TWO APPLICATIONS OF MACHINE LEARNING TO DATA FROM GALAXY SURVEYS

Viviana Acquaviva (CUNY City Tech, NYC)
Astrocosmology with galaxy surveys

Galaxies are awesome!
1. they have been around for 90% of the life of the Universe;
2. they probe cosmological expansion history $H(z)$ & structure growth $D(k,z)$
3. The evolution of galaxy properties through cosmic time teaches us about galaxy formation and evolution
4. Insert your favorite reason
Dark Energy with HETDEX

The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) is a blind spectroscopic survey that will target 900,000 Lyman Alpha Emitters at $1.9 < z < 3.5$ in a 420 square degree survey area.

Data: 2015-2018

GOALS

• Study the evolution of large-scale structure
• Determine $D_A$ and $H(z)$ at $z = 2.4$
• Measure of the curvature of the universe
• Constrain the dark energy equation of state and its evolution

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Problem: contamination from low-redshift emission line galaxies may hinder the cosmological signal.

The Ly-α line is at rest frame $\lambda = 1216$ Å.

Main contaminants are OII emitters (rest frame $\lambda = 3727$ Å) at redshift < 0.5.

Discrimination on the basis of line flux only is difficult.

Imaging survey in g band is ongoing, but further optimization is required.

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Cosmological requirement: contamination < 1%

What we know:
Emission line flux and continuum flux, hence equivalent width EW
+ EL wavelength

“Naïve” classification (shown above) assumed EW = 20 as discriminant, but there is obviously room for improvement

Goal: maximize amount of data kept (i.e., minimize incompleteness) while keeping contamination lower than 1% mark

LAEs on average have higher EW because:
1. Ly-\(\alpha\) line is stronger than [OII]
2. They are farther away and their continuum is dimmer

Image courtesy of Greg Zeimann (PSU)

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
A two-class classification problem

INPUT

MACHINE LEARNING

Bayesian Method
(Leung et al 2015 in prep)

OUTPUT

LAE

OII

Labels Available
For about 10% of data set

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
ML: Support Vector Machines (SVM)

Idea is to transform the data to higher dimensional spaces (so they are more separated), and find the hyperplane that maximizes the separation between classes, called decision boundary.

HYPERPARAMETERS:

**Kernel** (functional form of function describing the boundary)

**Gamma** (shape factor of Gaussian boundary)

**C** = Penalty function for misclassifications

**Class weight** i.e. different penalty for each class, useful for unbalanced data or skewed estimators (our case)

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Scaling, Feature Selection, Training and Optimization

Learning Set (10% of total) (Features scaled by subtracting median, dividing by MAD)

80% = Training Set

Used to find optimal hyperparameters Gamma, C, Class Weight via k-fold cross validation and to train algorithm

20% = Test Set.

Used to evaluate performance on “new” data (never in training)

Sample Variance: Random simulations of splits for learning, training, and test

Performance Check And Error Estimation

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
**Scaling, Feature Selection, Training and Optimization**

- **LEARNING SET (10% of total)**
  - (Features scaled by subtracting median, dividing by MAD)

- **80% = TRAINING SET**
  - Used to find optimal hyperparameters Gamma, C, Class weight via k-fold cross validation and to train algorithm

- **20% = TEST SET**
  - Used to evaluate performance on “NEW” data (never in training)

**Performance Check And Error Estimation**

- **Sample Variance**: Random simulations of splits for learning, training, and test

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Scaling, Feature Selection, Training and Optimization

**LEARNING SET (10% of total)**
(Features scaled by subtracting median, dividing by MAD)

**80% = TRAINING SET**
Used to find optimal hyperparameters gamma, C, class weight via k-fold cross validation and to train algorithm

**20% = TEST SET.**
Used to evaluate performance on “new” data (never in training)

**Performance Check And Error Estimation**

Sample Variance: Random simulations of splits for learning, training, and test

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Scaling, Feature Selection, Training and Optimization

LEARNING SET (10% of total)
(FEATURES SCALED BY SUBTRACTING MEDIAN, DIVIDING BY MAD)

80% = TRAINING SET
USED TO FIND OPTIMAL HYPERPARAMETERS GAMMA, C, CLASS WEIGHT VIA k-FOLD CROSS VALIDATION AND TO TRAIN ALGORITHM

20% = TEST SET
USED TO EVALUATE PERFORMANCE ON “NEW” DATA (NEVER IN TRAINING)

Performance Check And Error Estimation

Sample Variance: Random simulations of splits for learning, training, and test

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
LEARNING SET (10% of total) (FEATURES SCALED BY SUBTRACTING MEDIAN, DIVIDING BY MAD)

80% = TRAINING SET
USED TO FIND OPTIMAL HYPERPARAMETERS GAMMA, C, CLASS WEIGHT VIA k-FOLD CROSS VALIDATION AND TO TRAIN ALGORITHM

20% = TEST SET.
USED TO EVALUATE PERFORMANCE ON “NEW” DATA (NEVER IN TRAINING)

Sample Variance: Random simulations of splits for learning, training, and test

Learning Curves: Check for bias and variance

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Preliminary Results: How well can we measure the angular diameter distance?

<table>
<thead>
<tr>
<th></th>
<th>EW &lt; 20 A</th>
<th>ML (SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9 &lt; z &lt; 2.5</td>
<td>1e-4/0.32</td>
<td>2e-3/0.084</td>
</tr>
<tr>
<td>Contamination/Incompleteness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.9 &lt; z &lt; 2.5</td>
<td>1.23</td>
<td>1.065 ± 0.02</td>
</tr>
<tr>
<td>$\sigma(D_A)$ (arbitrary units)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 &lt; z &lt; 3.5</td>
<td>5e-2/0.26</td>
<td>.017/.27</td>
</tr>
<tr>
<td>Contamination/Incompleteness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 &lt; z &lt; 3.5</td>
<td>1.52</td>
<td>1.20 ± 0.04</td>
</tr>
<tr>
<td>$\sigma(D_A)$ (arbitrary units)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using ML approach reduces error on angular diameter distance by 20% both in low and high redshift bin.

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
Using additional emission lines to improve performance
Work in Progress (for me) and Open Questions (for you)

- Using additional emission lines to improve performance

![Plot](image)

- Observed-frame (Å)
- Flux (µJy)

**Figure 1:**

Left: The Lyman-$\alpha$ line from a $z=3$ galaxy observed exactly at the same wavelength as the OII line from a $z=0.3$ galaxy. The emission line flux (shown by the circle) is the same for these two galaxies in this example, but the intensity of the continuum (and therefore the equivalent width, EW) is different. On average, LAEs have much higher EWs than OIIs.

Right: Footprint of the two HETDEX survey areas, also showing overlap with other existing or planned data sets. There are many more ancillary data in one of the two regions; we will use them as a baseline to determine what additional data are needed to separate LAEs and OIIs in the other field.

Proportional to the square of its distance from us. Thus if we are looking at two galaxies in which the number of photons detected at the wavelength of the emission line is the same, but the first one has a much higher continuum than the second, it's likely that the first one is much closer to us (i.e., it is an OII emitter). This key ratio of (number of photons detected at the wavelength of the line)/(average number of photons detected at the wavelengths of the continuum) is called “equivalent width” (EW). LAEs have high emission line fluxes and dim continua (because they are farther away), so they have on average much higher equivalent widths.

The classification procedure so far adopted by the HETDEX team is to use a cut at EW = 20 Å (with objects whose EW is > 20 Å being classified as LAEs), but simulations show that this criterion induces a contamination fraction of over 4%, as shown in the middle panel of Fig. 2.

In the last few months, the PI has been working in collaboration with Prof. E. Gawiser and undergraduate student Andrew Leung at Rutgers to develop a better classification scheme based on Bayesian statistics. The main advantage with respect to the approach highlighted above is that additional information is used in separating LAEs from OII emitters, for example prior knowledge of how common LAE and OII galaxies of given mass are at their inferred redshifts. This is predicted to improve the contamination fraction, but is still limited by our imprecise and limited assumptions on the distribution of these objects in the Universe. For example, if the density of the environment had a strong correlation with the presence of LAEs or OIIs, but we didn’t assume a prior that takes it into account, the classification scheme would necessarily fail to capture this information.

**Proposed Research.**

The innovative approach that we propose here is to classify LAE/OII galaxies using supervised machine learning (SML). We plan to use ancillary data existing on some areas of the sky overlapping with the HETDEX survey (see Fig. 1) to identify a subset of galaxies for which a label (LAE or OII) can be assigned with a high degree of confidence. This subset is divided into a “training sample”, which is then used to “learn” what is the best rule to classify an object as an LAE or an OII emitter, and a “validation” sample, which is used to evaluate the performance of the classification procedure and make adjustments as needed. Contrary to the previous methods, no knowledge of the detailed physical model of galaxies is needed; the algorithm learns the classification rule exclusively by comparison with previous cases for which the correct answer was known.

Incidentally, this is a lot like our brain tends to learn; a simple example of how we “intuitively” apply machine learning in making decisions is described in the left panel of Fig. 2.

In principle, this method can perform better than any of those described above because there is no “assumption” involved in estimating whether an object is an LAE or an OII emitter, and there is no formal limit to how complicated the classification rule can be. On the other hand, the ability of the algorithm to learn the rule correctly depends crucially on the quality and size of the training sample, which needs to be a fair representation.
Work in Progress (for me) and Open Questions (for you)

I know how to do this

Please chip in

- Using additional emission lines to improve performance
- Trying other algorithms (SVMs are accurate and slow). Tried Random Forests, performance is slightly worse. Suggestions?
Work in Progress (for me) and Open Questions (for you)

- Using additional emission lines to improve performance
- Trying other algorithms (SVMs are accurate and slow). Tried Random Forests, performance is slightly worse. Suggestions?
- Data set is unbalanced and our decision criterion is skewed (less tolerant to type-I errors than type-II errors). Thoughts about using class weight as a hyper-parameter vs generating a classification probability and adjusting the threshold?
The Spectral Energy Distribution (SED) of a galaxy is a chart of the galaxy’s luminosity as a function of wavelength, $\lambda$

It contains information about the physical properties of the galaxy, such as mass, age of the stellar population, star formation history, and dust content.

The process of comparing models (whose properties are known) to the data in order to infer the properties of the data is known as SED fitting.
The Spectral Energy Distribution (SED) of a galaxy is a chart of the galaxy’s luminosity as a function of wavelength, $\lambda$.

It contains information about the physical properties of the galaxy, such as mass, age of the stellar population, star formation history, and dust content.

The process of comparing models (whose properties are known) to the data in order to infer the properties of the data is known as SED fitting.
The Spectral Energy Distribution (SED) of a galaxy is a chart of the galaxy’s luminosity as a function of wavelength, $\lambda$.

It contains information about the physical properties of the galaxy, such as mass, age of the stellar population, star formation history, and dust content.

The process of comparing models (whose properties are known) to the data in order to infer the properties of the data is known as SED fitting.
The Spectral Energy Distribution (SED) of a galaxy is a chart of the galaxy’s luminosity as a function of wavelength, $\lambda$.

It contains information about the physical properties of the galaxy, such as mass, age of the stellar population, star formation history, and dust content.

The process of comparing models (whose properties are known) to the data in order to infer the properties of the data is known as SED fitting.
How can we be accurate, fast and efficient?

1. By building faster and better algorithms:
   - GalMC (Acquaviva et al 2011)
   - SpeedyMC (Acquaviva et al 2011)
   - Many others (Others et al)
   
   [Link to SED fitting](http://www.sedfitting.org/SED08/Fitting.html)

2. By pre-processing data

   - In large data sets there will be many galaxies whose SEDs have the same shape **within the observational errors**.

   - If these galaxies can be grouped together *before* performing SED fitting, this affords a **factor of ~ N improvement in CPU time**, where $N$ is the average number of galaxies in each group.
Identifying Analog Galaxies

PRIOR WORK

☐ To my knowledge, the only method explored in Astrophysics literature so far for identifying analog galaxies is what could be called a “grid search algorithm” (Kriek et al 2012).

☐ In this method, the SED of every galaxy is compared to the SED of every other galaxy in order to find which galaxy has the most “analogs”.

☐ This group of galaxies is then removed from the list and the process is repeated in order to find the next largest group of analogs, until no more analogs are found.

OUR WORK

☐ We used a test catalog of data for a set of 5,000 CANDELS galaxies at $1 < z < 1.5$ (so they have similar RF coverage).

☐ We demonstrated that groups of analog galaxies exist, and they will have many more members for future larger surveys such as LSST.

☐ Problem: this method is very inefficient! ($20 \text{ hrs} + O(N^2)$)
Identifying Analog Galaxies

Idea: use Conceptual Hierarchical Clustering (COBWEB) to identify clusters of galaxies having similar shapes

GRID SEARCH

CLUSTERING

CLUSTERING 2.0

Classification is built in a few seconds;
Much better time complexity;
Allows a 10fold time saving in SED fitting

Satyanarayana
And Acquaviva,

Viviana Acquaviva (CUNY), Tools for Astronomical Big Data, March 2015
ML is able to find an intelligent classification with a 3-5% training set

I’d like some input on other algorithms worth trying besides SVMs

If you have strong opinions about class weight vs probabilistic classification, please let me know

Noise-weighted features will be implemented next

We still need to rigorously show that the properties of objects in clusters are similar

I believe this will be important not just to save CPU time, but for safe “stacking analyses” (see Vargas et al 2014)

More metrics could be implemented (e.g., color criteria)

My contact: vacquaviva@citytech.cuny.edu