Valentin Bernu
13/12

Ovarian tumor subtype prediction using Gene Expression Data

Home case study for Epigene Labs

Exploration

The data is taken from the dataset GSE6008 briefly described in the document “Epigene Labs - Data Science Challenge (Prediction)”. Train data is composed of 78 columns and 22286 rows. Test data is composed of 20 columns and 22286 rows.

The column “type” is the type of ovarian tumor. 22283 columns are gene expressions. The columns "Unnamed: 0" and "samples" are not considered in this study.

Tumor types

There are four types of ovarian tumor in the data: Serous, Mucinous, Endometrioid and ClearCel. The tumor type distribution in train data is 33 (42%) Serous, 10 (13%) Mucinous, 29 (37%) Endometrioid and 6 (8%) ClearCel. This is not representative of the observed repartition of ovarian tumor type. For instance Ovcare website accounts for up to 70% Serous types and about 10% for Endometrioid. Nonetheless the observed repartition varies with sources.

Gene expression

The graphs discussed below are in the images folder.

Gene expression columns are randomly sampled and plotted in histograms. The distribution shapes, the means and the standard deviations are not homogeneous.

Normal distributions are quite frequent. Shapiro-Wilk tests corrected with False Discovery Rate (FDR) correction quantifies this observation. For about half of the column, the hypothesis of a normal distribution is rejected.

Correlation matrices of gene expression shows that in spite of the very high number of columns, they are mostly not correlated.

Discussion

The very high number of columns compared to the number of rows makes the study belong to the theme of “the curse of dimensionality”. The focus will not be on training algorithms but on selecting features.

Half of the columns did not pass the normality test. Usual statistical tools relying on normal distributions should be treated with caution (e.g. Student t test, Gaussian Naive Bayes).

Feature selection

The first feature selection made is filtering columns with Kruskal-Wallis test corrected with Bonferroni correction.

Benchmark

The selected features are benchmarked with classifiers belonging to the weak learner category: Gaussian Naive Bayes (GNB), Support Vector Machine (SVM) and Logistic Regression (Lo). Random Forest (RF) classifier is added to the benchmark. They are compared to naive approaches (DS & DMF).

The benchmark shows that classifiers have approximately the same 3-fold cross validated accuracy on 10 iterations : 0.8. The boxplots of the benchmark are displayed in the images folder.

Sequential Feature Selection (SFS)

The weakest learner: Support Vector Machine classifier is used to perform a SFS study on the previously selected features. The result is displayed in the image folder. The SFS shows that the previous accuracy can be improved by selecting the adequate number of features with the SFS method.

Prediction and evaluation

The model chosen is SVM with 10 features selected by SFS. **Prediction on test data shows an accuracy of 0.8**. Confusion matrices show that most of the misclassifications are Endometrioid tumors classified as Serous for both train and test data.

The loss of accuracy points for test data compared to train data can be explained by the introduction of overfitting by the SFS method.

Scaling note: SFS is computationally expensive. With the number of rows increasing the number of selected columns with the Kruskal-Wallis test will increase. This will make the SFS computation time explode. I suggest setting the number of selected columns with Kruskal-Wallis to a constant integer that allows a reasonable computation time. The number of rows in itself is not likely to be an issue for the computation as the model chosen is SVM.