Data description

June 13, 2024

1 Explaination of data used in AUTOsurv

For any given dataset, 20% of each dataset were kept as testing set that did not participate in any of the model training process. The remaining 80% of the data will be split into 8/2 for training and validation. **Make sure the patient id for miRNA and gene dataset is consistant**.

1.1 Gene data(n x [m+7] dataset)

It consists 8 different parts, specific function **load_data** was defined to extract them individually.

1. Patient ID(nx1)|import as patient ID
2. OS(nx1): overall survival| import as yevent_
3. OS.time(nx1): overall survival time|import as ytime_
4. age(nx1)|import as age_
5. race_white(nx1)|import as race_white_
6. stage_i(nx1)|import as stage_i_
7. stage_ii(nx1)|import as stage_ii_
8. gene(nxm):contains m diferent gene's variable|import as x_train

4-8 were used in model

1.2 mirna data(n x [k+7] dataset)

All the format are the same as gene data except change the gene data to mirna data(k dimension)

1.3 pathway mask(m x a dataset)

Specific function **load_pathway** was used to load a bi-adjacency matrix of pathways(a dimension) and genes(m dimension), and then covert it to a Pytorch tensor.

The pathway mask data is the same for tune and overall data

2 Requirement list for our own dataset

1. Gene data with m dimension
2. Pathway data with m x n dimension where m corresponding to m different gene and n corresponding to n diferent pathway(matirx is binary)
3. Other RNA data
4. Other biomaker data