A Regularized Binomial Logistic Regression Approach to Cancer Classification Using Gene Expression

Joshua Horowitz and Mukund Jayaram

Abstract

Using data retrieved from the University of California, Irvine Machine Learning repository, we combined two datasets - that of individuals' cancer variants and their respective gene expression values. In turn, we used the 20,265 gene expression values given for each individual as predictors for the cancer variant. As Breast Cancer (BRCA) is the most common variant of cancer and currently faces significant hurdles in the diagnostics process, we focused on distinguishing between BRCA and non-BRCA variants. We then performed a L1-regularized logistic regression to determine the pertinent predictors of breast cancer. The results of our research culminated with testing the accuracy of our model, which successfully predicted the individual's breast cancer profile 93.7% of the time.

1 Background and Significance

Over 1.6 million people are diagnosed with cancer every year in the United States, and over 500,000 die from it (CDC). It is the second leading cause of death worldwide (World Cancer Day), accounting for 1 in 6 deaths around the world (WHO). Diagnosing and treating cancer costs 21 billion dollars every year (US News). Therefore, improving diagnosis of various cancer types is important from both a financial and humanistic perspective. Currently, cancer screening is often invasive. In the case of breast cancer, the primary method of diagnosis is through mammograms which painfully compress the breast (Armstrong et al., 2007). Additionally, the mammogram exposes individuals to radiation levels that contribute to higher incidence of breast cancer following screening (Corcos, 2020).

Through our research, we hope to create headway on an alternative method of diagnosing mammary carcinoma. In our work, we draw a correlation between the presence of specific genes and the BRCA variant of cancer. These efforts to trace tumor lineages will support existing research and development of cancer treatments personalized for individuals' genomic profile. Hopefully, as a result of our research, simple saliva swabs will be able to replace mammograms for women everywhere. Furthermore, this could be an opportunity to democratize access to cancer diagnostics, as mammogram machines could range into thousands of dollars whilst analysis of DNA swabs run a mere couple hundred dollars (NIH Human Genome Institute).

2 Methods

a. *Data collection.* The data was collected as part of the Cancer Genome Atlas Program (TCGA). It was made available by Samuele Fiorini (University of Genoa) on UC Irvine's Machine Learning Repository. The sample of RNA sequences was extracted from 801 people, and contains information on 20,531 genes. Each of the people were patients afflicted by one of the following five cancer variants:

- LUAD - A common lung cancer (18% of cases)
- BRCA - All breast cancer (37% of cases)
- PRAD - Prostate adenocarcinoma (17% of cases)
- KIRC - Kidney renal clear cell carcinoma (18% of cases)
- COAD - A common colon cancer (10% of cases)

b. *Variable creation.* There are 20,531 different genes sequenced as part of our data set. These are our predictors (independent variables). Each of the predictors is a gene expression value, which is the average number of cells that express the gene within a cluster of cells sampled. Therefore, these variables are all coded numerically. There is one response variable, the type of cancer, with 5 different levels, one for each of the cancer variants listed above. We combined the file for the gene expression values with the file for the cancer variants to get one consolidated data frame.

We recoded our data to obtain a binary categorical variable (if a person has BRCA or not). We used this as our response variable instead of the 5 cancer variants for two main reasons: 1) breast cancer is the most common form of cancer and 2) traditional mammogram screenings are painful and controversial due to the radiation exposure experienced by women. An alternative method for predicting breast cancer will be useful as a substitute for mammograms.

c. *Analytical Methods* We partitioned the dataset into training and testing data using a 80-20 split based upon the response variable. This maintains the proportion of cancer types in both groups. We performed a regularized multivariable binomial logistic regression to determine the likelihood of an individual having BRCA, given their combination of gene expression values. We did this using L1 regularization. We determined the predictive success of our model using testing data partitioned from the sample. In all of these methods, we extensively used the **glmnet** package in R.

3 Results

The optimal lambda (regularization parameter) used in our L1 regularization was found through cross-validation (*see Appendix Figure 2*). Thereafter, the regularization was performed, such that the number of non-zero coefficients shrunk from 20,265 predictors down to 162. Our L1-regularized logistic regression model successfully classified all 642 individuals from the training data. It also successfully classified 149/159 individuals based upon their gene expression values in the test data. These simulations yield accuracy rates of 100% and 93.7% respectively.

Table 1: Confusion Matrix: L1 Regularized Logistic Regression Model on Training Data

		Actual	
		Other	BRCA
Predicted	Other	402	0
	BRCA	0	240

Table 2: Confusion Matrix: L1 Regularized Logistic Regression Model on Testing Data

		Actual	
		Other	BRCA
Predicted	Other	93	4
	BRCA	6	56

4 Discussion

The model makes no errors upon classifying the training data; this would suggest that it may be overfitted. However, when it comes to classifying the testing data, the model performs very strongly as well. The model has a Type I error rate of 3.8% and a Type II error rate of 2.5%. This is intentional as we sought to minimize the number of false negatives, as cancer detection is a life-or-death matter, and if an individual is not diagnosed in early stages, the tumor may progress and spread to other organs within the body.

Previous research has shown that using convolutional neural networks (CNN) shows promise, but comes with a high risk of false positives and false negatives (Michigan Tech). Similarly, another team of researchers using data from the Federal University Fluminense in Niteroi, Brazil developed a deep learning model which successfully diagnosed breast cancer 99.33% based upon thermal imagery (Mohamed et al., 2022). However, our model is one of the first logistic regressions performed using gene expression values as predictors. We believe that our modeling will be able to contribute to the existing literature and augment our teams' efforts in successfully diagnosing this devastating disease. It is important to note that our research findings are limited by a lack of sample size, as we only evaluated 801 individuals and were given 20,265 gene expression values as predictors. If this model could be used by larger corporations, such as Ancestry or 23 and Me, which have access to exceptionally large datasets, perhaps, this model would have an accuracy rating on par with other teams.

An important caveat of our work is that our model is not readily interpretable due to 162 existing predictors, all of which are gene expression values. Due to our limited biology and genomics expertise, we are not qualified to further explain the significance of each gene.

5 References

R Packages

1. Friedman, Jerome, et al (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. Retrieved from: https://www.jstatsoft.org/v33/i01/.
2. Honaker, James (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. Retrieved from: https://www.jstatsoft.org/v45/i07/.
3. Kassambara, Alboukadel and Mundt, Fabian (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. Retrieved from: https://CRAN.R-project.org/package=factoextra.
4. Kuhn, Max (2021). caret: Classification and Regression Training. R package version.0-90. Retrieved fro: https://CRAN.R-project.org/package=caret.
5. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.R-project.org/.
6. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. Retrieved from: http://www.rstudio.com/.
7. Wickham, Hadley. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
8. Wickham, Hadley, et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. Retrieved from: https://doi.org/10.21105/joss.01686.
9. Wickham, Hadley, et al. (2021). readr: Read Rectangular Tex Data. R package version 2.1.1. Retrieved from: https://CRAN.R-project.org/package=readr.

Literature

1. Armstrong, Katrina, et al. (2007). *Screening Mammography in Women 40 to 49 Years of Age: A Systematic Review for the American College of Physicians.* Annals of Internal Medicine, vol. 146, no. 7, 2007, p. 516. Retrieved from: https://doi.org/10.7326/0003-4819-146-7-200704030-00008.
2. Brown, D. D. (1981). Gene Expression in Eukaryotes. *Science*, *211*(4483), 667–674. Retrieved from: http://www.jstor.org/stable/1685602.
3. Centers for Disease Control and Prevention (2022, February 8). *Cancer: CDC works to prevent cancer and improve the health of people with cancer.* CDC. Retrieved from: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/cancer.html.
4. Corcos D, Bleyer A (January 2020). *Epidemiologic Signatures in Cancer.* The New England Journal of Medicine. 382 (1): 96. Retrieved from: https://www.nejm.org/doi/full/10.1056/NEJMc1914747.
5. Fiorini, Samuele (2016). Gene Expression Cancer RNA-Sequence Data Set [CSV]. Retrieved from: https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq.
6. Michigan Tech News. *Machine Learning Reduces Uncertainty in Breast Cancer Diagnoses.* Michigan Tech University. Retrieved from: https://www.mtu.edu/news/2021/11/machine-learning-reduces-uncertainty-in-breast-cancer-diagnoses.html.
7. Mohamed EA, Rashed EA, Gaber T, Karam O (2022). *Deep learning model for fully automated breast cancer detection system from thermograms.* PLoS ONE 17(1): e0262349. Retrieved from: https://doi.org/10.1371/journal.pone.0262349.
8. U.S. News and World Report (2021, October 26). *Cancer Costs U.S. Patients $21 Billion a Year.* U.S. News and World Report. Retrieved from: https://www.usnews.com/news/health-news/articles/2021-10-26/cancer-costs-us-patients-21-billion-a-year.
9. Weinstein, John N., et al (2013). The cancer genome atlas pan-cancer analysis project. *Nature*, *45*(10), 1113-1120. Retrieved from: https://www.nature.com/articles/ng.2764.
10. Wetterstrand, Kris A (2021, November 1). *The Cost of Sequencing a Human Genome.* National Institute of Health: National Human Genome Institute. Retrieved from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
11. World Cancer Day. *What is Cancer?.* Union for International Cancer Control. Retrieved from: https://www.worldcancerday.org/about-us.
12. World Health Organization. *Cancer: Key Facts.* World Health Organization. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/cancer.

6 Appendix

Figure 1: Prevalence of Cancer Types

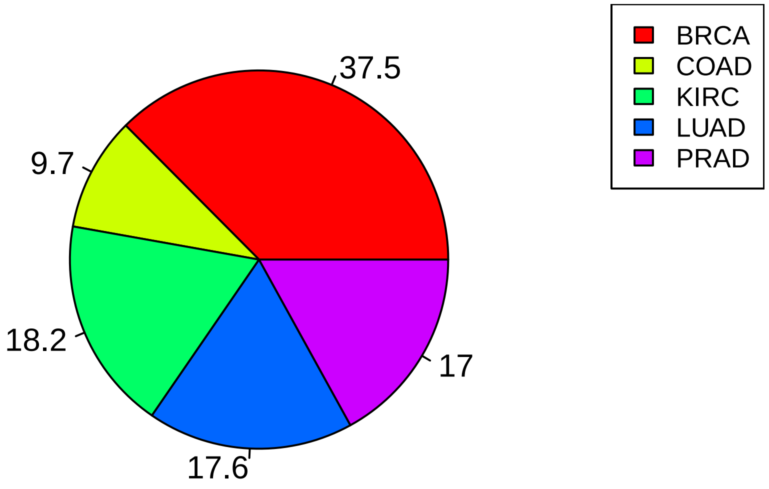
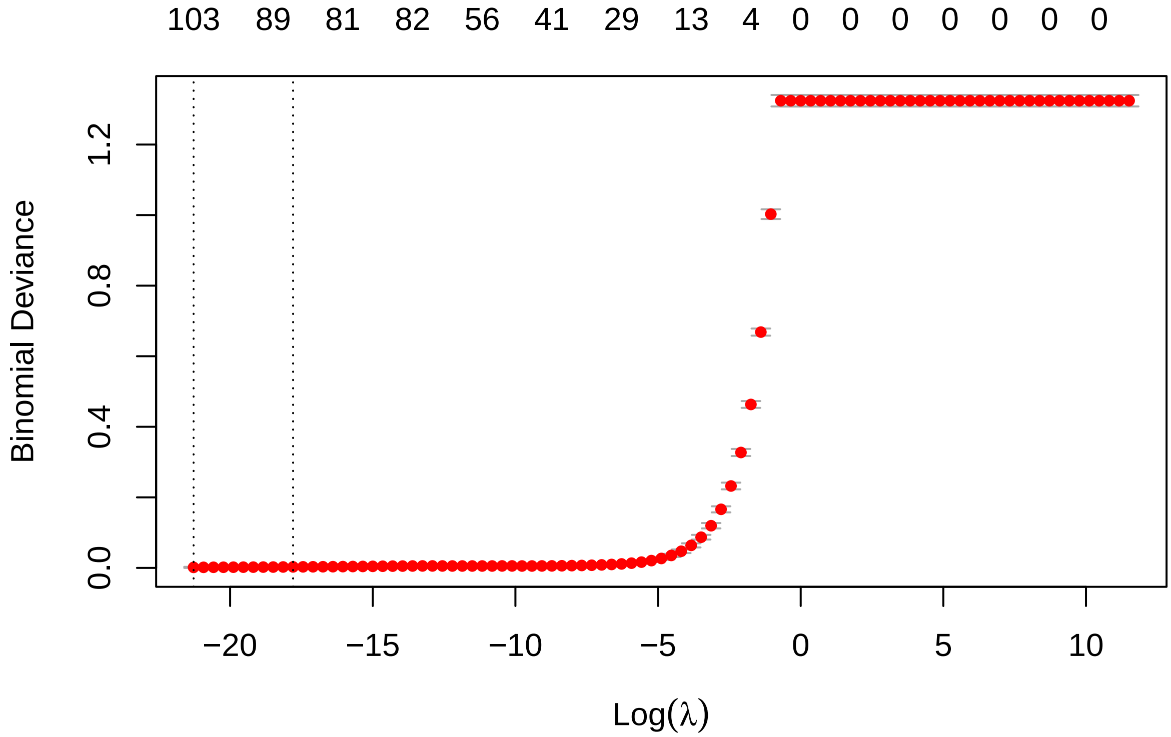


Figure 2: Determining the Optimal Lambda


Binomial Deviance
Log(λ)