CS 109A/Stat 121A/AC 209A: Final Project
Protopapas, Rader, Dave, Levine

Alzheimer's Disease and Cognitive Impairment Prediction

Problem statement

Alzheimer's disease (AD) is a type of dementia that causes problems with memory, thinking and behavior. AD is an irreversible process that typically begins after age 60. Once a patient has been diagnosed, their mental function usually declines until death. While some drug and non-drug treatments may help with both the cognitive and behavioral symptoms, currently Alzheimer's disease has no cure.

The personal, social, and economic impact of AD is profound. In the United States, more than 5 million people aged 65 and over suffer from the disease. AD is currently ranked as the sixth leading cause of death in the United States, and is the third leading cause of death for older people, just behind heart disease and cancer (NIA, 2017). The estimated national cost of patient care for Alzheimer's and other types of dementia was $236 billion in 2016. Therefore there is both a human and economic incentive to find an effective therapy for AD.

While scientists are still uncertain about the precise cause of AD, research has identified strong indicators for the disease. For example, an accumulation in the brain of Amyloid Beta protein (also referred to as Aβ protein or amyloid plaque) that is dense enough to meet a certain threshold (i.e., Amyloid Beta positivity) has been identified as a risk factor for AD (Dana, 2014). Leading biotechnology companies proactively developing therapies for Alzheimer's patients. Due to the currently irreversible nature of AD, it is believed that the effectiveness of any potential therapy greatly depends on early detection and treatment.

Project goal: The goal for this project is to propose, build and evaluate a data-driven model for a Alzheimer's or cognitive impairment related investigation using patient demographic information, medical records, and related data. A classic investigation would be centered around building a machine learning solution for evaluating a patient's risk of developing Alzheimer's based on medical history.

Other relevant projects could include evaluating imaging data or other associated data on cognitive measures or determining collinearity of various cognitive tests or generally investigating alternative target variables. Because of the variety of data collection methods, preprocessing, data cleansing and imputation will be a focus of the investigation. Since most of you might not be familiar with the disease or the tests used to diagnose it, you will also want to do some reading of relevant literature. You may use any skills learned in class (or not learned in class for the 209 component) such as preprocessing, including finding target values, feature selection, feature engineering, trying more than one learning models, and evaluating them at the end.

How to get the data

Your data will come from the ADNI (Alzheimer's Disease Neuroimaging Initiative) database

`http://adni.loni.usc.edu` (ADNI 2017)

The data is freely available. To download it one of you would need to register. To do so, visit

`https://ida.loni.usc.edu/collaboration/access/appLicense.jsp`

Review the ADNI Data Use Agreement and agree to the terms. Complete the 4 steps for sign up. You should receive a confirmation email with instructions to access your account within a week. We will get you started with some data since the process might take a few days.

Available features in the ADNI database include:

1. patients demographics
2. Medical history (disease and medication)
3. Lab records
4. Cognitive test score
5. Imaging data

For an overview of the data see

`http://adni.loni.usc.edu/data-samples/`

For a brief overview of the protocols used see

`http://adni.loni.usc.edu/about/`

One possible target variable is DX_BL which can take a number of values such as AD == Alzheimer's Disease, LMCI == low to mild cognitive impairment, CN == cognitively normal, etc. Other target variables might work for your definition of the problem.

High-level project goals

Use some of the the data in the ADNI database to construct a Baseline model. Some of the issues you will want to tackle are:

1. there is limited patient data available
2. data is collected according to multiple protocols (ADNI, ADNI Go, ADNI 2)
3. some of the data has a longitudinal component that adds complexity
4. labeling is not always reliable, e.g. it might be not clear whether it's AD or other form of dementia
5. early detection is key so finding the smallest and least expensive feature subset is important.