Introduction to Biomedical Informatics by Dr. Junhee Seok
Final Exam

Gene Expression Prediction Analysis

This assignment is to predict the phenotypes of colon cancer patients using gene expression profiles. You will use the same data set of the first assignment (GSE39582). This data set has 566 patient profiles in total. Among them, 443 are a discovery set and 123 are a validation set. The dataset annotation can be found in one of "!Sample_characteristics_ch1" fields. For each prediction problem, please use the discovery set as a training set, and the validation set as a test set. Please keep in mind that any information of a test set SHOULD NOT be used to build a predictor. The test set is used only for testing your predictor.

For this data, please answer the following questions. For each prediction problem, you need to

1. Describe the whole procedure,
2. Try at least three different ways of prediction (including feature selection, dimension reduction, and prediction algorithms)
3. Report the number of genes (or probe sets) in your predictors.
4. Report the prediction result of the cross validations using the training set,
5. Report the prediction result for the test set, and

The prediction result should be shown by ROC and AUC for discrete variables and RMSE for continuous variables. For each prediction problem, please ignore samples with missing values. The answer should include codes for the prediction.

1. (Sex) Predict the gender of samples. For the genes included in your predictor, find the significantly enriched gene sets among the C1 gene sets of MSigDB. Can you interpret the result?

2. (Continuous tumor stage) Please predict TMN tumor stages of patients. TMN tumor stage can be considered as a continuous variable.

3. (Binary tumor stage) Instead of continuous TMN tumor stages, we will consider two stages. Stage 0, 1, and 2 are "benign" stage, and stage 3 and 4 are "malignant" stage. Please predict the binary tumor stage.

4. (Reference Code) Review the given reference analysis code. Explain what this code does. If necessary, explain with the line numbers.

If you have any question, do not hesitate to ask me!!!