Team Meeting

11 May 2022 / 10:00 AM / ZOOM

Attendees

Ate, Stefan, Xavier, Jiadong, Zexi, Ni, Shuai

Agenda

Outline

1. Explain results (Ni),
2. Supplementary software (Zexi)
3. Ask biology term (Ate),
4. Jiadong will ask some questions as well
5. What can the results conclude?
6. Access to other datasets,
7. What is the application of these machine learning classifier (why is it important)?

New Business

Results:

- Supplementary data 13 and 14 are the two classifiers of the nature paper
- Supplementary data 13: MZ vs DZ
- Supplementary data 14: MZ vs (DZ & family)
- Zygosity: Classifier 13 MZ vs DZ
- MZ twin: Classifier 14 MZ vs (DZ & family)
- There are 1658 samples in E-Risk dataset, 194 NA, 852MZ, 612DZ
- 194 NA samples were removed when calculating the model performance
- If the missing values were not removed, the auc for classifier 13 would be 0.696
- Should we remove the missing values? 200 out of 1600,

A	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP
1							Test dataset 3 (whole blood 450k array, E-Risk)				
2 Description	N total	N MZ twins	N DZ twins	N non-twin	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
3 Zygosity 450k - standardized	1658	852	612	0							
4 Zygosity meta-analysis - standardized*	1658	852	612	0	0.728	0.684	0.683	NaN	0.316	0.317	NaN
5 MZ twin 450k - standardized	1658	852	612	0							
6 MZ twin meta-analysis - standardized	1658	852	612	0	0.739	0.621	0.740	NaN	0.379	0.260	NaN

ROC curve for classifier 13
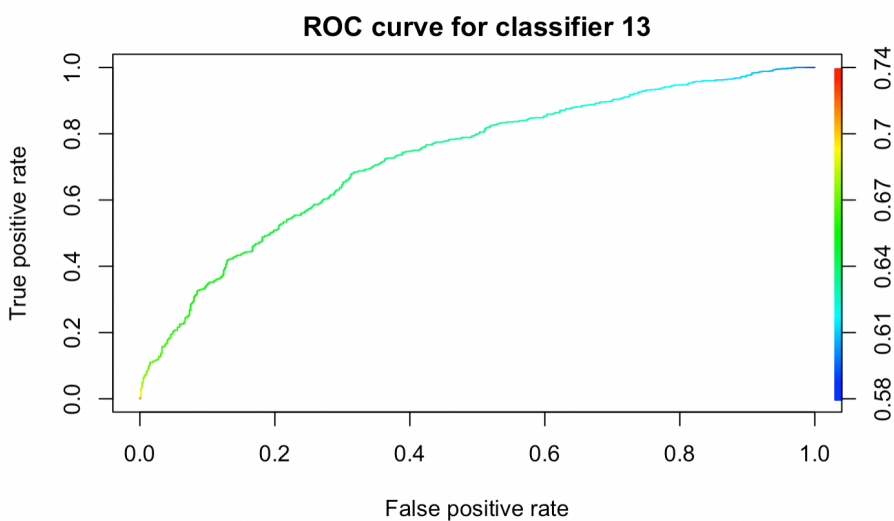

```
[1] "cutoff: 0.6392671063064"
Area under the curve: 0.7276
Confusion Matrix and Statistics


pred_b   0   1          Supplementary Data 13
     0 418 269                0: DZ
     1 194 583                1: MZ


               Accuracy : 0.6837
                 95% CI : (0.6592, 0.7075)
    No Information Rate : 0.582
    P-Value [Acc > NIR] : 6.783e-16

                  Kappa : 0.361

 Mcnemar's Test P-Value : 0.0005837

            Sensitivity : 0.6843
            Specificity : 0.6830
         Pos Pred Value : 0.7503
         Neg Pred Value : 0.6084
              Precision : 0.7503
                 Recall : 0.6843
                     F1 : 0.7158
             Prevalence : 0.5820
         Detection Rate : 0.3982
   Detection Prevalence : 0.5307
      Balanced Accuracy : 0.6836

       'Positive' Class : 1
```
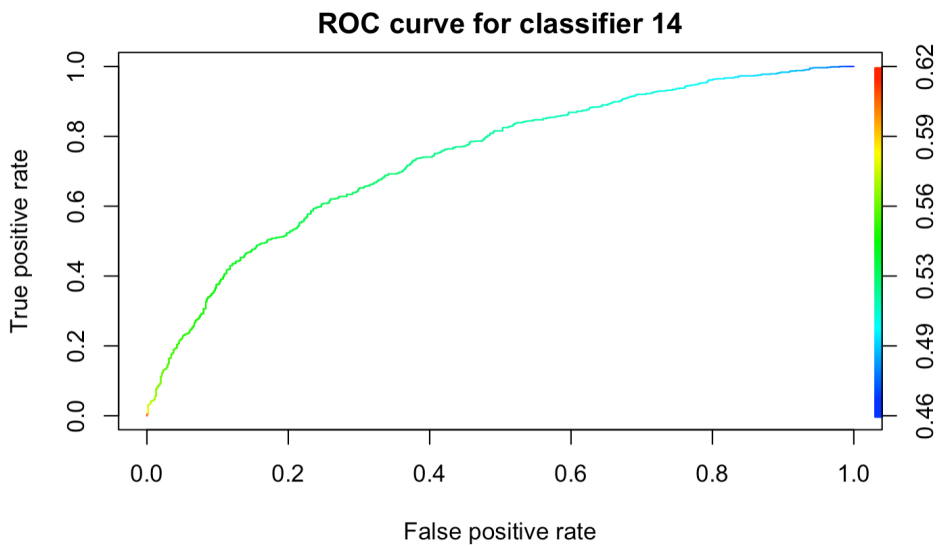
ROC curve for classifier 14


```
[1] "cutoff for supplementary data 14: 0.526967877315572"
Area under the curve: 0.7388
Confusion Matrix and Statistics


pred_b.s14   0   1
         0 453 323
         1 159 529
```

Supplementary Data 14

0: DZ
1: MZ

```
              Accuracy : 0.6708
                95% CI : (0.646, 0.6948)
    No Information Rate : 0.582
    P-Value [Acc > NIR] : 1.829e-12

                 Kappa : 0.348

 Mcnemar's Test P-Value : 1.133e-13

           Sensitivity : 0.6209
           Specificity : 0.7402
        Pos Pred Value : 0.7689
        Neg Pred Value : 0.5838
             Precision : 0.7689
                Recall : 0.6209
                    F1 : 0.6870
            Prevalence : 0.5820
        Detection Rate : 0.3613
  Detection Prevalence : 0.4699
     Balanced Accuracy : 0.6805

       'Positive' Class : 1
```

Notes

- Importance of studying zygosity: methodologies to infer, this project is only using methylation data method
- One twin vs pair of twins (Both of the MZ twins should be considered as they have different methylation levels)
- Genetic markers are almost the same for MZ
- Do not consider the R software file
- Data Linkage, algorithms
- Computing time
- Stabilized regression
- Generalisation - confusion on appearance when growing up

Action Items

1. Keep testing on the 4 other datasets.
2. Focus on the report and presentation

Next Meeting Agenda