**Jaccard similarity of selected features with p-value**

Analyze significance of features using Welch’s t-test per feature for each label against the healthy patients. Then used Bonferroni correction to correct for multiple hypothesis tests. Used Jaccard similarity to test the similarity of the features selected between different leukemia types. Learned that MDS and mature B-ALL with t(8;14) seem to be the most different from other Leukemia types. Other leukemia types seem to have selected features that are at least marginally similar to each other (~10%), with some reaching similarity levels of up to around 70%, which may give us some clue as to what type of mistakes our model will make in the future.

See Excel. Built with just the training data to not bias the model we train to

Used 1 v all

90%-10% split

**Data for cross validation with model types** for all 17788 features

Using L1 regularization

logistic

{'top1': 0.9238095238095239, 'top5': 1.0}

logistic - featurize by normalizing each gene expression level

{'top1': 0.9238095238095239, 'top5': 0.9952380952380953}

logistic - featurize by normalizing each gene expression level by "Non-leukemia and healthy bone marrow"

{'top1': 0.9047619047619048, 'top5': 0.9952380952380953}

|  |  |  |
| --- | --- | --- |
| Model | Top 1 from CV | Top 5 from CV |
| Logistic regression | 0.924 | 1.0 |
| Logistic regression with normalized variables | 0.924 | 1.0 |
| Logistic regression with normalized variables normalized by normal patients | 0.905 | 0.995 |

**Data for cross validation with model types** for selected 1225 features using p-value

Tries to avoid overfitting

logistic

{'top1': 0.9142857142857143, 'top5': 1.0}

logistic - featurize by normalizing each gene expression level

{'top1': 0.919047619047619, 'top5': 0.9952380952380953}

logistic - featurize by normalizing each gene expression level by "Non-leukemia and healthy bone marrow"

{'top1': 0.9095238095238095, 'top5': 0.9952380952380953}

Analysis of similarity of weights learned for each model

logistic

{'top1': 0.919047619047619, 'top5': 1.0}

logistic - featurize by normalizing each gene expression level

{'top1': 0.9238095238095239, 'top5': 0.9952380952380953}

logistic - featurize by normalizing each gene expression level by "Non-leukemia and healthy bone marrow"

{'top1': 0.9095238095238095, 'top5': 0.9952380952380953}

eval with significant features alpha = 0.10

logistic

{'top1': 0.8761904761904762, 'top5': 0.9904761904761905}

logistic - featurize by normalizing each gene expression level

{'top1': 0.8571428571428571, 'top5': 0.9809523809523809}

logistic - featurize by normalizing each gene expression level by "Non-leukemia and healthy bone marrow"

{'top1': 0.8428571428571429, 'top5': 0.9761904761904762}

eval with significant features alpha = 0.05

logistic

{'top1': 0.8714285714285714, 'top5': 0.9857142857142858}

logistic - featurize by normalizing each gene expression level

{'top1': 0.8523809523809524, 'top5': 0.9809523809523809}

logistic - featurize by normalizing each gene expression level by "Non-leukemia and healthy bone marrow"

{'top1': 0.8571428571428571, 'top5': 0.9761904761904762}

Data for

Generate

Number of shared important

|  |  |
| --- | --- |
| 16 | 1 |
| 15 | 17 |
| 14 | 31 |
| 13 | 30 |
| 12 | 51 |
| 11 | 64 |
| 10 | 75 |
| 9 | 110 |
| 8 | 101 |
| 7 | 128 |
| 6 | 111 |
| 5 | 103 |
| 4 | 116 |
| 3 | 125 |
| 2 | 130 |
| 1 | 215 |

Shared important at least 10

