Genomics Project

The topic of this project is a gene expression data study of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) [1]. Bone marrow samples were taken from each patient, RNA extracted, and probed for 6817 human genes (6568 of which are available for the project data analysis). The training dataset ("training_ALL_AML.csv") has 38 bone marrow samples (11 AML, 27 ALL). The independent (test) dataset ("independent_ALL_AML.csv") had 24 bone marrow and 10 peripheral blood samples and this was reserved in the original study for the evaluation of machine learning methods. The gene variables are the columns in each of the datasets. Both datasets contain the actual/true classifications in separate files ("trainClass.csv" and "indepClass.csv", respectively). All datasets are available via the Syllabus page in CLE (Collaborative Learning Environment) for Biostat 202: https://courses.ucsf.edu/course/view.php?id=3040.

You should follow the Project Guidelines as you work through the project. Some additional points specific to the project for each component are given below:

Component 1: You can treat this as a classification problem or if you want to ignore the class information you can treat it as a clustering/data reduction problem.

You will need to perform the following data processing steps to get the data ready for analysis,

1. Remove the Control variables from the dataset – these are identified as having "control" somewhere in the variable name and are all grouped together from the first variable ("AFFX-BioB-5_at (endogenous control)") to the variable "AFFX-HSAC07/X00351_M_st (endogenous control)", with the first variable of interest being "GB DEF = GABAa receptor alpha-3 subunit"
2. Merge the inputs and class data
3. Ensure that all variables have the correct type and role

Component 2: Use visualization tools to examine the dataset (both training and independent datasets). Are there extreme outliers? Are there any missing data?

Component 3: Fit a classification or clustering model/algorithm to the training dataset. Determine the classification accuracy or cluster separation within the training dataset. Did you try any other models/algorithm? Why did you choose the particular model/algorithm that you did?

Evaluate the classification accuracy/cluster reliability of the model/algorithm fitted to the training dataset on the independent dataset and compare to that of the training dataset.

Look again at visualizations of the independent dataset. Is there anything unusual about the observations that were not classified correctly or are there any cluster outliers in the independent dataset?

Did the classification accuracy or cluster reliability/separation in the independent dataset differ to that of the training dataset? Is this difference between training and independent datasets in the direction you would expect and why? Do you consider the classification method to be successful and why? Are there any limitations or potential biases that need to be considered when interpreting the results?

[1] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286.5439 (1999): 531-537.