Required Python packages

- Numpy (>=1.17.2)
- Pandas (>=0.25.1)
- sklearn (>=0.21.3)
- PyTorch (pytorch version >=1.5.0, torchvision version >=0.6.0)

Imputation

data_processing.py: Run data_processing.py to pre-preprocess the gene expression and SNP data for imputation.

$ python data_processing.py --data_directory /home/tanvir/Diabetes/data/raw_data/ --save_directory /home/tanvir/Diabetes/data/processed_data/

--data_directory: directory of raw omics data

--save_directory: directory to save processed data

imputation.py: Run imputation.py to perform gene expression imputation on processed data.

$ Python imputation.py --data_directory /home/tanvir/Diabetes/data/processed_data/ --save_directory /home/tanvir/Diabetes/data/imputed_data/ --n_epochs 100 --batch_size 32 --learning_rate 1e-3 --true_enc_hidden_size 100 --true_enc_batch_size 32 --true_enc_num_epochs 100 --true_enc_learning_rate 1e-4 --true_enc_gamma 0.99 --syn_enc_num_epochs 25 --syn_enc_batch_size 512 --syn_enc_learning_rate 1e-5

--data_directory: directory of input data

--save_directory: directory to save output

--n_epochs: number of epoch for C_1

--batch_size: batch size used in C_1

--learning_rate: learning rate used in C_1

--true_enc_hidden_size: size of the encoding in C_0

--true_enc_batch_size: batch size used in C_0

--true_enc_num_epochs: number of epoch for C_0

--true_enc_learning_rate: learning rate used in C_0

--true_enc_gamma: *gamma* value in learning rate shceduler for C_0

--syn_enc_num_epochs: number of epoch for C_2

--syn_enc_batch_size: batch size used in C_2

--syn_enc_learning_rate: learning rate used in C_2

Prediction

prediction.py: Run prediction.py to predicts IA status of participants using imputated data.

$ python prediction.py --hidden_size 200 --num_layers 3 --num_epochs 5 --batch_size 8 --learning_rate 0.00001 --end 24 --serial 16 --option 0 --imputed_data_dir /home/tanvir/Diabetes/data/imputed_data/ --processed_data_dir /home/tanvir/Diabetes/data/processed_data/ --raw_data_dir /home/tanvir/Diabetes/data/raw_data/

--hidden_size: size of the hidden representation in LSTM

--num_layers: number of layers in LSTM

--num_epochs: number of training epochs in LSTM

--batch_size: batch size in LSTM

--learning_rate: learning rate in LSTM

--end: IA time cutoff in months

--serial: gene expression cutoff in number of time steps

--option: run the model using combined data or only gene expression. Use 0 for combined data and 6 for only gene expression.

--imputed_data_dir: save directory for imputation.py

--processed_data_dir: save directory for data_processing.py

--raw_data_dir: directory of raw omics data

Datasets

Gene expression and SNP can be downloaded from this link:

https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001442.v1.p1