Thesis Project

Edgar Ortiz

2019

1 Introduction

We are drowning in information and starving for knowledge.
-Rutherford D. Roger

Astronomy has entered a Big-Data era Moravec et al. (2019) Siemiginowska (2019) Feigelson (2012) thanks to advances in instrumentation, new telescopes, and mainly because of large astronomical surveys. The trend is, it will just keep growing on and on at an exponential rate in the coming years. Just to mention a couple of these large surveys adding to this reservoir of data, we have the SDSS Gunn et al. (2006) Weijmans and Team (2015) and the future VERY large survey LSST Ivezić et al. (2008). Citing from LSST web page https://www.lsst.org/about :

"The goal of the Large Synoptic Survey Telescope (LSST) project is to conduct a 10-year survey of the sky that will deliver a 500 petabyte set of images and data products that will address some of the most pressing questions about the structure and evolution of the universe and the objects in it."

Within all this data lies hidden scientific truths, which, by traditional human and computational means will take many years to unveil. This new challenge posed by Big-Data can be tackled using new algorithms able to look for these truths Baron and Poznanski (2017) Tshitoyan et al. (2019). Algorithms of these types exist in the realm of Machine Learning (ML) Géron (2017). ML computing techniques differ from traditional ones in the way a problem is approached. Traditionally, a Rules-Based approach Patel (2019) is used, i.e, the computer is explicitly programmed to solve a

task. On the other hand, ML algorithms are able to solve a task learning from the data itself, i.e, they learn on their own without being explicitly programmed or guided. ML algorithms obtain insight from data. A final remark, but a very important one, is that the overall performance of ML techniques improves when more data is fed to the algorithms, a feature that makes it very attractive in this new era of astronomy.

ML algorithms can learn from data with supervision or without supervision, so a natural classification for ML appears Géron (2017) Patel (2019):

- supervised learning (SL), and
- unsupervised learning (UL).

Nonetheless, the division is blurred, since mixed approaches can be considered. The only thing we can talk about is the amount of supervision provided to the algorithm.

The difference between supervised and unsupervised learning in terms of the data used to feed the ML algorithm resides in the fact that in SL, training data must be given to the ML algorithm which already contains the relations between the features of the input data (input variables) and the labels assigned to them (output variables). These relations are usually computed by a human expert. In this framework, the algorithm adjusts its parameters to learn the relations implicitly and then it generalizes when unlabeled data is presented to it.

In a SL project, before delivering the final product, it is a common practice to split the training set (the one with labels) into three subsets:

1. Training set: with this one the ML algorithm acquires experience.
2. Validation set: the experience acquired before is tested by measuring the performance of the algorithm. If the performance is not acceptably good for the task, modifications are done and we go back to the training set.
3. Test set: once we have a working model, we test it against unseen data and see how well it generalizes.

From the previous workflow Géron (2017), while training a SL program, a weakness in SL can be spotted: if we present to it with data which is very dissimilar to the one used for the training, it would not give an appropriate answer (the algorithm can not generalized well for this data). This weakness propagates to another one: the amount of labeled data in the real world is small when compared with the amount of generated data (unlabeled), then an SL algorithm has severe limitations in broader contexts and adapting to incoming informationPatel (2019).

Regarding UL, here the algorithms do not have any guide during the learning process. UL algorithms have to find by themselves the patterns within the data. And here lies one of its strengths, it is able to map each feature in the data set to a generic label, then if more and more data is given, the pitfalls we had on SL are diminished Patel (2019).

Summarizing, the learning done in the supervised case is conditioned to the relation between the labels given to the input data in the training set, while in the case of unsupervised learning, the algorithm generates by itself the underlying relations within the data.

An obvious advantage of SL over UL is that for SL a performance measure can be established and optimized, and as a consequence, the algorithm can generalize to an acceptable degree according to the requirements of our task; in UL, that is impossible. This difference makes the problems to be tackled by each algorithm of a different nature. As an example, and going back to astronomy, one particular question we could ask to an astronomical data-set could be: What are the weirdest objects you have there? And by weird, we mean objects which are either extreme objects in our lore or unknown objects (not described by our current accepted knowledge). Well, it turns out to be a perfect task for UL Baron and Poznanski (2017), since it is impossible to provide a training data set containing unknown labeled objects for SL.

The end goal of these tools is knowledge discovery from the data, a term that is used in the fields of data mining and ML Ivezic et al. (2017). To accomplish this, we need to somehow make a working model for the data that can allow us to make predictions and better understand the problem at hand: either quantitatively or qualitatively. It

all sounds amazing, nonetheless, it is always good to have in mind that a model is just a representation of reality and that reality is more complex than that; that being said, when using a model to understand our data, we have to be careful about possible biases that might be introduced in our interpretation about the phenomenology in our problem. As an example, we can consider two very different approaches when modeling some data:

1. A linear model: one-parameters model
2. N-degree polynomial model: n-parameter model,

In the first case, assuming a linear behavior is a very strong assumption about the phenomenology in our problem which could lead to a very poor interpretation of it if the problem has a higher degree of complexity. On the other hand, in the second case, an n-degree polynomial could overfit the data, leading to a poor predictive power of our data, ie, our model doesn't generalize pretty well for points outside our training data. As we can see, there is a compromise between the complexity used in the model chosen to study the data. Finally, and also very relevant when trying to understand what is going on in our data is the very nature of it, this factor also has an impact on what model should be used to interpret it. An interesting discussion about this is found in Hastie et al. (2003) where this trade-off is discussed comparing a linear model against a KNN model for a classification problem.

2 Machine Learning in Astronomy

The traditional way to extract knowledge from data consists of using a predefined model to fit in it the data. The nature of the model depends on what information we want to extract from the data we have collected, as an example, we can consider a mono-atomic diluted gas confined in a recipient of volume V, at a temperature T and at a pressure P, for which we want to know how these features are correlated if they are modified. To get the data we have to measure pressure, volume, and temperature for different configurations of the system. Then we search for a toy-theoretical model based on sound physical laws to predict what would be the functional relation among all the variables. Once we have these predefined models which will depend on some parameters, we proceed to feed the data to it to see if it predicts correctly the correlations among the measured features, if not, we search for another model. If the prediction is quite accurate, then we can extend the model to consider more variables, like the number of particles in the recipient. This approach is just model fitting using some statistical considerations to handle the accuracy of the model.

The case stated before is quite simple since it is a system that can be controlled to explore the correlation among the variables, and additionally it is not a complex system since we have only three variables involved. In astronomy the story is completely different:

1. Experiments can not be performed on objects like stars or distant galaxies.

2. The number of measured variables is huge (multi-waveband), making its correlations highly complex and then tough to make a model describing them, see Delli Veneri et al. (2019).

In this case, the model-fitting paradigm is not the best approach to follow. Here appears the need for statistical tools that can learn those correlations on their own, helping us to gain more insight into the objects we study Baron (2019). These tools are machine learning algorithms that accomplish this task in two ways mainly: supervised and unsupervised learning. In what follows, supervised and unsupervised learning techniques are discussed, mentioning their current use in astronomy and their future perspectives.

2.1 Supervised learning (SL)

With the advent of computers and the digitalization of data, even the model-fitting paradigm can be done really quickly by programming a computer. The problem with this approach is that in it, a set of rules has to be programmed, according to the model considered in order to accomplish the task. Nonetheless, if the complexity of the correlation among the features of the data increases, as is the case with the advent of Big-Data in astronomy, the set of rules will be obsolete and a new set of rules will have to be written, however, and as mentioned in the introduction of this section, this requires a model which is pretty impossible to came up with if the complexity of the problem is huge. In order to avoid that, we have to program and train a ML algorithm. In a broad sense and citing Arthur Samuel (1959) Géron (2017), Machine Learning can be defined as "the field of study that gives computers the ability to learn without being explicitly programmed". SL is the case when the machine learns using domain knowledge from a human expert.

Regarding astronomy, we can picture the case where we have spectra from millions of galaxies and where our goal is to obtain, let's say: their star formation rates Delli Veneri et al. (2019), their metallicity Acquaviva (2016) Ucci et al. (2017), density, column density, and ionization parameter Ucci et al. (2017) using SL. In plain English, to make the machine learn the relation among the spectra and the quantities of interest, a data set with those relations (already computed by a human expert), is compulsory (that's the supervision). That knowledge is then fed to the algorithm so it learns the relations (in a black-box approach), tweaking some parameters according to the type of ML model used (hyper-parameters) and the nature of the studied data (model parameters) Baron (2019). After the relations are learned, we use the ML as a tool to make predictions of the target variables on new galaxies.

In the language of statistics, we can define the task of supervised ML with the following ingredients Hastie et al. (2003):

1. $\vec{X}$: a real-valued random input vector.
2. $\vec{Y}$: a real-valued random output vector.
3. $Pr(\vec{X}, \vec{Y})$: the joint distribution between $\vec{X}$ and $\vec{Y}$.

4. $\mathcal{L}(Y, f(\vec{X}))$: the loss function for penalizing errors in the predictions made by the machine

The ultimate goal is to find the optimal solution for $Y = f(X)$ such that our loss function is minimized. That's the mathematical formulation of the problem, but since reality is trickier, what one truly does is to attempt to achieve this, but as an approximation with the available data.

The loss function $\mathcal{L}(Y, f(X))$ is metric a that measures how much the prediction $f(\vec{X})$ deviates from the target variables $\vec{Y}$. Several metrics can be chosen, like the Mean Square Error (MSE) or the Mean Absolute Error (MAE), just to mention a couple of them:

$$
\begin{aligned}
MSE &= \frac{1}{N}\sum_{i=1}^{N}\left(\vec{Y} - f(\vec{X})\right)^2 \\
MAE &= \frac{1}{N}\sum_{i=1}^{N}\left|\vec{Y} - f(\vec{X})\right|
\end{aligned}
$$

Depending on the error metric chosen, we will get a solution to our optimization problem Hastie et al. (2003). The error metric has to be chosen according to the needs of our science case, for example, the MSE metric is better suited to consider the information contained in outliers, while the MAE is not. The loss function is optimized according to the hyper-parameters of the ML model so we obtain the better performing ML model Baron (2019)

References

Acquaviva, V. (2016), 'How to measure metallicity from five-band photometry with supervised machine learning algorithms', *Monthly Notices of the Royal Astronomical Society* **456**(2), 1618–1626.

Baron, D. (2019), 'Machine Learning in Astronomy: a practical overview'.
URL: *http://arxiv.org/abs/1904.07248*

Baron, D. and Poznanski, D. (2017), 'The weirdest SDSS galaxies: Results from an outlier detection algorithm', *Monthly Notices of the Royal Astronomical Society* **465**(4), 4530–4555.

Delli Veneri, M., Cavuoti, S., Brescia, M., Longo, G. and Riccio, G. (2019), 'Star formation rates for photometric samples of galaxies using machine learning methods', *Monthly Notices of the Royal Astronomical Society* **486**(1), 1377–1391.

Feigelson, E. (2012), 'Big Data in Astronomy', *International Astronomical Union Colloquium* **64**, 217–225.

Géron, A. (2017), *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, O'Reilly Media.

Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., Carey, L. N., Knapp, G. R., York, D. G., Boroski, W. N., Kent, S. M., Lupton, R. H., Rockosi, C. M., Evans, M. L., Waddell, P., Anderson, J. E., Annis, J., Barentine, J. C., Bartoszek, L. M., Bastian, S., Bracker, S. B., Brewington, H. J., Briegel, C. I., Brinkmann, J., Brown, Y. J., Carr, M. A., Czarapata, P. C., Drennan, C. C., Dombeck, T., Federwitz, G. R., Gillespie, B. A., Gonzales, C., Hansen, S. U., Harvanek, M., Hayes, J., Jordan, W., Kinney, E., Klaene, M., Kleinman, S. J., Kron, R. G., Kresinski, J., Lee, G., Limmongkol, S., Lindenmeyer, C. W., Long, D. C., Loomis, C. L., Mcgehee, P. M., Mantsch, P. M., Neilsen, E. H., Neswold, R. M., Newman, P. R., Nitta, A., Peoples, J., Pier, J. R., Prieto, P. S., Prosapio, A., Rivetta, C., Schneider, D. P., Snedden, S. and Wang, S.-i. (2006), 'THE 2 . 5 m TELESCOPE OF THE SLOAN DIGITAL SKY SURVEY', *The Astronomical Journal* (2001), 2332–2359.

Hastie, T., Tibshirani, R. and Jerome, F. (2003), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edn, Springer.

Ivezic, Ž., Connolly, A. J., VanderPlas, J. T. and Gray, A. (2017), *Statistics, Data Mining, and Machine Learning in Astronomy*.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., Axelrod, T. S., Bard, D. J., Barr, J. D., Barrau, A., Bartlett, J. G., Bauer, A. E., Bauman, B. J., Baumont, S., Becker, A. C., Becla, J., Beldica, C., Bellavia, S., Bianco, F. B., Biswas, R., Blanc, G., Blazek, J., Blandford, R. D., Bloom, J. S., Bogart, J., Bond, T. W., Borgland, A. W., Borne, K., Bosch, J. F., Boutigny, D., Brackett, C. A., Bradshaw, A., Brandt, W. N., Brown, M. E., Bullock, J. S., Burchat, P., Burke, D. L., Cagnoli, G., Calabrese, D., Callahan, S., Callen, A. L., Chandrasekharan, S., Charles-Emerson, G., Chesley, S., Cheu, E. C., Chiang, H.-F., Chiang, J., Chirino, C., Chow, D., Ciardi, D. R., Claver, C. F., Cohen-Tanugi, J., Cockrum, J. J., Coles, R., Connolly, A. J., Cook, K. H., Cooray, A., Covey, K. R., Cribbs, C., Cui, W., Cutri, R., Daly, P. N., Daniel, S. F., Daruich, F., Daubard, G., Daues, G., Dawson, W., Delgado, F., Dellapenna, A., de Peyster, R., de Val-Borro, M., Digel, S. W., Doherty, P., Dubois, R., Dubois-Felsmann, G. P., Durech, J., Economou, F., Eracleous, M., Ferguson, H., Figueroa, E., Fisher-Levine, M., Focke, W., Foss, M. D., Frank, J., Freemon, M. D., Gangler, E., Gawiser, E., Geary, J. C., Gee, P., Geha, M., Gessner, C. J. B., Gibson, R. R., Gilmore, D. K., Glanzman, T., Glick, W., Goldina, T., Goldstein, D. A., Goodenow, I., Graham, M. L., Gressler, W. J., Gris, P., Guy, L. P., Guyonnet, A., Haller, G., Harris, R., Hascall, P. A., Haupt, J., Hernandez, F., Herrmann, S., Hileman, E., Hoblitt, J., Hodgson, J. A., Hogan, C., Huang, D., Huffer, M. E., Ingraham, P., Innes, W. R., Jacoby, S. H., Jain, B., Jammes, F., Jee, J., Jenness, T., Jernigan, G., Jevremović, D., Johns, K., Johnson, A. S., Johnson, M. W. G., Jones, R. L., Juramy-Gilles, C., Jurić, M., Kalirai, J. S., Kallivayalil, N. J., Kalmbach, B., Kantor, J. P., Karst, P., Kasliwal, M. M., Kelly, H., Kessler, R., Kinnison, V., Kirkby, D., Knox, L., Kotov, I. V., Krabbendam, V. L., Krughoff, K. S., Kubánek, P., Kuczewski, J., Kulkarni, S., Ku, J., Kurita, N. R., Lage, C. S., Lambert, R., Lange, T., Langton, J. B., Guillou, L. L., Levine, D., Liang, M., Lim, K.-T., Lintott, C. J., Long, K. E., Lopez, M., Lotz,

P. J., Lupton, R. H., Lust, N. B., MacArthur, L. A., Mahabal, A., Mandelbaum, R., Marsh, D. S., Marshall, P. J., Marshall, S., May, M., McKercher, R., McQueen, M., Meyers, J., Migliore, M., Miller, M., Mills, D. J., Miraval, C., Moeyens, J., Monet, D. G., Moniez, M., Monkewitz, S., Montgomery, C., Mueller, F., Muller, G. P., Arancibia, F. M., Neill, D. R., Newbry, S. P., Nief, J.-Y., Nomerotski, A., Nordby, M., O'Connor, P., Oliver, J., Olivier, S. S., Olsen, K., O'Mullane, W., Ortiz, S., Osier, S., Owen, R. E., Pain, R., Palecek, P. E., Parejko, J. K., Parsons, J. B., Pease, N. M., Peterson, J. M., Peterson, J. R., Petravick, D. L., Petrick, M. E. L., Petry, C. E., Pierfederici, F., Pietrowicz, S., Pike, R., Pinto, P. A., Plante, R., Plate, S., Price, P. A., Prouza, M., Radeka, V., Rajagopal, J., Rasmussen, A. P., Regnault, N., Reil, K. A., Reiss, D. J., Reuter, M. A., Ridgway, S. T., Riot, V. J., Ritz, S., Robinson, S., Roby, W., Roodman, A., Rosing, W., Roucelle, C., Rumore, M. R., Russo, S., Saha, A., Sassolas, B., Schalk, T. L., Schellart, P., Schindler, R. H., Schmidt, S., Schneider, D. P., Schneider, M. D., Schoening, W., Schumacher, G., Schwamb, M. E., Sebag, J., Selvy, B., Sembroski, G. H., Seppala, L. G., Serio, A., Serrano, E., Shaw, R. A., Shipsey, I., Sick, J., Silvestri, N., Slater, C. T., Smith, J. A., Smith, R. C., Sobhani, S., Soldahl, C., Storrie-Lombardi, L., Stover, E., Strauss, M. A., Street, R. A., Stubbs, C. W., Sullivan, I. S., Sweeney, D., Swinbank, J. D., Szalay, A., Takacs, P., Tether, S. A., Thaler, J. J., Thayer, J. G., Thomas, S., Thukral, V., Tice, J., Trilling, D. E., Turri, M., Van Berg, R., Berk, D. V., Vetter, K., Virieux, F., Vucina, T., Wahl, W., Walkowicz, L., Walsh, B., Walter, C. W., Wang, D. L., Wang, S.-Y., Warner, M., Wiecha, O., Willman, B., Winters, S. E., Wittman, D., Wolff, S. C., Wood-Vasey, W. M., Wu, X., Xin, B., Yoachim, P., Zhan, H. and Collaboration, f. t. L. (2008), 'LSST: from Science Drivers to Reference Design and Anticipated Data Products', *The Astronomical Journal* **111**.
URL: *http://arxiv.org/abs/0805.2366%0Ahttp://dx.doi.org/10.3847/1538-4357/ab042c*

Moravec, E., Czekala, I. and Follette, K. (2019), 'Astro2020 APC White Paper: The Early Career Perspective on the Coming Decade, Astrophysics Career Paths, and the Decadal Survey Process', pp. 1–9.
URL: *http://arxiv.org/abs/1907.01676*

Patel, A. A. (2019), *Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data*, O'Reilly Media.

Siemiginowska, A. (2019), 'Astro2020 Science White Paper The Next Decade of Astroinformatics and Astrostatistics'.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G. and Jain, A. (2019), 'Unsupervised word embeddings capture latent knowledge from materials science literature', *Nature* **571**(7763), 95–98.
URL: *http://www.nature.com/articles/s41586-019-1335-8*

Ucci, G., Ferrara, A., Gallerani, S. and Pallottini, A. (2017), 'Inferring physical properties of galaxies from their emission-line spectra', *Monthly Notices of the Royal Astronomical Society* **465**(1), 1144–1156.

Weijmans, A.-m. and Team, o. b. o. t. M. (2015), 'MaNGA: Mapping Nearby Galaxies at Apache Point Observatory', pp. 1–7.
URL: *http://arxiv.org/abs/1508.04314*