1. Outputs:

* Logistic regression (calculating coefficients 50000 times):

Text

Description automatically generated

* Naïve Bayes:

Text

Description automatically generated

1. Coincidentally, both the models had the same accuracy (78.4553%) for the training data in the Titanic dataset. They also had the same sensitivity (0.695652) and specificity (0.862595). It took much longer to run the logistic regression model since the weights were calculated 50,000 times.
2. Logistic regression, which directly estimates the parameters of P(Y|X), is a discriminative classifier. On the other hand, Naive Bayes directly estimates parameters for P(Y) and P(X|Y) and is called a generative classifier (Mazidi, 141). [As stated by Chirag Goyal, “Discriminative models draw boundaries in the data space, while generative models try to model how data is placed throughout the space. A generative model focuses on explaining how the data was generated, while a discriminative model focuses on predicting the labels of the data” (Goyal, 1).] Generative classifiers don’t need as much data to train (as they make stronger assumptions) and can also work with missing data (Goyal, 1). However, they are less accurate than discriminative models (Goyal, 1). Both models have different applications.

If the Naive Bayes independence assumptions hold, and the number of training examples grows towards infinity, the Naive Bayes and logistic regression converge toward similar classifiers (Mazidi, 141). Both classifiers can be used to create classification models, but they work in different ways. They both use conditional probability to classify data.

Sources:

Goyal, Chirag. “Deep Understanding of Discriminative and Generative Models.”

Analytics Vidhya, 19 July 2021, <https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/#:~:text=Discriminative%20models%20draw%20boundaries%20in,the%20labels%20of%20the%20data>.

Mazidi, Karen. Machine Learning Handbook Using R and Python

1. Research in machine learning is reproducible if another researcher can use the same source files to obtain the same results as the original researcher (by implementing the same analysis and models) (Ding, 1 and Goodman et al, 1). It is obviously a common practice for researchers, and an important one because it can ensure that the results achieved by a specific machine learning algorithm/model are correct (Ding, 1). This way, false claims can be debunked easily (Ding, 3). In addition, it can help others understand what was done, as well as reduce the chance of errors happening (Ding, 1).

According to Ding, there is more than one type of reproducibility (Ding, 2). Methods reproducibility is “the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results as in an original work” (Ding, 2). Essentially, enough detail about the procedures in the study (at least the code, data, and overall result) should be provided for it to be replicated accurately (Ding, 2). Results reproducibility deals with producing the same experimental results after following the same procedures as the research we’re trying to replicate (Ding, 3 and Goodman et al., 2). For these two types of reproducibility, changes in data and code need to be recorded. In inferential reproducibility, the independent replication of a study leads to similar results as the original study (Ding, 4). All these reproducibility types have their pitfalls.

Sources:

Ding, Zihao. “5 - Reproducibility.” Machine Learning Blog | ML@CMU | Carnegie

Mellon University, 24 Aug. 2020, https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/.

Goodman, Steven N., Daniele Fanelli, and John PA Ioannidis. “What does

research reproducibility mean?.” Science translational medicine 8.341 (2016): 341ps12-341ps12.