Team Meeting

29 July 2022 / 3:00 PM / Zoom

Attendees

Ate, Stefan, Xavier, Jiadong, Zexi, Ni

Agenda

Machine Learning Models

To Do

Models

Dataset trained - E-Risk

1. Size: 1648 * 833 (Obtained from meta analysis)
2. Removing NaN gives 1464 * 833
3. Train vs Dev - 75% vs 25%

Models have been checked:

1. SVM (SVC)
2. Logistic Regression, Penalized Regression (L2 Norm - Ridge)
3. Bayesian Algorithm (MNB) - Low AUC
4. Adaboosting - Low AUC
5. Gradient Boosting - Slow
6. Random Forests

Models tuned (still need to be improved, only hyperparameters that deemed important are tuned) :

1. Logistic Regression - 80.04%
2. SVM - 82.36%
3. Random Forest - 79.49%
4. Compare with the paper

Training data (whole blood 450k array, NTR)

N total	N MZ twins	N DZ twins	N non-twins	minimum lambda	AUC	SE	N nonzero CpGs
1989	1256	733	0	0.002239	0.8124	0.008228	1792
1989	1256	733	0	0.01507	0.8007	0.01219	249
2155	1258	738	159	0.002392	0.8281	0.009245	1867
2155	1258	738	159	0.01611	0.7989	0.007947	232

Test dataset 1 (whole blood 450k array, NTR)

N total	N MZ twins	N DZ twins	N non-twins	AUC	proportion MZ twins correctly predicted	proportion DZ twins correctly predicted	proportion non-twins correctly predicted	proportion MZ twins incorrectly predicted	proportion DZ twins incorrectly predicted	proportion non-twins incorrectly predicted
1100	522	339	239	0.687	0.785	0.472	0.310	0.215	0.528	0.690
1100	522	339	239	0.751	0.843	0.469	0.444	0.157	0.531	0.556
934	520	334	80	0.715	0.773	0.512	0.538	0.227	0.488	0.463
934	520	334	80	0.766	0.808	0.572	0.625	0.192	0.428	0.375

Test dataset 3 (whole blood 450k array, E-Risk)										
N total	N MZ twins	N DZ twins	N non-twins	AUC	proportion MZ twins correctly predicted	proportion DZ twins correctly predicted	proportion non-twins correctly predicted	proportion MZ twins incorrectly predicted	proportion DZ twins incorrectly predicted	proportion non-twins incorrectly predicted
1658	852	612	0							
1658	852	612	0	0.728	0.684	0.683	NaN	0.316	0.317	NaN
1658	852	612	0							
1658	852	612	0	0.739	0.621	0.740	NaN	0.379	0.260	NaN

Models yet to be checked:

1. Deep learning - Pytorch / Tensorflow
2. Stacking / Embedding / Voting

Note

1. Full dataset: 450k * 1400
2. The size of removing list: 70k
3. Training using full dataset: 400k * 1400 => Spartan
4. The normalization methods of each dataset are different

To Do

1. Continue tuning the models with highest performance (hyperparameter, overfitting...) - Ni, Ate
2. Check the AUC on the other four datasets (the cleaning step for the four datasets are nearly done)
3. Logistic regression adjusting proportion (elastic net); more values for C (hyperparameters) - also try C smaller than 1
4. Stacking / Embedding / Voting - Ni, Ate
5. Random forest - more trees higher n_estimators
6. Deep learning - Jesse, Auto-encoders
7. Spartan - Stefan, Xavier
8. Dimensionality reduction on 833 - pls : select the number of components., cca, (ask the client if cares about interpretation? auc?)
9. Variable selection?