MACHINE LEARNING WITH BIOMEDICAL DATA
BioML CLASS PROJECT
OPTION A - TADPOLE
Fall 2017

Instructor: Mert R. Sabuncu, ECE/BME, Cornell

Introduction and Description

This project topic and the dataset have been derived from an on-going data science challenge (**The Alzheimer's Disease Prediction Of Longitudinal Evolution, TADPOLE,** https://tadpole.grand-challenge.org/).

This project represents a typical clinical scenario, where on each individual subject, various sources of information may or may not be available. Moreover, the individual might have visited the clinic multiple times at different time intervals (longitudinal). For example, the study might acquire a brain MRI scan (and compute certain measurements of neuroanatomy) and conduct a cognitive test, but not draw a blood sample during one visit. On another day, the study might omit the brain MRI for that subject but conduct a blood test.

Our goal is to make a *future prediction* (forecast) about clinically-relevant variables based on all the historical data we have (both on the given individual and all other individuals). A very naive baseline strategy might be to simply use the very last relevant observation that contains the target clinical variable and assume nothing will change going forward. This strategy will ignore historical trends (assuming the individual has visited the clinic several times). Moreover, this naïve approach will not exploit data from other individuals.

Data

The data are derived from a publicly available, large-scale Alzheimer's study (called ADNI - adni.loni.usc.edu). These are de-identified/anonymized human subject data, and as such, you should be extremely careful about their safety and security. You should NEVER attempt to: 1) identify the individuals in the data, 2) merge the provided files with other datasets, and 3) use the provided files for purposes other than the class project. You should always keep the files on a secure, password-protected computer that has controlled access. No person who is not a member of your class project team should be able to read these files. Importantly, you should permanently DELETE all data files after the completion of the class project and *no later than December 15, 2017.*

You will have several .csv files (text files formatted as comma separated variables).

TADPOLE_InputData.csv: This is the main dataset. Each row is a subject visit/exam[1]. PTID_Key is an anonymized patient ID you need to use as key, in order to link to the "target" data. Each column is a biomedical/clinical variable. If it is missing (and it often is), it was not collected during that visit. For missing variables, one approach might be to impute, e.g., by interpolating based on the same subject's (individual's) other visits and/or exploiting other subjects' data. Brief field/variable descriptions are available in TADPOLE_Dictionary.csv. You can refer to ADNI or TADPOLE for further details.

Target Files include the future target variables you will aim to predict. We have chosen one categorical variable (diagnosis: healthy control, mild cognitive impairment –MCI-, or Alzheimer's disease), and three quantitative/continuous variables which are biomarkers/correlates of Alzheimer's disease. The categorical variables are encoded in three columns, where an entry of 1 indicates corresponding clinical diagnosis. Note that these are mutually exclusive clinical categories and as such in each row the diagnosis variables should contain two zeros and one 1 (if they are not missing). MCI is a pre-dementia stage that may or may not lead to Alzheimer's disease. The quantitative targets are: Mini Mental State Exam (MMSE) test score, Alzheimer's Disease Assessment Scale (ADAS13), and head size-normalized volume of the brain ventricles, an MRI-derived marker of aging and dementia. All target variables are strongly related, thus one can treat this as a multi-task prediction problem where the machine learning approach takes advantage of similar prediction tasks.

You are given target variable values for two non-overlapping datasets: train and test. You should use the train dataset to train your machine learning model and test dataset to compute test accuracy and monitor generalization error and overfitting.

TADPOLE_TargetData_train.csv and TADPOLE_TargetData_test.csv are the corresponding target files. Each row is a subject's visit. Missing values are encoded as NaN. For training, you might want to pre-process these files and fill-in some of the missing values. For example, if a subject is diagnosed as healthy for visit 1 and 4, while diagnostic variables are missing during visits 2 and 3, it is safe to assume the subject was healthy during the intermediate visits.

Your ultimate objective is to fill in the TADPOLE_PredictTargetData_valid.csv file. You need to replace each NaN with a number (prediction) computed with your machine learning algorithm. This is a non-overlapping, independent dataset that we will use to independently validate your model(s). *Note that for the three diagnostic variables, I encourage you to assign (non-negative) probabilistic estimates that sum up 1.*

You are strongly encouraged to submit your best prediction (TADPOLE_PredictTargetData_valid.csv) file by 5 pm on November 9, 2017 (email msabuncu@cornell.edu with subject line: "BioML Target Predictions, Team X"). I will use these submissions to rank teams based on performance. The team(s) ranking highest in each category

[1] Strictly speaking, each row is a cluster of very close visits (within days of each other) that can be treated as one visit.

(diagnosis, MMSE, ADAS13, and Ventricle Volume) will receive a 5% bonus on their final grade and be encouraged to make an official submission to the Tadpole challenge (that deadline is November 15).

Performance Metrics (Mean squared error on quantitative variables, and cross-entropy on diagnosis) will be computed on all available validation data. I encourage computing these on the test dataset as well, as test performance *should be* very close to the validation performance.