Austin Girouard and Aarya Patil

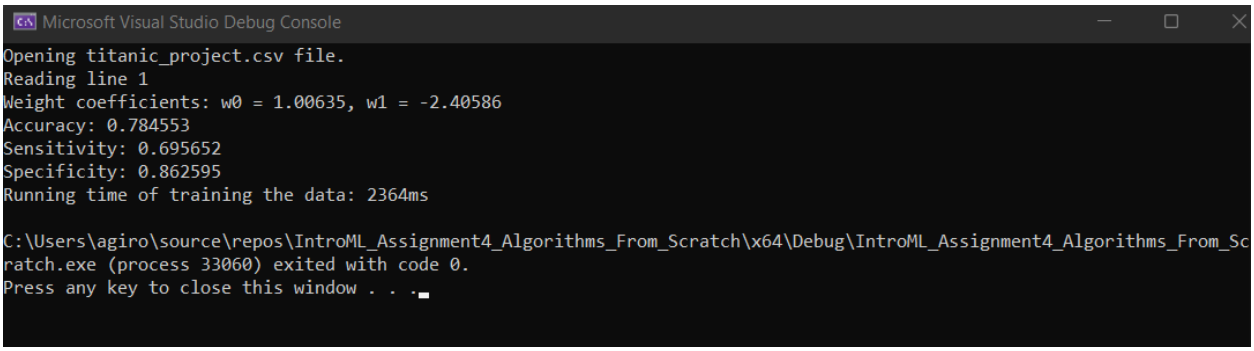
ATG180001 / AAP180014

Dr. Karen Mazidi

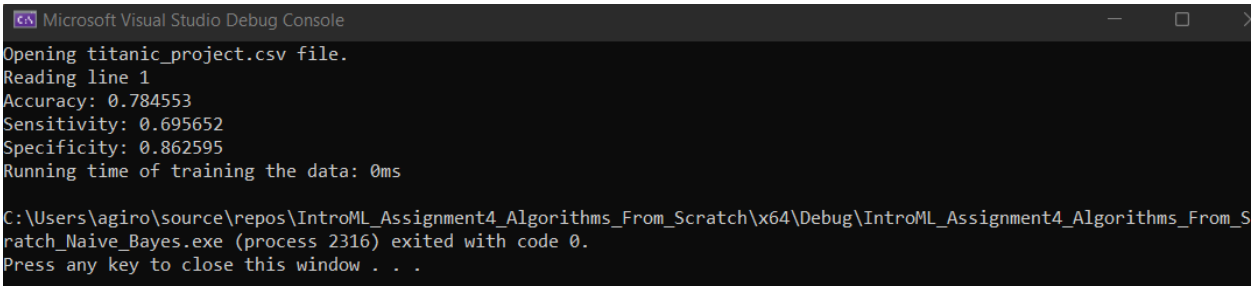
Assignment 4: ML Algorithms from Scratch

a)

Example output for Logistic Regression



Example output for Naïve Bayes.



b)

Both the Logistic Regression and the Naïve Bayes algorithms produced the exact same probabilities, which is interesting because we were at least expecting a small difference in calculations. Both algorithms produced high enough accuracy scores to prove that the model performs much better than 50% correctness (selecting an option at random).

c)

Both Generative and Discriminative Models can both be used for developing learning models based around generating conditional probability. The main difference is that “discriminative models divide the data space into classes by learning the boundaries”, while “generative models understand how the data is embedded into the space” [1]. This makes these two methods of generating learning models very different in practice.

Generative Classifiers “try to model features of the classes” and “predict which class would have most likely generated the given observation,” such as Naïve Bayes [2]. Discriminative Classifiers “learn what features in the input are most useful to distinguish between the various possible classes”, such as Logistic Regression [2].

Sources:

[1] https://www.turing.com/kb/generative-models-vs-discriminative-models-for-deep-learning

[2] https://medium.com/@akankshamalhotra24/generative-classifiers-v-s-discriminative-classifiers-1045f499d8cc#:~:text=Generative%20Classifiers%20tries%20to%20model,likely%20generated%20the%20given%20observation.

Google this phrase: reproducible research in machine learning. Using 2-3 sources, at least one of which should be academic, write a couple of paragraphs of what this means, why it is important, and how reproducibility can be implemented. Cite your sources using any format

d)

Reproducible research in machine learning is “the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator” [1]. So, reproducible research should be able to be duplicated and further investigated by other researchers through an exact documentation on the data, code, and metrics/parameters used in the analysis.

Reproducible research is important because “computational science is facing a credibility crisis: it’s impossible to verify most of the computational results presented at conferences and in papers today,” which do not exhibit reproducibility [2]. In a way, non-reproducible research is the same as uncited work: there is no way to prove whether or not the presented information holds any substantial backing.

Reproducibility can be implemented in research by having researchers include all “parameter values, function invocation sequences, and other computational details [which are] typically omitted from published articles but are critical for replicating results or reconciling sets of independently generated results” [2].

[Sources on next page]

Sources:

[1] https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/

[2] https://staff.washington.edu/rjl/pubs/cise12/CiSE12.pdf