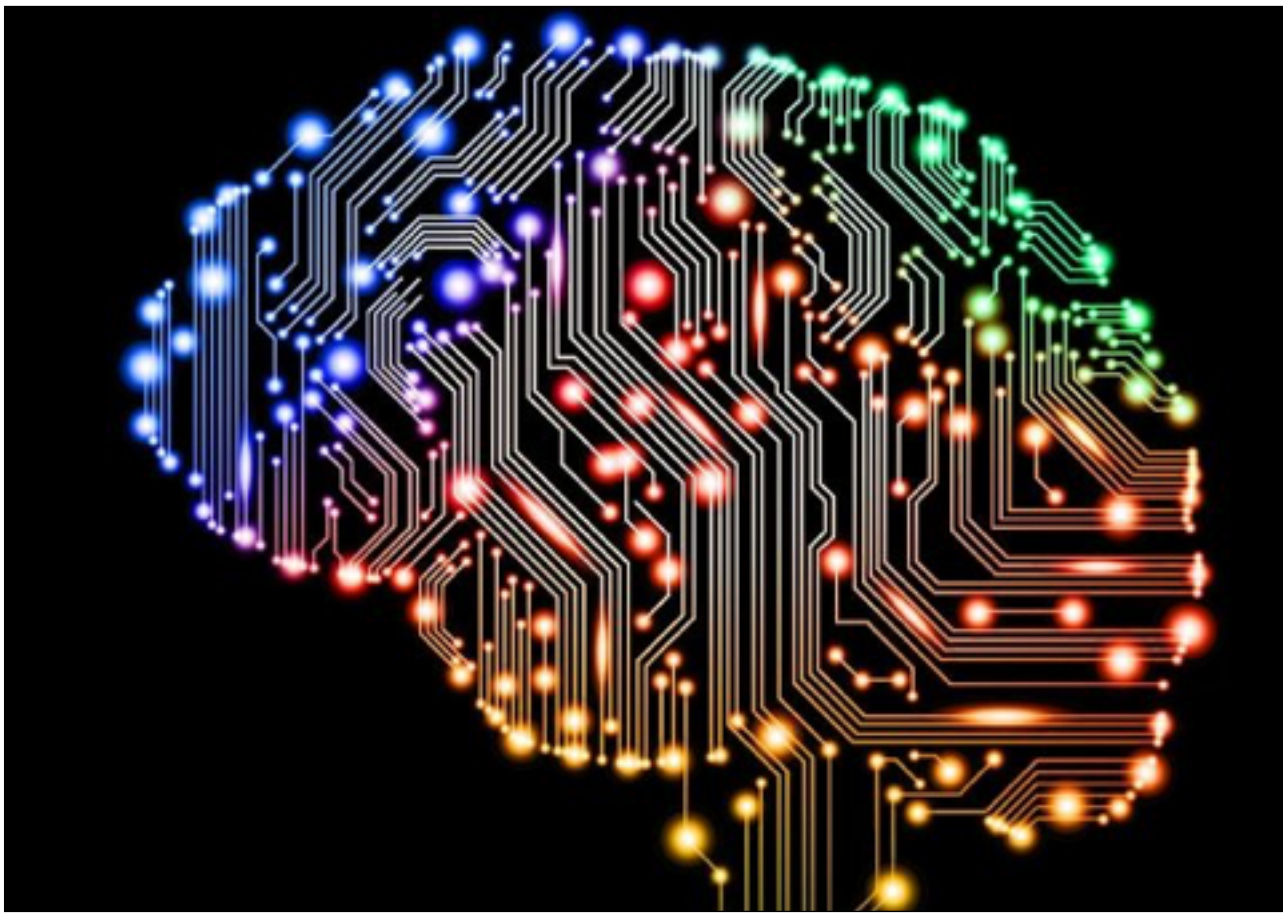
Evaluations

Results from the model

Yuntong Bai - 17/3/27



The optimization problem we aim to solve is as following,

$$\sum_{i=1}^{p}(f_i - y_i)^2 + \mu \sum_{i=p+1}^{n} f_i^2 + c\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

Focusing on the special case where μ=1, these three terms degenerate to the following terms,

$$\min_{f,\gamma}(f-y)^T(f-y) + cf^T Lf$$

Where the matrix L is graph Laplacian matrix defined as $L = \mathrm{diag}(\Sigma_j w_{ij})-W$, and $y=(y_1,\ldots y_p,0,\ldots,0)^T$, where $y_1 \sim y_p$ are the labels of the training group, and labels of testing group are set to be zeros. The solution of this problem is obtained as

$$f = (I + cL)^{-1} y$$

where I is an identity matrix.

The updated MatLab file is the new 'starter' file for implementing the algorithm using Methylation data and the difficult part is selecting suitable model for calculating adjacency matrix w and suitable value of c. And one way of solving this problem is by cross-validation. The test error is 35.56% while ROC score is 0.60. This is not very ideal result, probably because of simple choice of values of parameters during cross-validation, which reduces computational time. And if given more choices of parameter values to train in cross-validation, the results are supposed to be improved. And this can be solved by running this implementation on Cypress.