기계학습을 이용한 유전자 기능 예측 - 연구 일지 3-4주차

컴퓨터공학부 소프트웨어전공

4학년 박한솔, 박창호

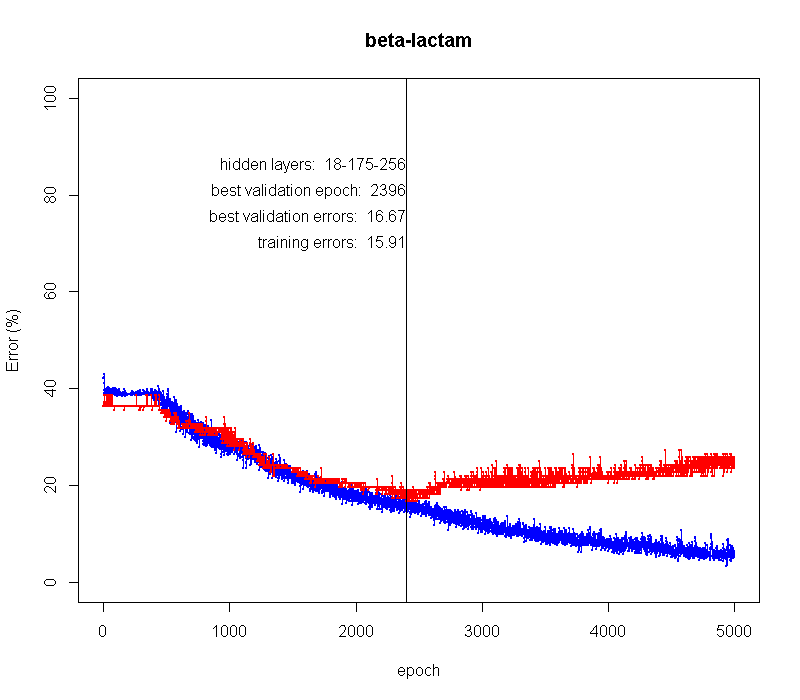
1. 목표

- MLP Classifier 성능 개선 , chembl350의 feature 추출

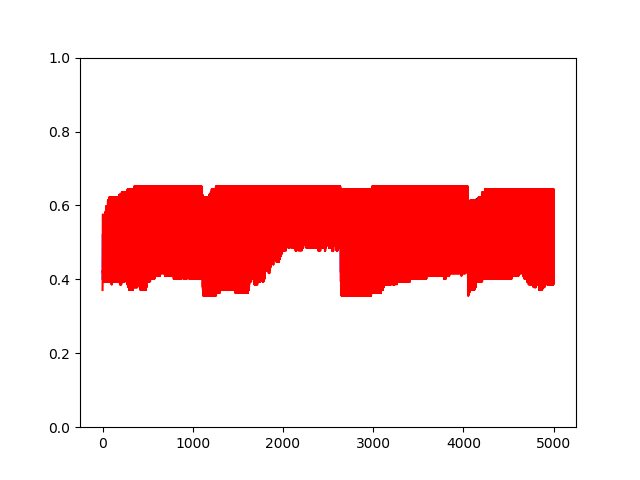
2. 진행 사항

2.1.  multi-layer-perceptron 연습

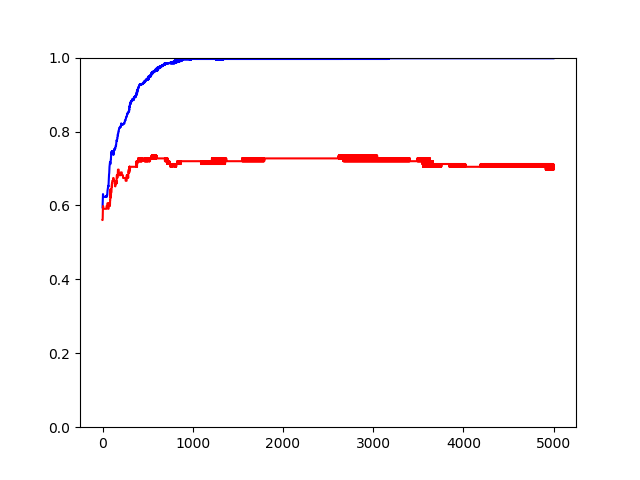
MLP Classifier는 1) adam optimizer, 2) softrelu, 3) dropout=0.5, 4) dropout=0.5을 이용을 했고, parameter는 1) learning rate=0.0001, 2) weight decay=0.001, 3) batch size=132를 이용해서 신재홍 박사님이 도출한 결과는 다음과 같습니다.



저희는 batch size를 training data의 크기인 528로 변경해서 accuracy를 측정해보았습니다. 그 결과, 다음과 같은 그래프를 도출 할 수 있었습니다.



결과를 보면, training이 제대로 이루어지지 않는다는 사실을 알 수 있었습니다. 그래서 training과 test set를 normalize를 하고, 다시 측정을 해보았습니다. 그 결과는 다음과 같습니다.



파란색이 training set을 test한 결과이고, 빨간색이 test set을 test한 결과로 약 70%의 정확도를 보였습니다.

2.2. 데이터의 feature 추출 및 프로젝트 모듈화

데이터의 갯수가 7000개 이상이어서 데이터들에 대한 molecular factor들만 계산하는 것이 최소 30분은 걸리게 되었습니다. 그래서 항상 mol factor들을 계산하는 것이 아닌 한번만 계산해서 파일로 저장할 수 있도록 추출기를 만들었습니다. 만든 것은 다시 mlp classifier로 실행해 봄으로써 정상적인 파일이 만들어 졌다는 것을 확인했습니다. 