ML Algorithms from Scratch

Code Output

Logistic Regression from Scratch

Weights:
3.41074
-2.41086
Accuracy = 0.784553
Sensitivity = 0.695652
Specificity = 0.862595
Elapsed training time in seconds: 34s

Naïve Bayes from Scratch

Prior probability for not surviving and surviving, respectively:
0.61
0.39

Likelihood values for p(pclass|survived):
0.172131 0.22541 0.602459
0.416667 0.262821 0.320513

Likelihood values for p(sex|survived):
0.159836 0.840164
0.679487 0.320513

Perished Survived
0.421277 0.578723
0.793977 0.206023
0.871095 0.128905
0.226246 0.773754
0.145841 0.854159

Accuracy = 0.784553
Sensitivity = 0.695652
Specificity = 0.862595
Elapsed training time in nanoseconds: 160900ns

Discriminative vs Generative Classifiers

Discriminative classifiers are used to directly estimate the parameters for P(Y|X) using the training data, while generative classifiers estimate the parameters for P(Y) and P(X|Y), and use Bayes Theorem to find P(Y|X).

Discriminative models want to find decision boundaries between classes in classification in hopes of finding a function that takes any input and returns a class label. During training, the conditional probability of P(Y|X) is maximized. In logistic regression, the parameters found are the coefficients.

Generative models focus on how classes are distributed and use algorithms to find a pattern on distribution for the data. During training, the joint probability of P(X, Y) is maximized. These models can make new data points and need less data but are significantly affected by outliers and always assume conditional independence.

Reproducible Research in Machine Learning

Reproducibility is ensuring that a researcher's findings can be replicated by other researchers using the same methods and obtaining the same results. It's important because it encourages transparency and gives researchers confidence that the findings are correct. Reproducibility increases reliability by lowering the possibility of errors. There are three types of reproducibility: methods, results, and inferential.

One issue with attempting to present reproducible research is that sometimes the raw data used in training algorithms are protected information (e.g. health info) and sharing it in a non-identifiable way is infeasible. Additionally, sometimes organizations that have spent years collecting, curating, and maintaining datasets without support from researchers are reluctant to publicize them. Finally, though the code is made public, it may not be possible to simply run it. Specialized workstations, software, and hardware may have been originally used as well as significant pre-processing of the data.

To implement reproducible research, the researcher should record the platform, data, code, and results at the bare minimum. The algorithm, its complexity analysis, the data collection process, the allocation of data, and the computing infrastructure should be described in great detail. If possible, reviews of articles that oppose the author's hypothesis should be included and enabling comments and reviews post-publication to help foster discussion is encouraged for objective evaluation.

Sources

Carter, R. E., Attia, Z. I., Lopez-Jimenez, F., & Friedman, P. A. (2019, May 22). *Pragmatic considerations for fostering reproducible research in Artificial Intelligence*. Nature News. Retrieved March 4, 2023, from https://www.nature.com/articles/s41746-019-0120-2

Ding, Z. (2020, August 24). *5 - reproducibility*. Machine Learning Blog | ML@CMU | Carnegie Mellon University. Retrieved March 4, 2023, from https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/

Goyal, C. (2023, February 14). *Machine learning models: Descriptive & generative ML models*. Analytics Vidhya. Retrieved March 4, 2023, from https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/