Column Editors:
Donald A. Smith, Guilford College, Greensboro, NC; tptastronotes@gmail.com
Janelle M. Bailey, Temple University, Philadelphia, PA

Machine learning in introductory astrophysics laboratory activities

Donald A. Smith, Guilford College, Greensboro, NC

A working knowledge of Artificial Neural Networks[1] is rapidly becoming critical for navigating the modern world. Although the last few years have seen an explosion of the use of these tools in research, and there are many do-it-yourself articles on the web,[2] they have not yet filtered down to wide implementation in introductory courses. I report here on my integration of machine learning activities into a general education course on galaxies and cosmology. I describe four lab activities for image classification, and I reflect on the strengths and weaknesses of using these tools in the context of online instruction during the 2020-21 pandemic academic year.

Commerce, law enforcement, and entertainment are all being shaped by machine learning tools in ways both inspirational and problematic. Scientific research, astrophysics in particular, has seen an explosion in both the enormous data sets and computer processing power appropriate for machine learning tools. Figure 1 shows the exponential growth of the number of papers each year that the SAO/NASA Astrophysics Data System[3] returns on simple searches for abstracts with the text strings "machine learning" and "deep learning."

So far, the teaching of these techniques has been mostly limited to advanced courses in data analytics and computational physics. An identical abstract search in *TPT* and *AJP* returns almost no hits at all, beyond a hilarious 2017 April fool's joke.[4] In the last few years, presentations on machine learning have begun appearing at AAPT regional and national meetings, and two technological developments have made it practical to start teaching machine learning to beginning undergraduates in general education courses: free packages in the Python language such as Keras[5] and TensorFlow[6] have become much easier for beginners to use, and Google has made free high-power computing available through their Colaboratory environment. These factors lower the barriers to beginners' engagement.

In the fall semester of 2020, I taught an intensive three-week online course on Machine Learning for a general education audience (i.e., no specific math or programming prerequisites).[7] Our plans to offer a hands-on observational astronomy course in spring 2021 had to be scrapped due to the pandemic, and we decided to offer an online course in galaxies and cosmology instead. Lab activities were either carried out with household materials or used computer archives and simulations. I incorporated machine learning activities into this course.

To scaffold this topic, I used four lab sessions. In week two of the semester, students completed a worksheet on the Galaxy Zoo citizen science project.[8] I adapted this from Slater, Slater, and Lyons,[9,10] whose instructions for this activity no longer match the Galaxy Zoo website. That website shows students images of galaxies from the Sloan Digital Sky Survey (SDSS)[11] and asks them to try to classify the type of galaxy in each image. Students learned about the different kinds of galaxies (spiral, elliptical, and irregular) and were challenged to consider the dangers in extrapolating from a small sample of data.

In week eight, we returned to the subject of galaxies to study the properties of their distribution in space. I adapted to Python an Excel activity by D. Smith (no relation)[12] that

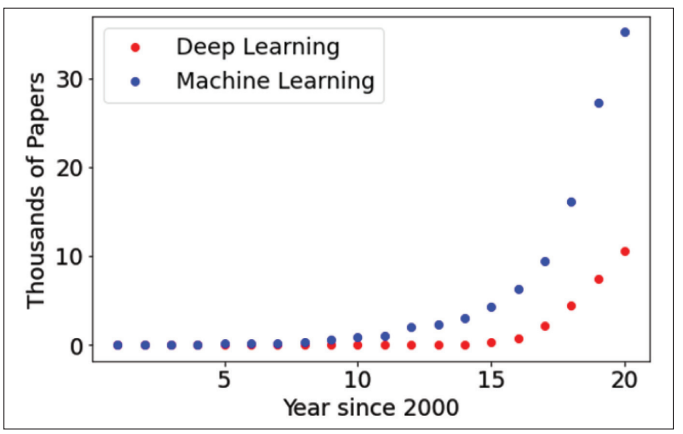

Fig. 1. Number of papers per year (in thousands) returned by Harvard's Astrophysics Data System abstract archive. Blue dots show results of a search on "machine learning" while red dots show results of a search on "deep learning." No detailed analysis was performed to filter results on actual paper content—numbers should be interpreted as indicative, not exhaustive.

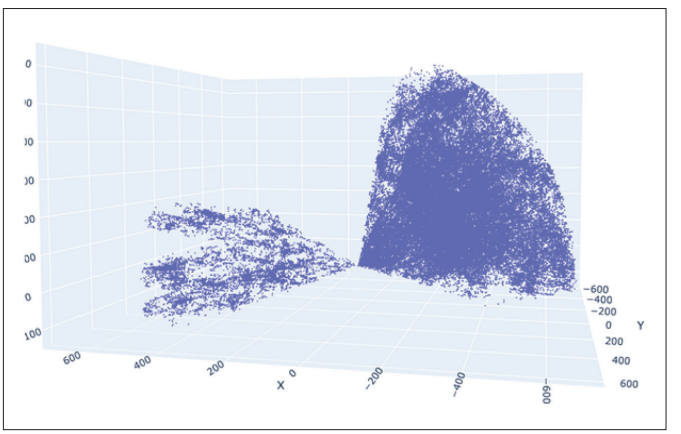

Fig. 2. Static version of plotly 3D scatter plot of 100,000 galaxies out to z = 0.05 (axis numbers in MLY) from the SDSS. Viewed live through the Jupyter Notebook, the plot can be rotated around any axis as well as zoomed in and out. Putting the mouse over any single dot will yield a pop-up box with the *xyz*-coordinates. Students can identify filaments and voids as well as understand how the survey observing strategy leads to large gaps in coverage.

 DOI: 10.1119/10.0006925

taught students to extract and plot locations of SDSS galaxies. Using Jupyter Notebooks[13] for this activity had several advantages over Excel: the plotly library allows for interactive 3D graphs that can be rotated and zoomed to give a better sense of the galaxies' locations in space than a 2D slice would. One need also not restrict oneself to a slice, but all 100,000 galaxies can be plotted in a single graph, as shown in Fig. 2. In the notebook, galaxy location numbers pop up when you put the mouse over each dot.

This activity refreshed students' memory about galaxies, enabled them to explore the properties of large scale structure, and also gave them working experience with Jupyter Notebooks, which they would need for the next two sessions, bringing in machine learning for the last two weeks of the semester. The students spent a week working through a notebook that covered the building blocks of machine learning tools for image classification: how you can use an array of millions of numbers to calculate the probability that a given input image is associated with one of a given array of output labels.

The final lab activity, based on the work of the Galaxy10 project,[14] brought the students full circle back to Galaxy Zoo, but this time the students used 18,000 SDSS images to train a program to classify galaxies into 10 morphological types. I wrote a Jupyter Notebook that allowed the students to import those images, train a network to classify them, and test their own instincts for classification (reaching back to their experience in week two) against the performance of their network.

One challenge in trying to develop pedagogy for machine learning is that the fast-moving nature of the field means that tools quickly become obsolete. At the time I needed it in the course, the Galaxy10 code as written by H. Leung & J. Bovy was no longer consistent with the most recent version of Tensorflow. I had to rewrite it myself. Worse, I finished and successfully tested the galaxy classification network two weeks before the class date, but in class, the galaxy images exceeded the available RAM of the virtual machine created by the Google Colaboratory. Google seems to, exactly during those two weeks, have chosen to restrict the available RAM offered to users. I had to very quickly rewrite the code to downsample the galaxy images. If you are going to try to use machine learning tools in the classroom, you have to be prepared for external/cloud resources to change without warning.

Although students were able to complete my activities, I would not claim they gained a robust understanding of AI concepts. Many of them balked at even trying to make sense of the quantitative output (the probabilities assigned to each possible classification label). For a very rough introduction to the topic, it was moderately successful, but to really gain an understanding of machine learning concepts, I would need more activities and more time.

Although the tools are not ready for easy distribution and implementation yet, if you have the time and the energy to help the students get through them (and if you're ready to be nimble when the command libraries change under you), it's a rewarding experience. Students found it fascinating, albeit challenging. Several spoke to me of making connections between what we were doing and stories about machine learning that they were hearing in the news. The students clearly recognized that knowing something about what machine learning tools do would be useful for their future.

Acknowledgments

The author would like to acknowledge W. Hahn for programming help and inspiration, and also to thank former Column Editor Joe Heafner and Column Co-Editor Janelle Bailey for providing useful input on this paper. This research has made use of NASA's Astrophysics Data System.

References

1. There are some fantastic introductory videos and transcripts at https://www.3blue1brown.com/topics/neural-networks, accessed Aug. 9, 2021, that explain the basic concepts and methods of machine learning.
2. The site medium.com has many articles, such as https://carterrhea93.medium.com/astronomical-images-for-machine-learning-applications-b50e7f298337, accessed May 18, 2021. Github and Kaggle also provide many, many resources. Any details I might explain here would probably be obsolete by the time this article is printed.
3. http://ads.harvard.edu/, accessed June 1, 2021.
4. J. P. Davis and W. A. Price, "Deep learning for teaching university physics to computers," *Am J. Phys.* **85**, 311 (April 2017).
5. https://keras.io/, accessed June 2, 2021.
6. https://www.tensorflow.org/, accessed June 1, 2021.
7. https://www.guilford.edu/news/2020/10/programming-future, accessed June 1, 2021.
8. C. J. Lintott et al., "Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Mon. Not. R. Astron. Soc.* **389**, 1179 (2008).
9. S. Slater, T. Slater, and D. J. Lyons, *Engaging in Astronomical Inquiry* (W.H. Freeman & Company, 2010).
10. E. Prather, T. Slater, J. Adams, J. Bailey, L. Monowar-Jones, and J. Dostal, "Research on a Lecture-Tutorial approach to teaching introductory astronomy for non-science majors," *Astron. Educ. Rev.* **3** (2), 122 (2004)
11. https://www.sdss.org/, accessed Aug. 9, 2021.
12. http://physics.scsu.edu/~dms/cosmology/Earth_Sci_LSS_Lab_05-16-2016.pdf, accessed May 17, 2021.
13. A Jupyter Notebook is a web-browser-based interactive computing environment that uses the Python language to carry out programming tasks. It also supports HTML and LaTeX markup languages to present text and image-based commentary on and explanations of the material in the notebook. The result is an interactive web page with images, graphics, text, equations, and active Python code. Jupyter Notebooks are used extensively in programming, physics, and data science research and commerce.
14. https://astronn.readthedocs.io/en/latest/galaxy10.html, accessed May 17, 2021.