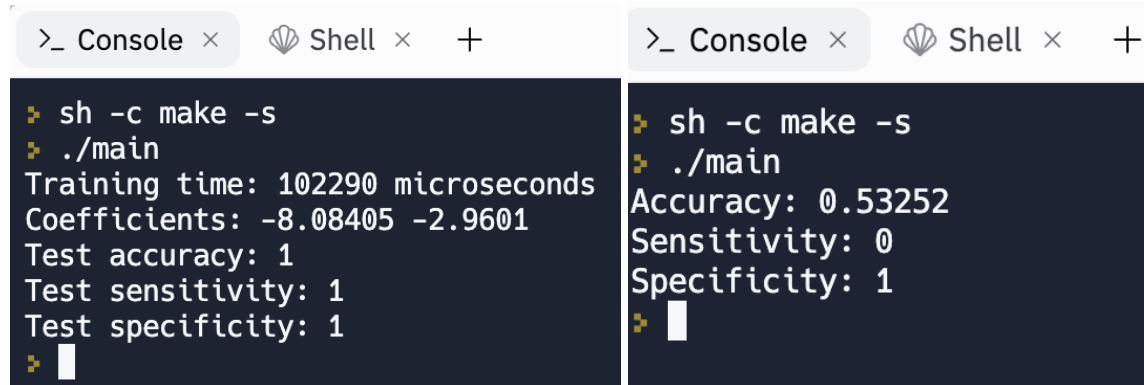


Anurag Diwate - ASD190004

Gaurang Goel - GXG190015

a.



The image shows two terminal windows side-by-side. Both windows have tabs labeled '>_ Console' and 'Shell'. The left terminal shows the output of running 'sh -c make -s' and './main', displaying training time, coefficients, and perfect test accuracy, sensitivity, and specificity. The right terminal shows the output of running 'sh -c make -s' and './main', displaying accuracy, sensitivity, and specificity.

```
>_ Console x Shell x +
```

```
> sh -c make -s
> ./main
Training time: 102290 microseconds
Coefficients: -8.08405 -2.9601
Test accuracy: 1
Test sensitivity: 1
Test specificity: 1
> 
```

```
>_ Console x Shell x +
```

```
> sh -c make -s
> ./main
Accuracy: 0.53252
Sensitivity: 0
Specificity: 1
> 
```

b. Using the synthetic data, the logistic regression model achieved a test accuracy of 1, suggesting that it was able to properly categorize the test data. Nevertheless, the accuracy of the Titanic data was just 0.53252, demonstrating that the model did not perform as well on the real-world dataset. The sensitivity was zero, suggesting that the model was unable to accurately detect positive instances, which in this case would be persons who died. The specificity was one, suggesting that the model accurately identified the negative instances, in this case, the ones who lived. There might be various reasons why the model did not perform as well with Titanic data. One reason might be that the synthetic data was created with a clear distinction between the two groups, but real-world data is likely to be more complicated and lack such a distinction. Another cause might be that the model's features were insufficiently informative to discriminate between the two classes, and new features or more advanced feature engineering may be required.

c. Machine learning methods used for classification applications include generative classifiers and discriminative classifiers. Generative classifiers employ Bayes' rule to determine the conditional probability of the class label given the input features by modeling the joint probability distribution of the input features and the class labels. In contrast, discriminative classifiers explicitly predict the conditional probability of the class label given the input characteristics. Generative classifiers have the benefit of being able to represent the whole probability distribution of the input characteristics and class labels, which can be valuable for tasks such as missing data imputation or synthetic data synthesis. They also perform well when there are few training examples available since they can estimate the class label using the complete probability distribution. When the focus is on the classification job itself, discriminative classifiers perform better since they explicitly predict the likelihood of the class label given the input characteristics. Since they do not have to represent the complete probability distribution, discriminative classifiers can be more computationally efficient.

d. Reproducibility in the context of Machine Learning is the ability to recreate a machine learning workflow or algorithm to reach the same conclusion as the original machine learning workflow [1]. The significance of reproducibility is that other researchers are able to then replicate the workflow and achieve the same results. The advantages of this are that other researchers can validate the conclusions and methodology, which boost trust in the findings [1]. It also “reduces or eliminates variations when rerunning failed jobs or prior experiments, making it essential in the context of fault tolerance and iterative refinement of models [2]. But there are pitfalls that need to be overcome to achieve reproducibility, such as data leakage [3]. This can lead to overoptimistic determinations of an algorithm’s performance. Ultimately, to achieve reproducibility, we should use version control systems, document as many steps taken to achieve the workflow as possible, and use open source platforms to host detailed information about the workflow.

References:

- [1] Hemant, Priti. "Reproducible Machine Learning: A step towards making ML research open and accessible", 17 February 2020,
<https://towardsdatascience.com/reproducible-machine-learning-cf1841606805>.
- [2] Villa, Jennifer, and Zimmerman, Yoav. "Reproducibility in ML: why it matters and how to achieve it", 25 May 2018, <https://www.determined.ai/blog/reproducibility-in-ml>.
- [3] Kapoor, Sayash, and Narayanan, Arvind. "Leakage and the Reproducibility Crisis in ML-based Science", 14 July 2022, <https://reproducible.cs.princeton.edu/>.