Data description

There’s no denying that heart is indispensable to people. It makes sense to know the systolic and diastolic volumes, which indicate the capacity of the heart. Now, we have a dataset of 500 subjects. For each subject, we have several observations, every observation consists 30 cardiac MRI images in [DICOM](https://en.wikipedia.org/wiki/DICOM) format. These 30 2D cine images make up the cardiac cycle.  To be noticed, for different subject, we have different number of observations.

Data processing

The general preprocessing is first extracting all .dcm files from zip file, then construct a function to import all .dcm files into array in the python environment. We prefer deep learning to deal with the dataset because deep learning methods always consider images as 2D or 3D arrays or matrices and it will bring a lot of convenience. We use Object Oriented Programming to process the folders in batch. After the processing on all the train folders done, we can access many attributes of the Dataset objects which is stored as elements of 'dset' list. We are able to load data of 493 subjects with 7 subjects dropped.

To simplify this problem, we just take 'sax' images and ignore '2ch' and '4ch' images. Roughly, for each subject, we just resize and take average of all these images to make this problem easier. The first purpose of these operations is scaling the number of pixels in output for each subject . And the second purpose is extracting more representative feature from multiple images. Through reshaping, we will get a 394\*64\*64 matrix. We can interpret it as every subject has 64\*64 features.

However, the above method to deal with the X matrix will always decline the precision of our model for it ignores the properties of periodicity. Another way to deal with the data is that we take average of images in the same stage for every subject instead of taking average of all the images. Through this method, we will get a 394\*30\*64\*64 matrix, which still reserve the property of periodicity. We can interpret it as every subject has 30\*64\*64 features.

Then we split the training data into two parts, with 330 subjects as new training data and 163 subjects as testing data.

Also, we use one hot encoding on target value. In terms of the evaluation method, we set the result to be related to cumulative probabilities. To be specific, P0 represents the probability the volume is less than or equal to 0 mL, P1 represents the probability the volume is less than or equal to 1 mL and is larger than 0 ml, etc. So we round the target value to int and extend the target variable to one-hot vector of length 600.

Evalution