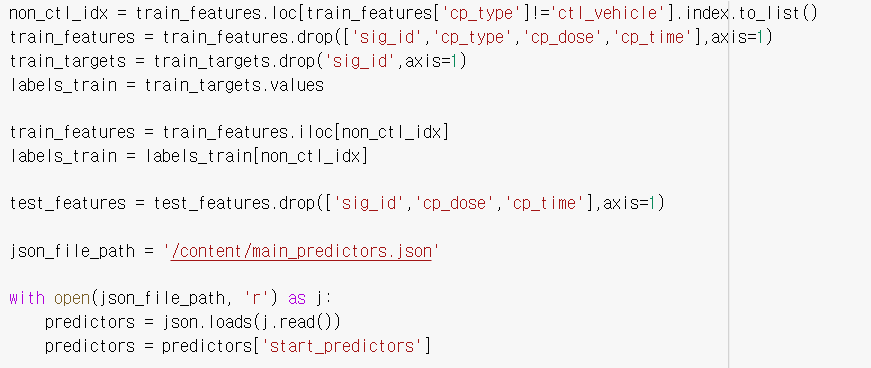
As is customary, the dataset has been split into testing and training subsets. Hence, your task is to use the training dataset to develop an algorithm that automatically labels each case in the test set as one or more MoA classes. Note that since drugs can have multiple MoA annotations, the task is formally a multi-label classification problem.

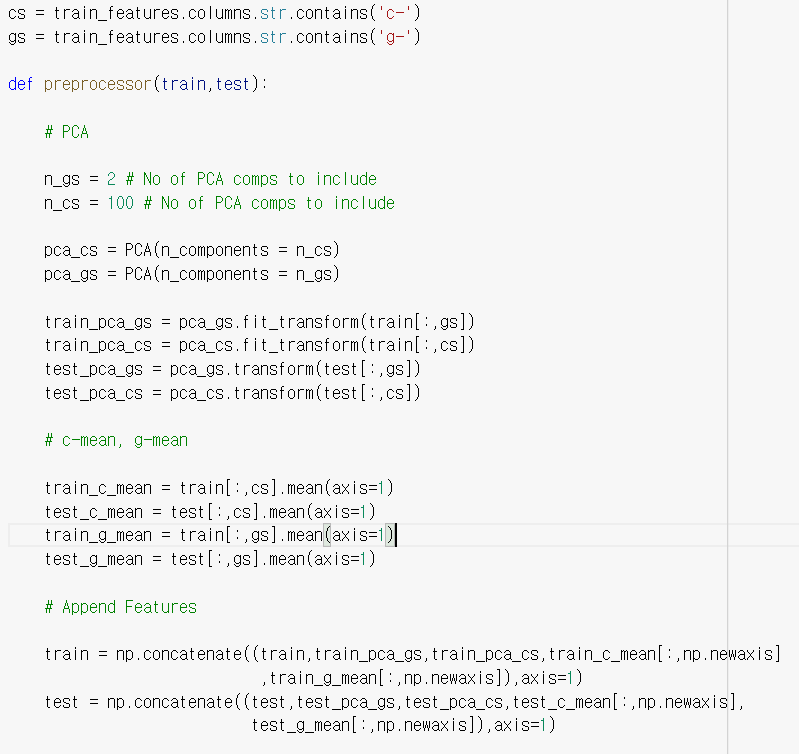
1. Data Fields

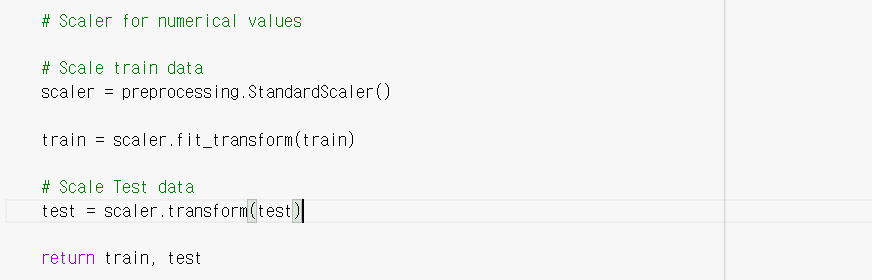
* g-feature: 유전자 발현 데이터
* c-feature: 세포 생존 가능성(능력) 데이터
* cp\_type: 화합물로 처리된 표본 (cp\_type이 ctl\_vehicle일 때 MoAs는 항상 0 (어떻게 처리할지, 행을 삭제할지 생각해야함))
* cp\_time: 치료기간 (24시간, 48시간, 72시간)
* cp\_dose: 복용량

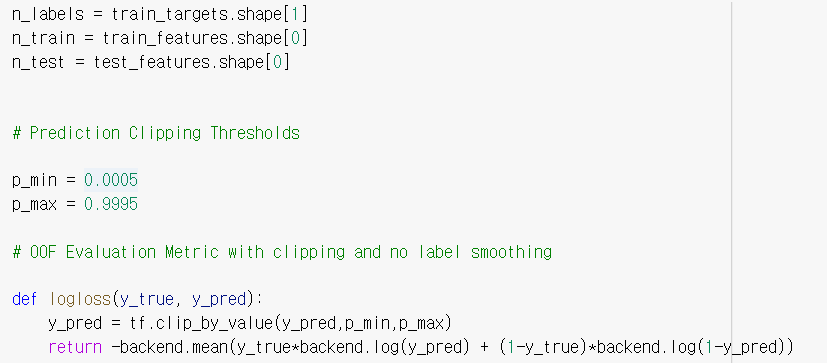
2. Kaggle 필사

Import packge

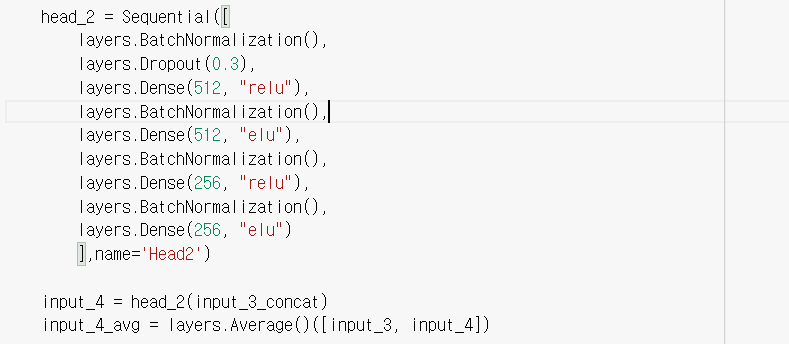
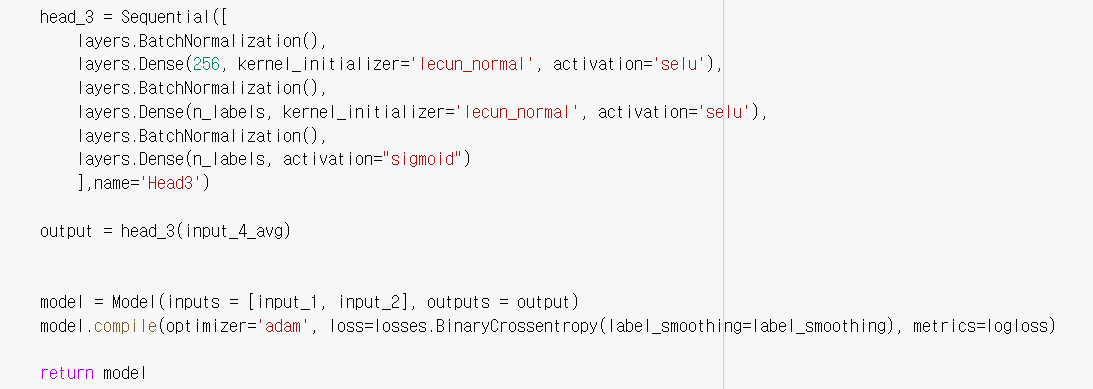
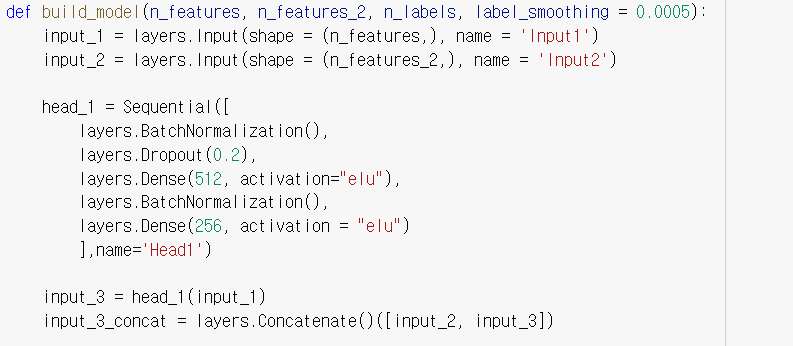
Data preprocessing여기선 cp\_type이 ctl\_vehicle인 경우를 제거하고 cp\_type, cp\_dose ,cp\_time을 제거한다.



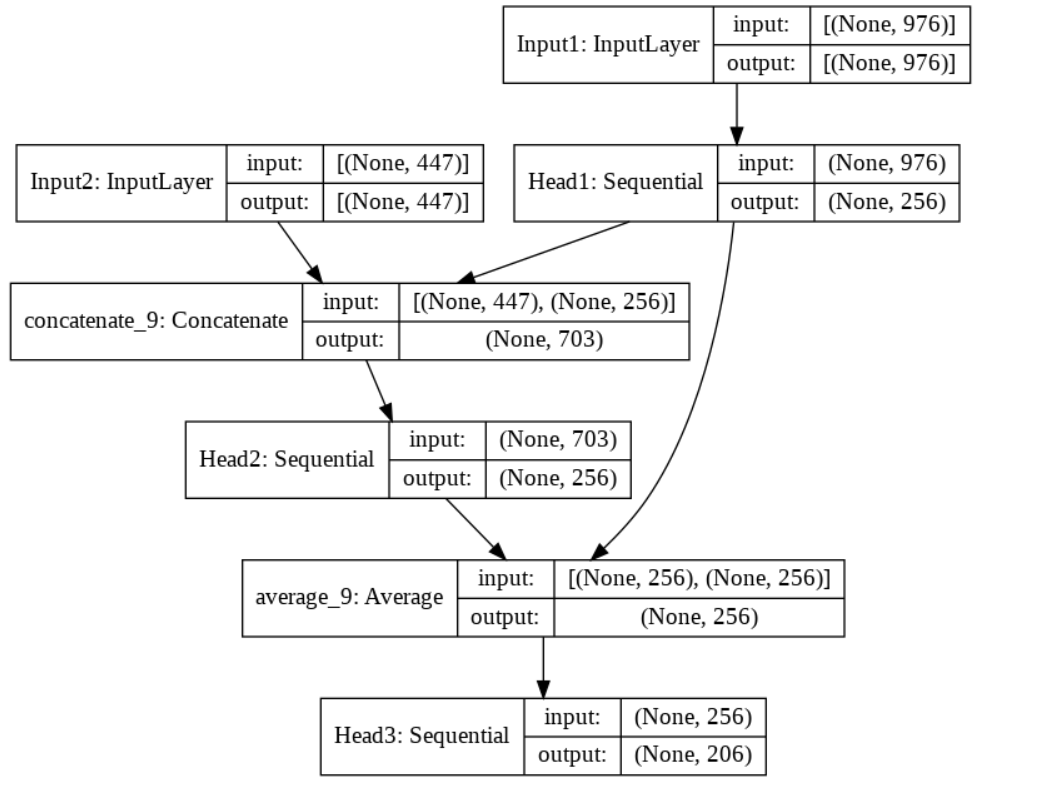
여기까지는 g-feature, c-feature를 PCA를 이용해 각각 2, 100으로 feature를 줄인다. 또한 feature를 g-feature(772) , c-feature(100), PCA\_g(2), PCA\_c(100), g-mean, c-mean로 총 976개로 이후 모델의 input1이 된다.

이는 oof evaluation을 위한 logloss function이다.

Model



Model Summary



이 모델은 두개의 input을 가진 Resnet 구조이다. Input1에선 g-feature(772) , c-feature(100), PCA\_g(2), PCA\_c(100), g-mean, c-mean로 총 976개이고, Input 2에선 p-value 0.0.1이하의 predictor로 447개이다. predictor 참고:<https://www.kaggle.com/demetrypascal/t-test-pca-rfe-logistic-regression/output>

Train





학습은 random의 seed값을 7개를 정하여 각 seed값 별로 10-fold 검증을 실시한 후 각 모델을 저장한다. 이때 callback함수로 reduce lr, earlystopping을 이용한다. 이렇게 검증한 결과 OOF score는 0.0165592507이 나왔다.