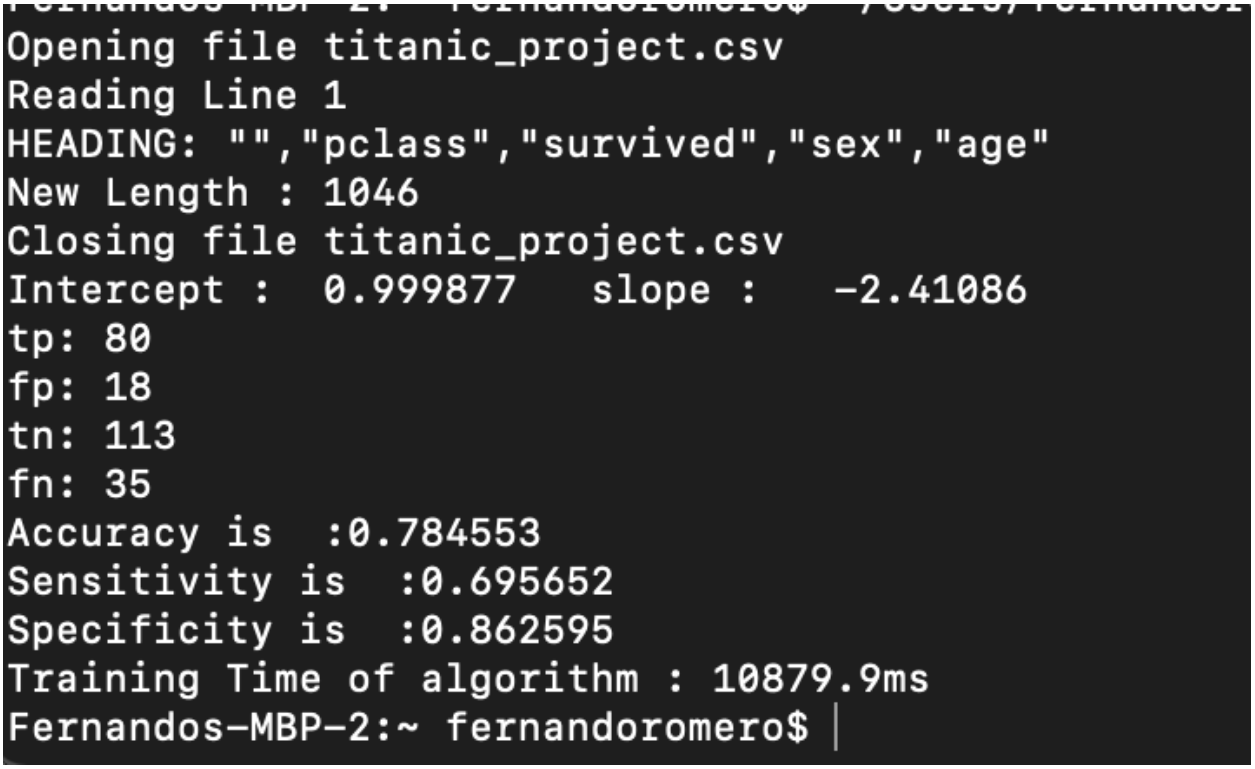
CS 4375 Intro to Machine Learning

Manuel Romero
Sagar Darnal

Project 1: Logistic Regression



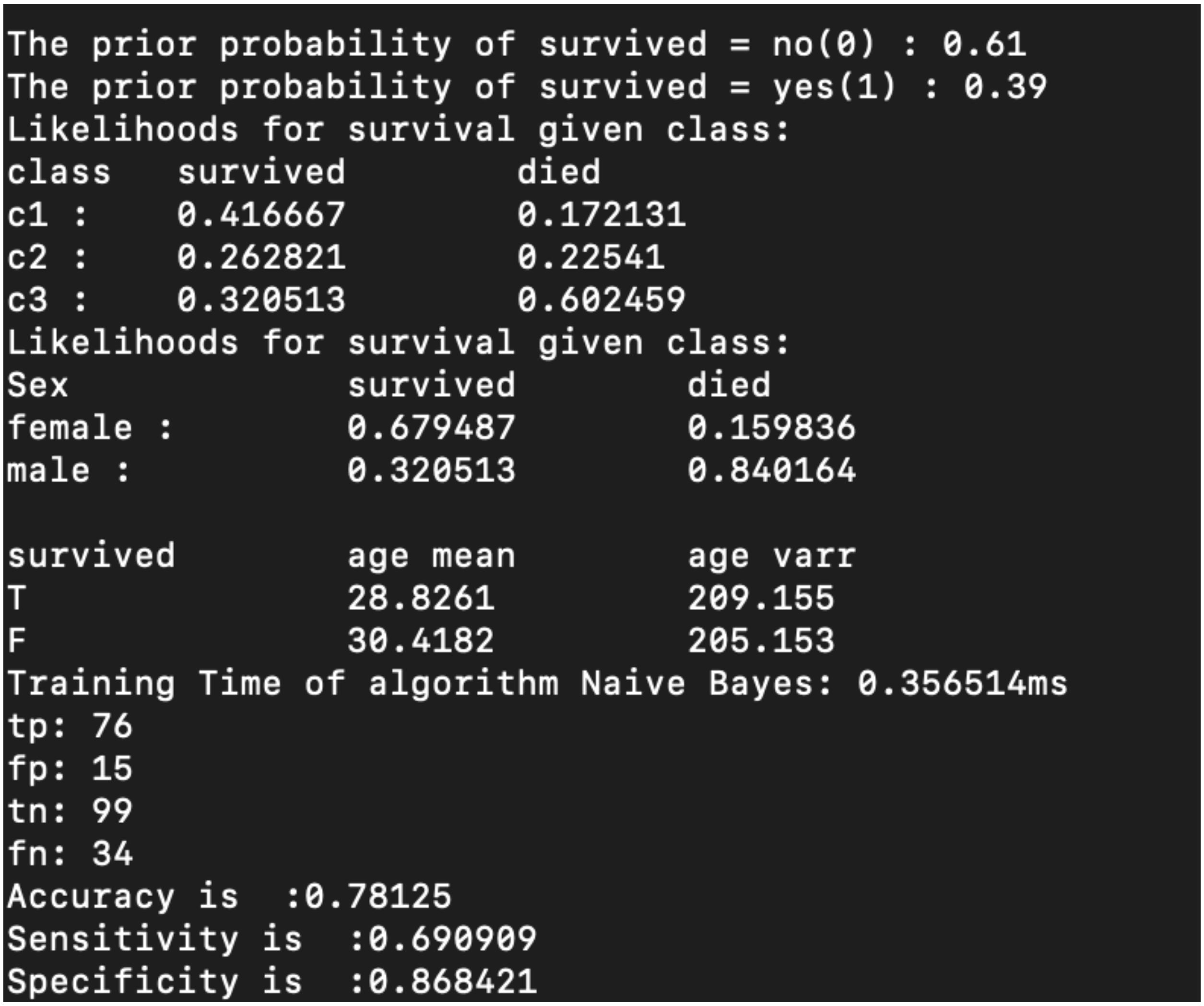
```
Opening file titanic_project.csv
Reading Line 1
HEADING: "","pclass","survived","sex","age"
New Length : 1046
Closing file titanic_project.csv
Intercept :  0.999877   slope :   -2.41086
tp: 80
fp: 18
tn: 113
fn: 35
Accuracy is  :0.784553
Sensitivity is  :0.695652
Specificity is  :0.862595
Training Time of algorithm : 10879.9ms
Fernandos-MBP-2:~ fernandoromero$ 
```

The statement refers to the application of logistic regression to predict whether a passenger survived or not based on their sex. The intercept and slope values of the model were found to be .999877 and -2.41086, respectively. These values were used to compute the accuracy, sensitivity, and specificity of the algorithm.

The accuracy of the algorithm was found to be 78.4553%, which indicates that the model was able to make correct predictions for approximately 78% of the data. The sensitivity and specificity values of 69.5652% and 86.2595%, respectively, indicate the algorithm's ability to correctly identify true positives and true negatives.

The statement concludes that the algorithm is decent and moderately effective in making accurate predictions based on the given data. It also states that an accuracy of 78% is decent in this context. Additionally, the run time of the algorithm was 10871.9 ms, which may provide insight into the efficiency of the algorithm.

Program 2: Naïve Bayes



```
The prior probability of survived = no(0) : 0.61
The prior probability of survived = yes(1) : 0.39
Likelihoods for survival given class:
class   survived        died
c1 :    0.416667        0.172131
c2 :    0.262821        0.22541
c3 :    0.320513        0.602459
Likelihoods for survival given class:
Sex             survived        died
female :        0.679487        0.159836
male :          0.320513        0.840164

survived        age mean        age varr
T               28.8261         209.155
F               30.4182         205.153
Training Time of algorithm Naive Bayes: 0.356514ms
tp: 76
fp: 15
tn: 99
fn: 34
Accuracy is  :0.78125
Sensitivity is  :0.690909
Specificity is  :0.868421
```

The statement discusses the performance of Naive Bayes in predicting whether a passenger survived or not based on their sex. The model's accuracy, sensitivity, and specificity were found to be 78.125%, 69.0909%, and 86.8421%, respectively. The performance of Naive Bayes was found to be comparable to logistic regression, with slightly lower accuracy and sensitivity, but slightly higher specificity.

However, the statement highlights that Naive Bayes had a significantly shorter training time compared to logistic regression. This difference in training time may be attributed to the fact that Naive Bayes is a generative model, while logistic regression is a discriminative model.

CS 4375 Intro to Machine Learning

Paragraphs comparing Generative classifiers versus Discriminative classifiers.

Generative and discriminative classifiers are two broad categories of classification algorithms used in machine learning. A generative classifier models the joint distribution of the input features and the class labels. It estimates the probability density function of the features for each class and uses Bayes' theorem to compute the probability of a particular class given the input features. On the other hand, a discriminative classifier directly models the conditional probability of the class labels given the input features. It learns a decision boundary that separates the classes in the feature space.

One major difference between generative and discriminative classifiers is their ability to handle missing or noisy data. Generative models can handle missing data by estimating the probability distribution of the features from the available data. Discriminative models, on the other hand, require complete data and may not work well when there is missing data. Another difference is their ability to model complex data distributions. Generative models can model complex data distributions more effectively as they model the joint probability distribution of the input features and the class labels. Discriminative models, however, are simpler and may work better for high-dimensional feature spaces where modeling the joint distribution may be difficult.

In conclusion, both generative and discriminative classifiers have their advantages and disadvantages, and the choice of the classifier depends on the nature of the problem at hand. Generative classifiers work well for handling missing data and modeling complex data distributions, while discriminative classifiers are simpler and work well for high-dimensional feature spaces. A combination of both approaches may also be used to improve classification performance.

Sources:

"*Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*" by Xindong Wu and Vipin Kumar

"*Generative vs. discriminative classification*" by Saeed Izadi, Kevin McGoff, and Wlodek Zadrozny in IEEE Intelligent Systems.

CS 4375 Intro to Machine Learning

Reproducible Research in Machine Learning

Reproducibility is an important concept in machine learning, referring to the ability to run an algorithm and get the same or similar results repeatedly. This means that regardless of the machine or software environment, the results obtained should be consistent. However, this is not always the case as different machines or software environments can cause variations in the results.

To ensure reproducibility, it is important to document the runs of the algorithm and include all important details so that if someone else were to use the algorithm, they could reproduce the same results. This is like scientific experiments, where results need to be valid and replicable.

For instance, if we use a dataset and run an algorithm, the results obtained can vary based on the random sample used, mean, variance, etc. To obtain consistent results, we need to pass on the exact subset of data used for the first run of the algorithm so that others can reproduce the same results. Alternatively, we can specify the number of data points used and their order so that others can account for differences in the result.

In summary, reproducibility is essential in machine learning, as it helps to ensure that the results obtained are consistent and valid, regardless of the machine or software environment used. This can be achieved through proper documentation and sharing of important details about the algorithm and its runs.

Sources:

https://www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applications/#:~:text=Reproducibility%20with%20respect%20to%20machine,reporting%2C%20data%20analysis%20and%20interpretation.

https://www.determined.ai/blog/reproducibility-in-ml