Classifying Leukemia Using Logistic Regression and Lasso Regression

Team: Allan Li, Sarah Trần, Veronica Peng {shilun, sdtran, tpeng24}@stanford.edu

Prediction

Hence, our project aims to increase the diagnostic accuracy for leukemia subtypes by exploring the possibility of diagnosing leukemia based on the patient's genes. We used two-sample Student's t-tests and PCA to preprocess the data, applied logistic regression and lasso regression to our dataset to make predictions, and uses hierarchical agglomerative cluster analysis (HAC) to visualize the data. All algorithms other than the baseline yields a prediction accuracy higher than 90% on test set, with the highest reaching 100%.

Data

Golub dataset (1999): The first dataset has the expression levels of 7,129 genes assayed with HGU6800 gene chips from 72 subjects, and we aimed to classify AML and ALL patients based on the genes.

Chiaretti dataset (2004): The second dataset has 128 subjects with ALL along with the expression levels of 12,625 genes assayed using HGU95aV2 chips, and we aimed to classify patients with ALL into two subcategories: ALL affecting B cells and ALL affecting T cells.

Feature

Different features for 3 different algorithms

Baseline:

- randomly selecte 10 genes from the original datasets.

Logistic regression with feature selection:

Need to avoid overfitting

- select the top 10 most informative genes with the lowest p-value using two-sample Student's t-tests
- decompose those selected features using PCA.

Lasso regression:

automates feature selection - no worries about overfitting

- use top 100 most informative genes from two-sample Student's t-tests.

Model

We used logistic regression and lasso regression, to make predictions, and uses hierarchical agglomerative cluster analysis (HAC) to visualize the data.

Logistic regression: optimizes

$$\min_{w,c} \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

Lasso regression with L1 norm: optimizes

$$\min_{w,c} ||w||_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

C is the reciprocal of the regression strength

HAC with Euclidean distance metric: metric formula

$$||a-b||_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Results

Golub dataset:

38 data points for the training set and 34 for the test set

Chiaretti dataset:

90 data points for the training set and 38 for the test set.

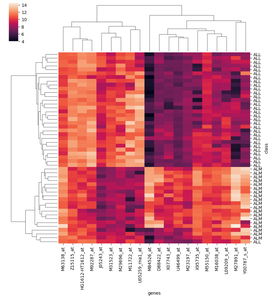
	Golub Dataset		Chiaretti Dataset	
	Training Accuracy	Test Accuracy	Training Accuracy	Test Accuracy
Logistic Regression with Random Feature Selection	82.3%	67%	89.5%	15.8%
Logistic Regression with T-test	100%	100%	100%	96.9%
Logistic Regression with T-test and 10-fold cross validation	--	95.7%	--	98.8%
Logistic Regression with T-test, PCA	100%	97%	100%	100%
Logistic Regression with T-test, PCA and 10-fold cross validation	--	92%	--	98.8%
Lasso Regression	100%	95.5%	100%	100%

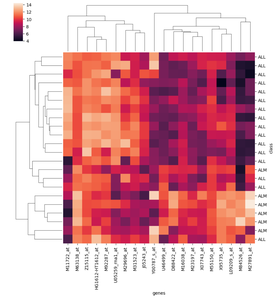
Discussion

Overall, our algorithm yields a high accuracy in predicting subtypes of leukemia, with most of them reaching 90%

Golub dataset:

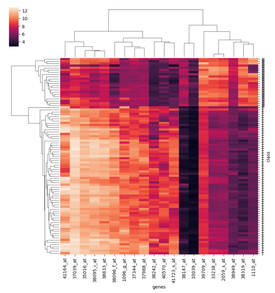
Feature selection with t-test finds new genes helpful in identifying different leukemia subtypes, many of them have functions intuitively related to classifying two types of cancers.

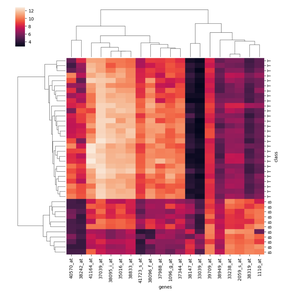




Charetti dataset:

The highest accuracy rate yielded by our algorithm on the test set is 100% when we applied lasso regression.





Future Work

Our algorithm performs very well in classifying subtypes of leukemia. Thus, given more time, instead of spending more time increasing the prediction accuracy, we may apply our algorithm on other types of cancer prediction based on patient's gene to help improve the diagnosis reliability。