ECE5970 Project Proposal

Team member: Jiaohao Li (jl3838), Quan Sun (qs84), Xiaowen Wang (xw453)
Dataset: The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE)

Background

Alzheimer's disease (AD) is the most common cause of dementia in the world. It is not only the 6th cause of death in the US, but also a chronic progressive disease: the patient's cognitive functions are impaired throughout years, which significantly reduces the quality of life. Currently, Alzheimer's disease has no cure. However, treatment can be induced effectively at early status of AD, by which disease progression can be slowed down. Therefore, the ability to predict Alzheimer's disease at its early stage can improve patient's life quality dramatically.

Literature Review and Data Introduction

The cause of Alzheimer's is believed to be dependent on many factors. Many hypotheses have been put forward, including particular genotype, i.e. alipoprotein E4 (APOE E4), the plaques of amyloid and the tangles of tau protein. Besides, other factors such as age, gender, lifestyle, heart health and particular disease, i.e. Down Syndrome and past head trauma, may have certain contribution into the disease development.

Unfortunately, there is still no gold standard in AD diagnosis. Doctors use several physical and neurological examination to determine dementia and further identify AD. Brain imaging is an effective tool to delve into the diagnosis, including structural MRI measurements into brain atrophy, FDG-PET measures into cell metabolism, AV45-PET into the levels of abnormal amyloid beta, AV1451-PET into the levels of abnormal tau protein, diffusion tensor imaging (DTI) into white matter degeneration, cerebrospinal fluid (CSF) measures, etc.

All of these mentioned diagnoses and relevant patients' clinical records are available on TADPOLE, in sum to 8717 patients' data available along with around 1890 records, while data missing occupies quite a lot. Training set is split as 2506 patients, validation set includes 867, and test set has 802 patients. Our goal in this project is to predict clinical status, Alzheimer's Disease Assessment Scale (ADAS-Cog13) and brain ventricles volume, using the data given by TADPOLE.

Machine Learning Approach

1. *Machine learning problems*

This is a supervised learning problem with both regression and classification parts. For predicting ADAS-Cog13, MMSE, and volume of brain ventricles, it is a regression problem since the targets are continuous variables. For the clinical status, this is classification since the target is categorical (normal, mild cognitive impairment, Alzheimer's disease).

2. *Data Cleansing/Feature Selection*

A lot of missing data in both training and validation sets, we will drop or fill in some values based on the same individual’s other visits and the quantity of missing data. Choosing important features and dropping those weak features for prediction. For better model performance, augment the quantity of data by introducing other datasets is also considered.

3. *Model*
For this supervised learning problem, there are several potential models, such as linear regression, SVM, K-Nearest Neighbor (KNN), Gradient Boosting, Gradient Boosting Decision Tree (GBDT), RandomForest, Deep neural network (DNN), etc. For DNN, we may design different structures with different optimizers, such as Stochastic Gradient Descent (SGD), Adam, Stochastic Gradient Descent with Warm Restarts(SGDR). SGDR sets learning rate with periodical change by cosine decay. To find better hyperparameters, plot with learning rate and loss function can be helpful in fine-tuning. Training different models by above methods and ensembling usually gives better performance, such as snapshot ensembling.

4. *Metrics*
For categorical target (clinical status) we will evaluate our performance by combining cross entropy, multiclass area under the receiver operating curve (mAUC), and the overall balanced classification accuracy (BCA). For quantitative targets (ADAS-Cog13 score, MMSE test score, and head size-normalized volume of the brain ventricles), we will use mean absolute error (MAE), the weighted error score (WES) and the coverage probability accuracy (CPA).

5. *Visualization*
Visualization is a substantial step to show results and determine the effect of models. We will visualize input and output data to show models performance.

Source of Information
Other than materials from lectures and support of TAs and the instructor, we will primarily use the papers and forum mentioned in TADPOLE main website. We will also search for recent research literature in machine learning and optimization if any difficulties outside of the scope of this course arises.

References

1. TADPOLE main website: https://tadpole.grand-challenge.org/Home/
2. NIH webpage on Alzheimer’s disease: https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet
3. Mayo Clinic on Alzheimer’s disease: https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447
4. Zhihua Zhou, “Ensemble Methods Foundations and Algorithms”. Chapman & Hall/CRC. 2012.