Computational Molecular Medicine

HW 4 |Wei Jiang | April 12th, 2016

Steps for this HW project:

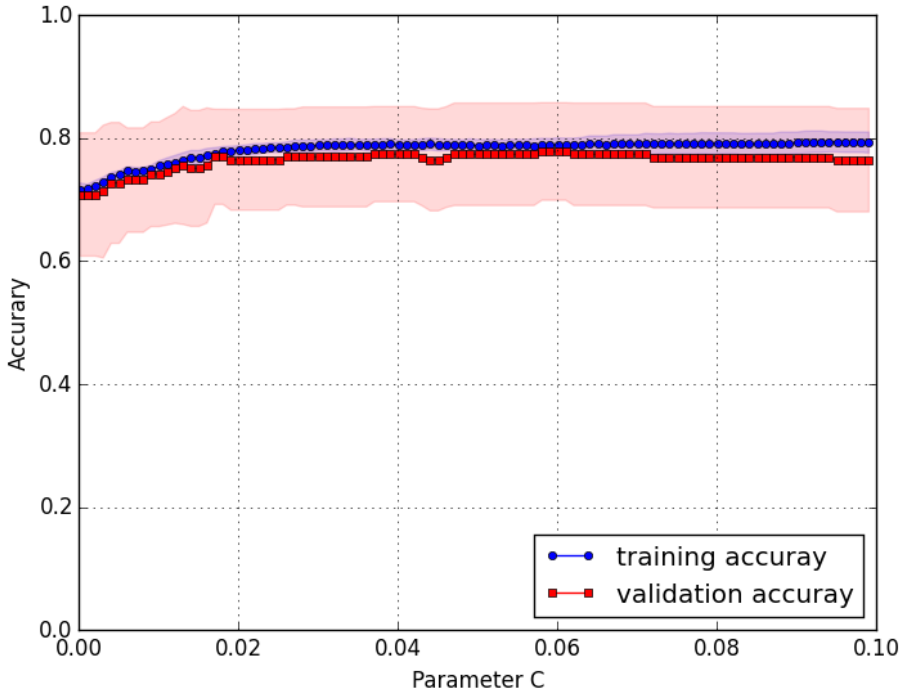
1. First, I used Wilcoxon rank sum test to filter the genes. I tried retaining 10 vs 100 vs 1000 genes.
2. I trained my model only using the training data.
3. I used logistic regression with l2 regularization.
4. Then I trained the model using training data and tuned the hyper-parameter of logistic regression using 10 fold cross validation. The performance measure is accuracy. For the cross validation, data was sampled using stratified random sampling so that the original class balance was retained. In other words, the ratio of number of relapsed to not relapsed patients was maintained across the different folds as the original data set.
5. Then I used the tuned hyper-parameter to fit the entire training data again. The optimal hyper-parameter I found is 0.06 (using the model which has 10 genes). After that, I did prediction on the test data set using the trained model and plotted the ROC curve.
6. By looking at the ROC curve, we can chose the specificity on the test dataset that can be achieved while maintaining 80% sensitivity. Specificity is just 1 minus false positive rate.
    - Using 1000 genes, the AUC score is 0.7 and the corresponding specificity is $1 - 0.62 = 0.38$
    - Using 100 genes, the AUC score is 0.64 and the specificity is 0.4
    - Using 10 genes, the AUC score is 0.65 and the specificity is 0.35

    So actually, using less number of genes gave better performance in this case. However, the AUC score is not high, so more complex and nonlinear model should be able to get better results.

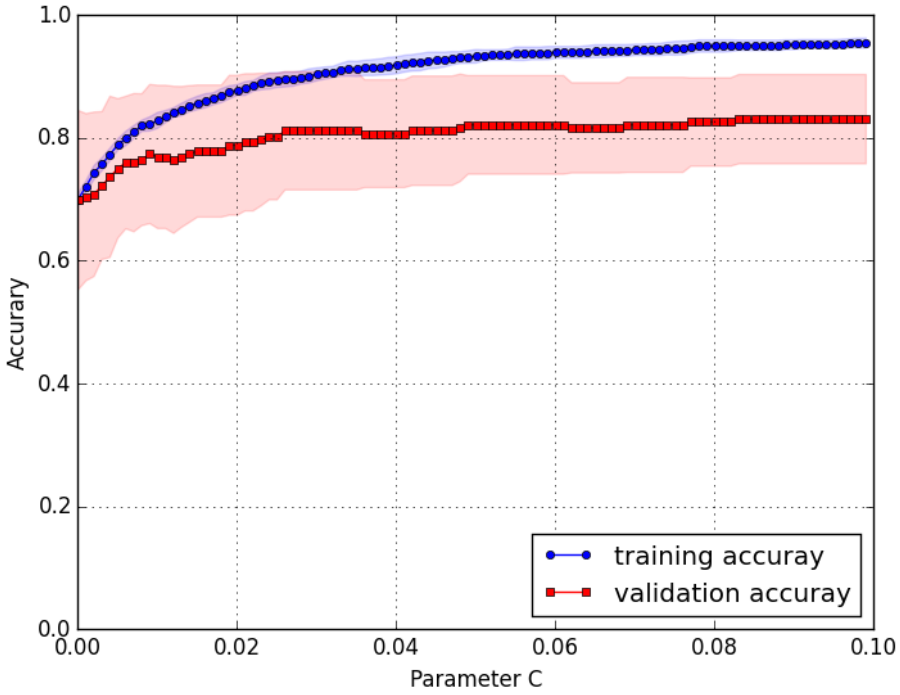
Note: I also used principle component analysis to compress the data into two-dimensional spaces for visualization and see whether the data can be separated well using linear decision boundaries. It turns out the separation was not bad using linear decision boundary so I chose to use simple logistic regression model with l2 regularization.

Following figures shows the validation curve for tuning parameter, principle component plots and ROC curve corresponding to using 10, 100 and 1000 genes:

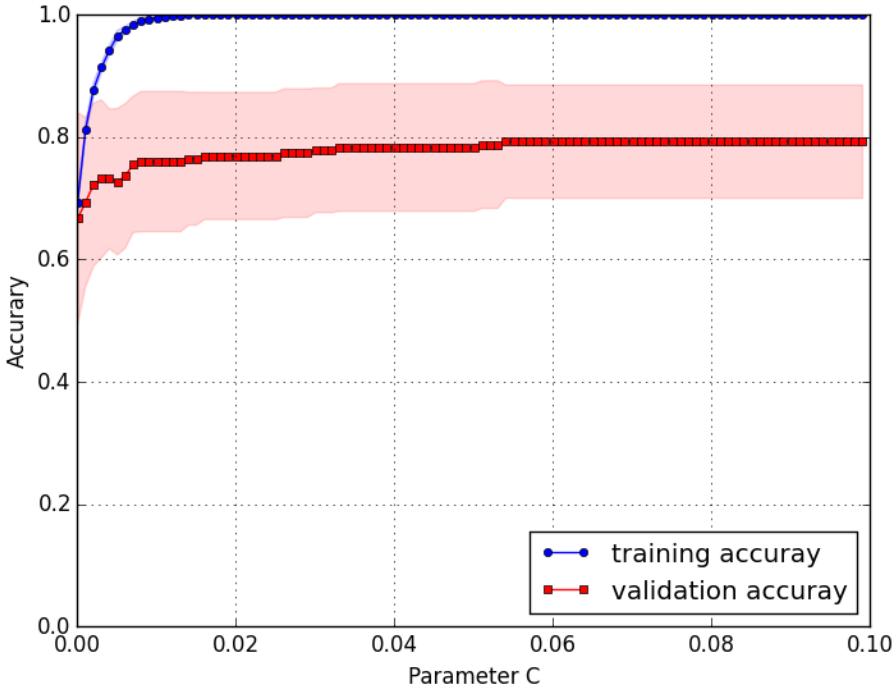
1. Validation curve for model selection (tuning hyper-parameter)
   We can clearly see that the model overfits the data when using 100 or 1000 genes. The training accuracy is systematically higher than the validation accuracy. So I chose the use 10 genes which has low bias but a little bit high variance.
    - Using 10 genes



- Using 100 genes

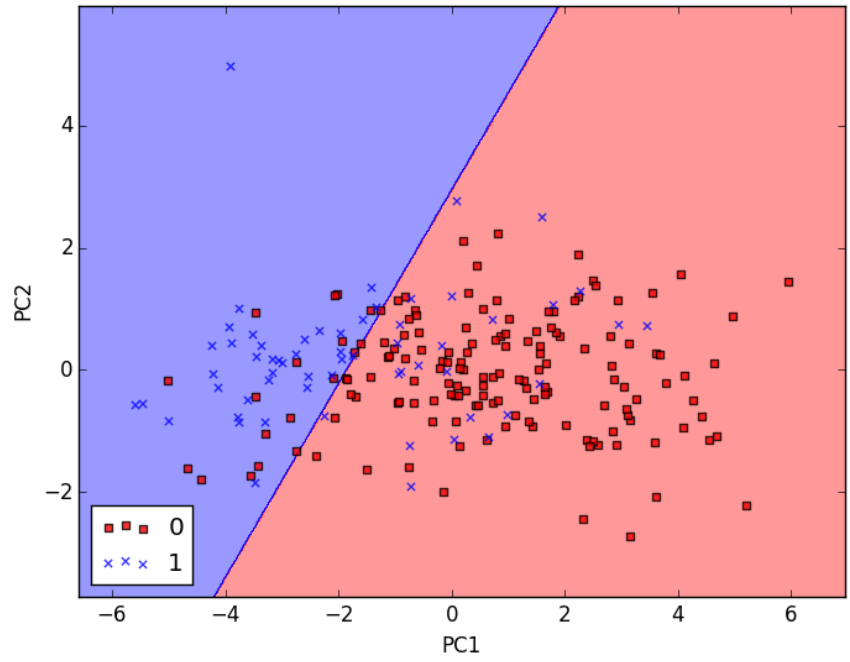


- Using 1000 genes



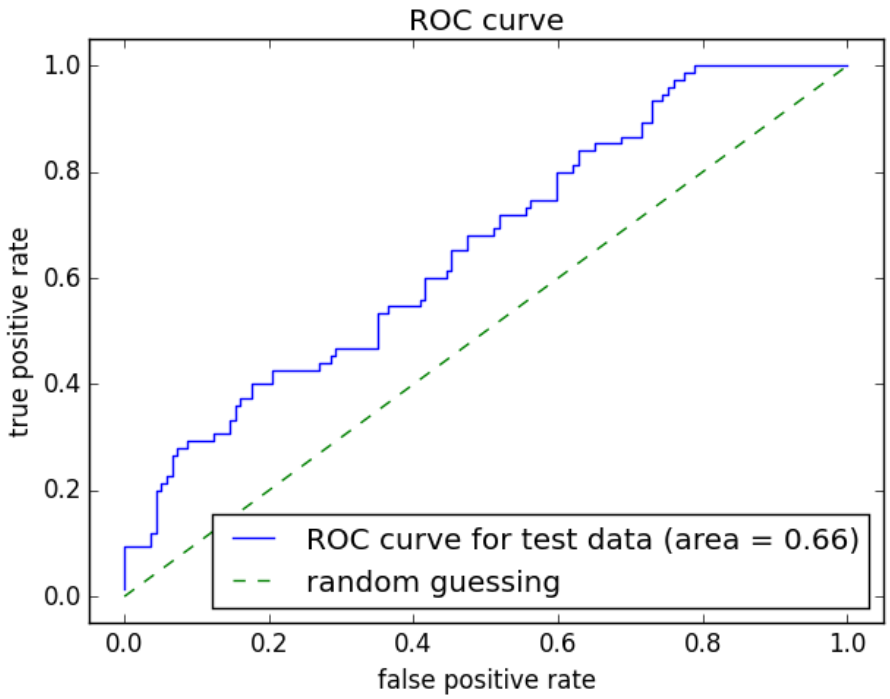
Principle component plot and logistic regression decision boundary trained using two principle component. These plots are most just for data visualization:

- 10 genes:

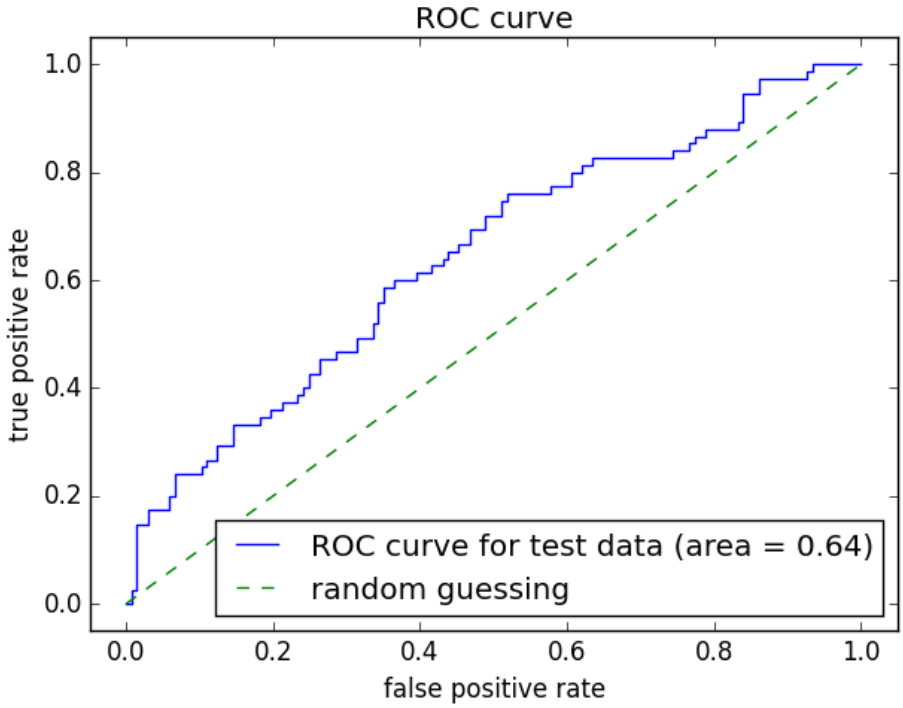


ROC curves for test dataset:

- 10 genes:



- 100 genes:



- 1000 genes: