My title*

My subtitle if needed

Harrison Jones Liam Wall

December 8, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Problem Statement

1.1 Research Question

The objective of this study is to classify cancer patients into three distinct subtypes using gene expression data, with a dataset containing 12,000 gene expression variables for 880 cancer patients. Moreover, we aim to be able to identify a smaller subset of genes that play a larger role in determining cancer type.

1.2 Introduction

Cancer subtype classification based on gene expression is crucial for understanding the molecular differences between cancer types, guiding personalized treatment plans, and advancing our knowledge of cancer biology. Accurate classification models can aid in improving clinical outcomes by enabling early detection and targeted therapies.

2 Statistical Analyses

2.1 Data

Describe the training and test data used in this dataset. How it comes from kaggle but explain what is in the project documentation about this dataset. Also try and explain what the gene

*Code and data are available at: https://github.com/Lwall02/Cancer-Type-Classification.

expression data is. All data cleanign and analysis was done using the open source statistical programming language R Core Team (2023).

2.1.1 Data Cleaning

Removed non-feature columns such as IDs and other irrelevant information. Ensured that the data is clean and free from missing values or inconsistencies. Made the cancer column into a 3-level factor of 1, 2, and 3.

2.1.2 Data Transformation

The downloaded training and test data both had the gene expressions in log space. This is especially useful for our use of the glmnet model. This is a model that performs Ridge and Lasso regression, where standardization of the input variables helps with the regularization of our parameters. The log transformed data will reduce the effect of outliers and will stabilize the variance. It also makes the data more normally distributed. Lastly, we still maintain interpretability through log data as an increase/decrease of 1 unit in log corresponds to a multiplicative increase/decrease on the original scale. Summary: Standardized the gene expression data to normalize the features and ensure all variables are on the same scale, which is crucial for machine learning models like glmnet.

2.1.3 Feature Selection

No explicit feature selection techniques were applied, as the glmnet model inherently performs regularization and variable selection during training.

2.2 Model

2.2.1 Model Choice

We utilized Multinomial Logistic Regression implemented through the `cv.glmnet` function, a robust method well-suited for high-dimensional datasets where the number of gene expression features far exceeds the number of samples. This method is particularly advantageous in the context of genomic data, where the vast majority of features may be irrelevant or redundant. The LASSO approach, which penalizes the absolute magnitude of coefficients, is particularly effective for feature selection, as it encourages sparsity by shrinking the coefficients of irrelevant features to zero. This inherent feature selection capability makes LASSO an ideal candidate for our classification task. In addition, the flexibility of the cv.glmnet function allows us to compare LASSO and Ridge regularization approaches, giving us the opportunity to determine which method best fits the data and enhances model performance.

2.2.2 Cross Validation

To optimize the regularization parameter, (lambda), and to determine the most appropriate regularization method (LASSO vs. Ridge), we employed 10-fold cross-validation. Cross-validation is a critical technique for assessing the model's ability to generalize to unseen data, reducing the risk of overfitting and providing a more robust estimate of its predictive performance. By performing cross-validation on a wide range of regularization parameters and comparing the performance of both LASSO and Ridge, we ensure that the selected model achieves optimal performance without overfitting to the training set.

2.2.3 Model Justifiation

The choice of the cv.glmnet function is grounded in its ability to handle datasets with a large number of predictors efficiently, providing a versatile framework for regularized regression. The ability to perform LASSO and Ridge regularization allows for flexibility in addressing the complex, high-dimensional nature of genomic data. The regularization of LASSO is particularly beneficial for sparse models, where many of the features may be irrelevant or redundant, making feature selection critical. On the other hand, regularization of Ridge works well when it is assumed that most features contribute to the model, allowing for shrinkage without completely eliminating any features. By utilizing the cv.glmnet function without explicitly specifying whether to use LASSO or Ridge, we ensured an unbiased comparison of both methods to determine the most appropriate model for our dataset.

2.2.4 Training and Evaluation

The model was trained using the given training data to predict the cancer subtype outcome (multinomial classification). Once the optimal value was identified via cross-validation, the final model was trained and used to generate predictions on the testing data. Model performance was evaluated based on its ability to accurately classify patients into the correct cancer subtypes. The effectiveness of the model was gauged by measuring the classification accuracy, ensuring that the selected model generalized well to new data. This iterative approach of training and validation allowed for the identification of the model configuration that best balanced bias and variance, ultimately leading to robust and interpretable predictions. This process of regularization selection and cross-validation ensures that the model is both accurate and generalizable, making it a powerful tool for classifying cancer subtypes based on gene expression data.

2.3 Evaluation Metrics

The primary metric used to evaluate the performance of our models is the testing misclassification error, as it directly reflects the model's ability to correctly classify unseen data.

Misclassification error is calculated as the proportion of incorrect predictions to the total number of predictions, with a lower value indicating better model performance. This metric is particularly relevant in the context of our study, where the goal is to classify cancer patients into three distinct subtypes based on gene expression data. A model that minimizes the misclassification error will be more reliable and applicable in clinical settings, where accurate classification is critical for patient care. Although misclassification error provides a comprehensive measure of performance, we are also interested in selecting the model that is both effective and interpretable. In scenarios where multiple models achieve similar classification accuracy, simplicity and interpretability become important factors in model selection. A model that achieves similar predictive performance but with fewer variables or a clearer structure is preferable, particularly when considering real-world applications in clinical practice. Simplifying the model enhances its usability and interpretability, which is critical for clinicians who need to understand and trust the decisions made by the model. Given this, while we will first focus on the misclassification error to determine which model offers the best classification accuracy, we will also consider the complexity and interpretability of the models. If multiple models achieve comparable results in terms of misclassification error, the less complex and more interpretable model will be favored. In particular, Lasso regression, with its ability to perform feature selection and identify the most relevant genes, may offer valuable insights into the key factors contributing to the cancer subtype classification. This interpretability, in turn, allows for better understanding and potential clinical application of the model, providing a pathway to personalized treatment strategies and further research. Therefore, while the misclassification error is the most important evaluation criterion, model complexity and interpretability will be key considerations in our final selection process.

3 Results and Conclusion

3.1 Model Performance

The model developed using multinomial logistic regression with regularization through cross-validation achieved 100% classification accuracy on the testing dataset. This perfect prediction performance suggests that the model is highly effective in distinguishing between the three cancer subtypes based on the gene expression profiles of the patients. The training and testing process involved rigorous cross-validation, which allowed for optimal tuning of the regularization parameter (lambda), ensuring the model's ability to generalize well to unseen data. The accuracy of the predictions on the testing dataset further supports the robustness of the model, as it indicates the model's consistent ability to correctly classify each cancer subtype without overfitting the training data. Both LASSO and Ridge models were evaluated using the same dataset, and interestingly, both models achieved identical performance in terms of classification accuracy on the testing data, with 100% correct classification. This outcome underscores the strength of both regularization techniques in handling high-dimensional data, where the number of features (gene expression variables) substantially exceeds the number of observations (patients). These results suggest that both methods are highly effective in this context, managing to identify the underlying patterns and relationships in the gene expression data with remarkable precision.

3.2 Conclusion

The results from this analysis demonstrate that multinomial logistic regression, when coupled with regularization techniques such as LASSO and Ridge, can yield highly accurate models for the classification of cancer subtypes based on gene expression data. The perfect classification rates observed with both LASSO and Ridge models provide compelling evidence of the model's accuracy and reliability in distinguishing between cancer subtypes, even in the face of high-dimensional data. Although both LASSO and Ridge models achieved the same level of performance in terms of classification accuracy, LASSO's inherent feature selection capabilities make it particularly valuable. LASSO's ability to shrink coefficients to zero for less relevant features allows for a more interpretable model, enabling the identification of the most influential genes involved in cancer subtype classification. This interpretability is a key advantage, as it provides actionable insights into the genetic markers that contribute to the differentiation of cancer subtypes. By pinpointing the genes most relevant to cancer classification, this model could facilitate further research into the molecular underpinnings of cancer and inform the development of targeted therapies. Given the excellent performance of the models and the interpretability offered by LASSO, this approach holds significant potential for deployment in clinical settings. With continued validation and refinement, such a model could support more accurate cancer diagnoses, inform personalized treatment plans, and contribute to a deeper understanding of the genetic factors driving cancer progression. The results of this study, therefore, represent an important step toward the integration of machine learning models in precision medicine, particularly in the context of cancer diagnostics and subtype classification.

4 Discussion

4.1 Limitations

While the model achieved perfect classification accuracy on the test data, it is essential to acknowledge several limitations that may impact the robustness and generalizability of these findings. One of the primary concerns is the potential for overfitting, especially given the high-dimensional nature of the dataset. In this study, the number of features greatly exceeds the number of samples, which increases the risk of the model fitting to noise or spurious patterns present in the data. Although cross-validation was employed to mitigate this risk, the simplicity of the dataset and relatively small sample size in comparison to the number of features still raise concerns about the model's generalizability to other populations. Another important limitation lies in the lack of external validation. While the model exhibited perfect performance on the provided test dataset, its ability to generalize to data from different sources or patient cohorts remains unverified. External validation using independent datasets—preferably with diverse cancer populations—would be necessary to assess the true robustness of the model and its ability to maintain high classification accuracy across varying conditions. Without such validation, there is a risk that the model's performance may not be representative of its practical applicability in diverse clinical settings. Furthermore, the gene expression profiles used in this study may not fully capture the complex and multifactorial nature of cancer. Cancer subtypes are influenced by a wide range of genetic, clinical, and environmental factors. The dataset used in this analysis is limited to gene expression data, and does not include other potentially significant information such as patient demographics, clinical history, or environmental exposures. This limitation may reduce the model's ability to account for all potential variations in cancer subtypes, as these non-genetic factors can also play a significant role in tumor progression and differentiation.

4.2 Future Directions

Given the results of this, several avenues for future research and model improvement are worth exploring. One potential direction is the investigation of more complex machine learning models, such as deep learning architectures or ensemble methods (e.g., random forests, gradient boosting), which may improve classification accuracy and enhance the model's ability to generalize to more complex datasets. These methods, while computationally more intensive, have the potential to capture nonlinear relationships and higher-order interactions within the data that simpler models like logistic regression may miss. Another promising direction is to integrate additional forms of clinical data into the model. By incorporating features such as patient demographics, clinical history, and treatment responses, the model could achieve a more holistic view of cancer progression. This multimodal approach could potentially improve classification accuracy and make the model more robust, as it would consider a broader range of factors beyond gene expression. Moreover, combining genomic and clinical data has the potential to yield insights into the complex interplay between genetics, environment, and disease, providing a more nuanced understanding of cancer subtypes. Finally, external validation of the model is a critical step in assessing its practical applicability in real-world scenarios. Performance evaluation on independent and diverse datasets from different geographic regions, cancer cohorts, and medical institutions is essential to ensure the model's robustness and to determine its effectiveness in clinical practice. Such external validation will help identify potential biases in the data, assess the model's ability to generalize across populations, and confirm whether the model's 100% classification accuracy on the testing dataset can be replicated in broader, more varied contexts. In summary, while the model developed in this study shows promise, addressing the limitations discussed above and exploring these future research directions will be essential for improving its robustness, interpretability, and practical application in clinical settings. By extending this work to incorporate additional data sources and more complex modeling techniques, future research can contribute significantly to the development of more accurate and reliable cancer classification tools, ultimately enhancing patient outcomes and advancing personalized cancer treatment strategies.

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.