Using Machine Learning to find galaxies with the [O_{III}] λ4363 line in the Sloan Digital Sky Survey (SDSS)

Pablo Lechón

September 10, 2019

Abstract

This manual explains the first version of a machine learning algorithm that determines whether a galaxy presents the [O_{III}] λ4363 line or not based on some of its photometric properties. We use two data sets for this purpose; the training set, made of galaxies with fitted line emissions, and the data set of photometric features of galaxies we want to predict. After a feature selection and training the algorithm we have found approximately 100 galaxies with the [O_{III}] λ4363 line that were not in the training set. To increase this number, expected to be 10 to 20 times higher, future work will include cleaning the data set, performing a better feature selection by improving the features we already have with PCA analysis and including new spectroscopic features into our analysis.

1 Motivation

Calculating abundances in galaxies is key to understand their evolution. One way to estimate galaxy abundances is by looking at its emission lines. Particularly, the [O_{III}] λ4363 line, also known as the *Golden Standard for Metallicity*, is the most useful line to calculate abundances. Therefore it would be ideal to have a method to systematically find these galaxies within a given survey of galaxies. Finding such a method is not easy because the auroral line [O_{III}] λ4363 is extremely faint, so it shows up in a very small fraction of spectra.

This manual comprises the first approach to create a recipe that finds galaxies with the [O_{III}] λ4363 line. Due to the size of our dataset, we have limited our analysis merely to photometric properties, leaving out any spectroscopic information. This avoids having to fit the spectra of SDSS DR15 survey, which add up to more than $2 \cdot 10^{6}$ spectra. Moreover, the [O_{III}] λ4363 line is so faint that we have to perform a very accurate representation of the underlying stellar population to be able to subtract it from the spectra and uncover the [O_{III}] λ4363 line. This makes fitting spectra a very computationally expensive task.

2 Data

Our data comes from the SDSS 15th Data Release (DR15) [1]. We have analyzed three datasets. The first one is a set of nearly $8 \cdot 10^5$ spectra fits which details can be found in the set of papers [2–4]. We have used the signal-to-noise ratio (S/N) for the [O_{III}] λ4363 line.

The second one is a set of $2 \cdot 10^5$ fits per-

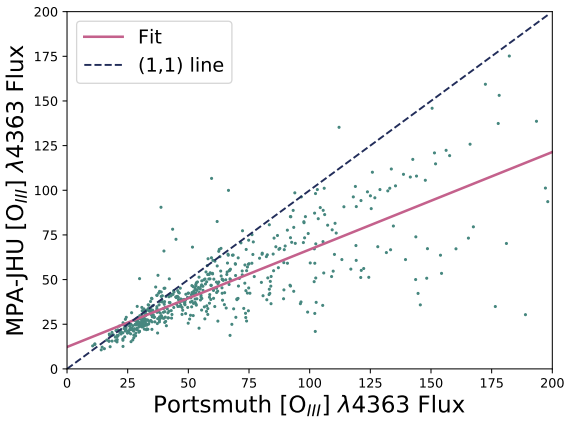

Figure 1: Comparison between the [O_{III}] λ4363 line flux from the Portsmuth group and Max Plank for Astrophysics - Johns Hopkins University (MPA-JHU).

formed by [5], from which we have drowned the same information as for the first data set. We have assumed that both datasets yield equivalent results because they are strongly correlated, as we can see in figure 1. It can be seen that there is a slight misalignment between what we would expect and the fit. This may be caused by a slightly different subtraction of the subjacent stellar population by both groups. This is not a big problem, because we are only worried about galaxies with high S/N in the [O_{III}] λ4363 line, not about the specific numerical value. Therefore, the only condition that must satisfy in order for us to consider both data sets as equivalent is that they are correlated.

The third dataset is a combination of several tables contained in the SDSS Sky Server data base, namely, GalSpecLine, SpecObj, Galaxy, and SpecPhoto. To create the three datasets we use a SQL query that is sent to the data base through CasJobs. These queries can be found at section 6.

3 Machine Learning

The concept of machine learning, first introduced in [6], refers to the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions. There are several types of machine learning algorithms, and using one or another depends on their approach, the type of data they input and output, and the type of task or problem that they are intended to solve. The two main types of machine learning algorithms are supervised learning and unsupervised learning. The former being the type of algorithm in which we know the output possibilities, and the latter being the one in which we don't know neither what nor how many categories the output has. Unsupervised learning algorithms are also known as clustering algorithms.

In our case of study, we wish to predict whether a given galaxy will present the [O_{III}] λ4363 line or not. Thus, this is a supervised learning problem, because we know the output categories. It is also discrete, since there is a finite number of output categories (only 2: yes or no).

There are several machine learning algorithms that deal with these kind of problems. In this manual we use the random forest algorithm [7] from Python's ScikitLearn [8], which is an ensemble method involving numerous tree predictors.

Tree Predictors, or Decision Trees (DTs) are supervised learning methods used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Tree predictors have a tree-like structure (see figure 2). A decision tree is drawn upside

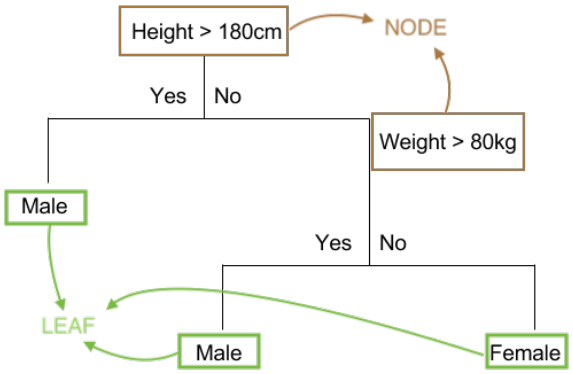

Figure 2: DT structure

down with its root at the top. The condition or nodes, are the way by which the tree splits into branches. The end of the branch that doesn't split anymore is the decision or leaf. DTs's main input arguments are: 1. `max_depth`: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure [1] or until they contain less than `min_samples_split` samples; 2. `min_samples_split`: The minimum number of samples required to split an internal node; 3. `min_samples_leaf`: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least `min_samples_leaf` training samples in each of the left and right branches. it is clear that we run into risk of overfitting our data if `max_depth` is high enough or `min_samples_split` is sufficiently low.
If we combine the predictions of several tree estimators, we have an Ensemble Method. This is convenient because it allows for an improvement in robustness and generalizability compared to a single estimator. Random forest, the method we use in our analysis, is an ensemble method formed by combining several DTs. Each tree in the ensemble is built from a sample drawn (with replacement) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

[1] A leaf is pure if it comes from a 100 % pure node. A node is 100% pure when all of its data belongs to a single class and 100% impure when a node is split evenly 50/50 between the two classes.

4 Methodology and Predictions

Before applying our algorithm we need to create a training set and perform a feature selection. The training set is the data that we will fit our classifier to. This means we need to know the classification/labels for that dataset; in other words, we need to know whether the galaxies in the training set present the [O$_{III}$] λ4363 line or not. In order to get this information, we use our first and second data sets mentioned in section 2. Through a CasJobs SQL query we select all the galaxies from both data sets with a S/N $\geq$ 5 in the [O$_{III}$] λ4363 line flux, which is obtained from the fits to the spectra. However, since the [O$_{III}$] λ4363 line is very faint, we expect that a significant fraction of galaxies with S/N $\geq$ 5 do not actually present such a strong line because it has been misfitted. Therefore, we apply a second filter: spectra with S/N $\geq$

5 is visually analyzed to see if the fit has done its job correctly or not. This tedious but rewarding analysis brings valuable answers: We discovered three types of spectra that had S/N $\geq$ 5: a) the ones which actually present the line; b) those that presented it but in an odd proportion due to an AGN[2]; c) misfits. We ruled out of our analysis groups b) and c).

This two step analysis provides the algorithm with a pretty robust set of labels for those galaxies with S/N $\geq$ 5. The assumption that for those galaxies with S/N $\leq$ 5 there is no visible [O_{III}] λ4363 line, can be made without a significant loss in accuracy. To justify this assumption, the visual inspection of the spectra revealed that for 5 $\leq$ S/N $\leq$ 6 there were almost no galaxies with a visible [O_{III}] λ4363 line.

After getting the labels, we have to perform a feature selection. This selection tries to answer the question whether a certain feature is or is not discriminatory towards determining the presence of the [O_{III}] λ4363 line. The initial future selection was inspired by [9]. We also plotted a color magnitude diagram for the galaxies in the training set. Histograming the Petrosian radius revealed that it was also a discriminatory measure. After preparing the training data set and selecting features, we proceed to train the algorithm.

Fitting data to a model always entails the risk of overfitting it. To avoid this, we subtract from the training set a fraction (30 % of the initial training set in our case) of data that is not fed to the model: the test data set. As we make our model more complex by increasing `max_depth` or decreasing `min_samples_split`, the accuracy of the predictions on the training set will always increase. However, the accuracy on the test set will increase with the model complexity until a certain point, where it will start to decay. This is the optimal point: any lower complexity is underfitting our data, whereas a higher one implies overfitting it. This process can be visualized in figure 5. The accuracy increases on the training set until `max_depth` = 8, where it doesn't increase anymore.

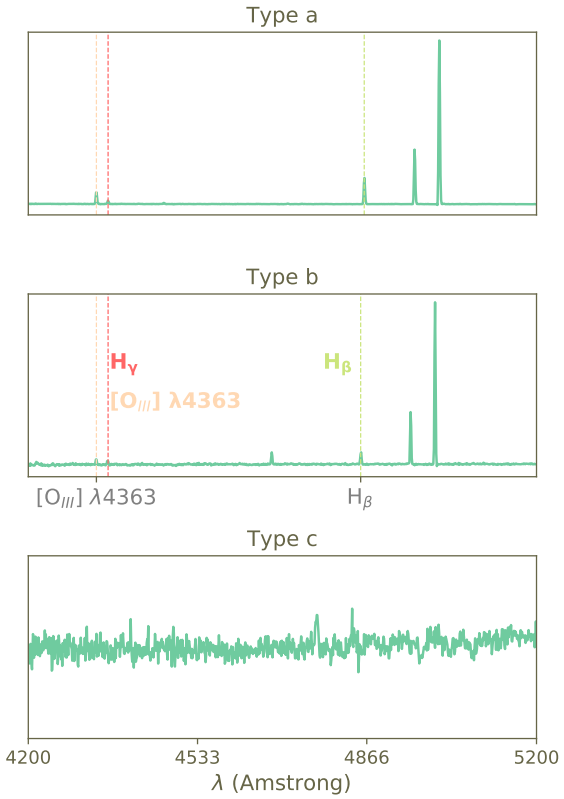

Figure 3: Types of spectra found in the visual analysis.

[2]The difference between spectra types a) and b), as we se in figure 3, is that the ration between lines is different. Both H_γ and H_β are several orders of magnitude higher that [O_{III}] λ4363 in type a) compared to type b). When plotting type b) spectra in a diagnostic diagram, we can see that they are in the AGN region. Nucleus of AGN galaxies (Black Holes) is capable of strongly ionize the oxygen. That is why the [O_{III}] λ4363 line appears to be much stronger.

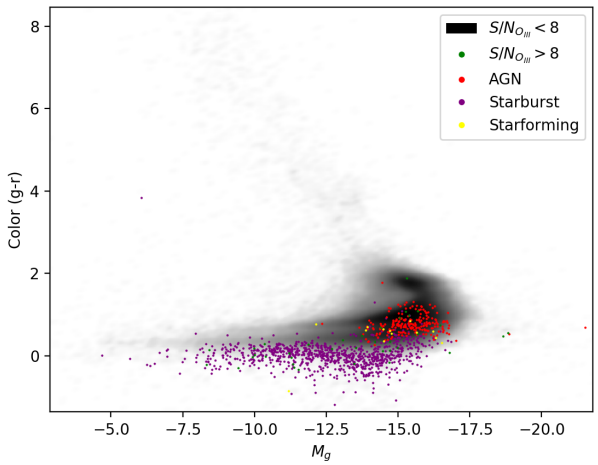

Figure 4: Color-Magnitude diagram for data from Portsmuth and MPA-JHU group

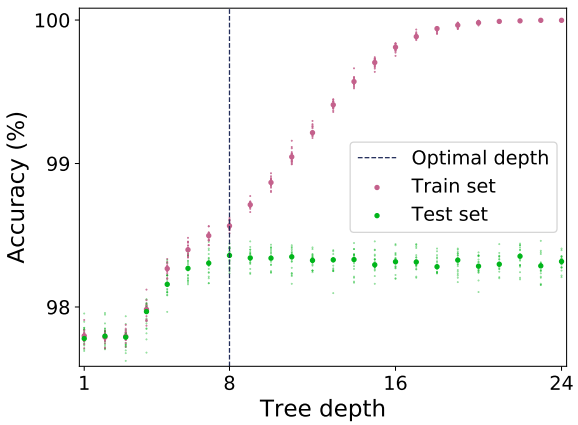

Figure 5: Accuracy of the training and test data sets as `max_depth` increases. The uncertainties were obtained via Monte Carlo simulations. Note that they are higher in the test set, because it is a smaller set. Also, the standard deviation decreases in the training set when the tree depth increases, this is due to overfitting.

With our trained model we are ready to predict outputs for the $2 \cdot 10^6$ galaxies with measured spectra. The output of the algorithm were 100 positive flags. Upon visual examination, a majority of them (93 galaxies) had a very strong [O_{III}] λ4363 line. We believe the number of galaxies with a positive flag can be significantly improved (by a factor of 10, or 20) by making some corrections in the data preparation.
The limitation of predicting only galaxies that have a measured spectra is temporary. Since the algorithm is not perfect yet, we want to make sure that the positive outputs of our algorithm are indeed galaxies with the [O_{III}] λ4363 line. For this purpose, we need to have a reference spectrum which to compare our results. Once the algorithm predictions are consistently right, we would trust it enough to start predicting the presence of the [O_{III}] λ4363 line in galaxies with no measured spectra. This procedure would provide us with interesting SDSS targets to point at.

5 Future Work

Several things can be done in order to find more galaxies with the [O_{III}] λ4363 line. First, we need to clean the training set and explore different variables than the ones we have. To clean the data set, we want to get rid of spectra without emission lines (type c in figure 3). Some of these spectra are flagged positively by the algorithm, which means that the photometric features we are considering are not sufficient to find only galaxies with the [O_{III}] λ4363 line. To solve this problem we can do two things; include more features (maybe consider extending our analysis to the spectroscopic domain too) and improve the features we have by carrying out a PCA analysis on our features [10]. This would create a similar set of variables to the ones we had before that discriminate more efficiently between

galaxies with and without the line. If we decided to extend analysis to the spectroscopic domain, we can still avoid fitting a all the spectrum of each galaxy while get very useful information. The intensity of the H_β line is strongly correlated to the presence of the $[O_{III}]$ $\lambda 4363$ line, so fitting only this line, and using it as one of the machine algorithm features would significantly improve our results. Finally, trying other classification algorithms might yield better conclusions. In [11] there is an in depth comparison.

6 Appendix

List of queries to CasJobs for datasets groups 1, 2, and 3 respectively mentioned in section 2.

```
SELECT s.plate, s.fiberid, s.mjd, ga.oiii_4363_flux,
ga.oiii_4363_flux_err, (sph.dered_g-sph.dered_r) as col_g_r,
sph.dered_g-5*log10(4.28E+08*sph.z)+5 as M_g, sph.petroMag_g,
g.petroRad_g

FROM GalSpecLine as ga, specobj as s, Galaxy as g,
SpecPhoto as sph

WHERE ga.oiii_4363_flux <5*ga.oiii_4363_flux_err
AND s.class = 'GALAXY'
AND s.specobjid = ga.specobjid
AND s.specobjid = g.specobjid
AND sph.specobjid = g.specobjid
AND sph.specobjid = ga.specobjid
AND ga.oiii_4363_flux_err > 0
AND g.specobjid > 0
AND sph.z > 0
into MyDB.t_5_t
```

```
SELECT s.plate, s.fiberid, s.mjd, ga.Flux_OIII_4363,
ga.Flux_OIII_4363_err, (sph.dered_g-sph.dered_r) as col_g_r,
sph.dered_g-5*log10(4.28E+08*sph.z)+5 as M_g, sph.petroMag_g,
g.petroRad_g

FROM emissionlinesport as ga, specobj as s, Galaxy as g,
SpecPhoto as sph

WHERE ga.Flux_OIII_4363 <5*ga.Flux_OIII_4363_Err
AND s.class = 'GALAXY'
AND s.specobjid = ga.specobjid
AND s.specobjid = g.specobjid
AND sph.specobjid = g.specobjid
AND sph.specobjid = ga.specobjid
AND ga.Flux_OIII_4363_err > 0
AND g.specobjid > 0
AND sph.z > 0
into MyDB.t_5_p
```

```
SELECT s.plate, s.fiberid, s.mjd,
```

```
(sph.dered_g-sph.dered_r) as col_g_r,
sph.dered_g-5*log10(4.28E+08*sph.z)+5 as M_g, sph.petroMag_g,
g.petroRad_g


FROM  specobj as s, Galaxy as g, SpecPhoto as sph

WHERE s.class = 'GALAXY'
AND s.specobjid = g.specobjid
AND sph.specobjid = g.specobjid
AND g.specobjid > 0
AND sph.z > 0
into MyDB.tsdss
```

References

[1] D. S. Aguado, Romina Ahumada, Andrés Almeida, Scott F. Anderson, and Brett H. Andrews. The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library. *The Astrophysical Journal Supplement Series*, 2019.

[2] J. Brinchmann, S. Charlot, S. D. M. White, C. Tremonti, G. Kauffmann, T. Heckman, and J. Brinkmann. The physical properties of star-forming galaxies in the low-redshift Universe. *Monthly Notices of the Royal Astronomical Society*, 351(4):1151–1179, jul 2004.

[3] Christy A. Tremonti, Timothy M. Heckman, Guinevere Kauffmann, Jarle Brinchmann, Stephane Charlot, Simon D. M. White, Mark Seibert, Eric W. Peng, David J. Schlegel, Alan Uomoto, Masataka Fukugita, and Jon Brinkmann. The Origin of the Mass-Metallicity Relation: Insights from 53,000 Star-forming Galaxies in the Sloan Digital Sky Survey. *The Astrophysical Journal*, 613(2):898–913, oct 2004.

[4] Guinevere Kauffmann, Timothy M. Heckman, Simon D M White, Stéphane Charlot, Christy Tremonti, Jarle Brinchmann, Gustavo Bruzual, Eric W. Peng, Mark Seibert, Mariangela Bernardi, Michael Blanton, Jon Brinkmann, Francisco Castander, Istvan Csábai, Masataka Fukugita, Zeljko Ivezic, Jeffrey A. Munn, Robert C. Nichol, Nikhil Padmanabhan, Aniruddha R. Thakar, David H. Weinberg, and Donald York. Stellar masses and star formation histories for 105 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 2003.

[5] D. Thomas, O. Steele, C. Maraston, J. Johansson, A. Beifiori, J. Pforr, G. Strömbäck, C. A. Tremonti, and D. Wake. Stellar velocity dispersions and emission line properties of SDSS-III/BOSS galaxies. *Proceedings of the International Astronomical Union*, 8(S295):129–132, aug 2012.

[6] John R. Koza, Forrest H. Bennett, David Andre, Martin A. Keane, and Frank Dunlap. Automated synthesis of analog electrical circuits by means of genetic programming. *IEEE Transactions on Evolutionary Computation*, 1997.

[7] Leo Breiman. Random Forest. *Machine Learning*, 2001.

[8] Fabian Pedregosa, Normalesuporg Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Peter Prettenhofer, Jake Vanderplas, Matthieu Brucher, Matthieu Perrot an Edouard Duchesnay, Al Matthieu Brucher, Matthieu Perrot, and Cea F Edouard Duchesnay. Scikit-learn: Machine Learning in Python Gaël Varoquaux. *Journal of Machine Learning Research*, 2011.

[9] C. Hoyos and A. I. Díaz. The impact of the visibility of the [O III]λ4363 line on the general properties of H II galaxies in the Local Universe. *Monthly Notices of the Royal Astronomical Society*, 2006.

[10] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901.

[11] Joris Van den Bossche, Loïc Estève, Thomas J Fan, Alexandre Gramfort, Olivier Grisel, Yaroslav Halchenko, Nicolas Hug, Adrin Jalali, Guillaume Lemaitre, Jan Hendrik Metzen, Andreas Mueller, Vlad Niculae, Joel Nothman, Hanmin Qin, Bertrand Thirion, Tom Dupré la Tour, Nelle Varoquaux, Gael Varoquaux, and Roman Yurchak. Scikit-Learn. Machine Learning in Python.