Characterizing Overlapping Galaxies Using Machine Learning Techniques[*]

Luis Alvarez
Stanford University

(Dated: December 12, 2014)

As next generation telescopes enhance our ability to see dimmer objects in outer space, the number of optically overlapping galaxies in astronomical images increases. Measurements of galaxy shapes must now account for blended objects. This research attempts to study the classification of blended objects using machine learning techniques.

I. INTRODUCTION

Gravitational lensing, the bending of light through gravitational fields, provides insight into current problems in cosmology concerning the nature of dark matter and dark energy, mysterious components of the universe that do not emit light yet play a fundamental role in the time evolution of the universe.

Gravitational lensing studies are able to estimate the amount of dark matter and luminous matter by observing light emitted from galaxies distorted by dark matter; galaxy shapes are the key observables. Optically overlapping galaxies present a challenge to existing shape algorithms therefore classifying and resolving blended profiles into respective components are critical for lensing studies that will collect information on billions of galaxies such as the Large Synoptic Survey Telescope (LSST) beginning in 2019.

A. Definitions

In this paper we use the following terminology:

- A *postage stamp* describes an image that bounds a particular object of interest; a representative size is about 30 x 30 pixels depending on the size of the object.
- An *object* may consist of a blend of two or more distinct profiles or one profile.
- Each *object* contains a *flux density* that describes the number of counts per pixel.
- The *flux* is the sum of flux values over all pixels, or the total count.
- A *catalog* denotes the set of all objects, with corresponding postage stamps.

II. METHODOLOGY

A. Goal

We aim to predict whether the object in a given postage stamp is composed of two or more profiles or only one profile. This corresponds to two or more galaxies or only one galaxy.

B. Data Collection

We use an open-source galaxy simulation package called Galsim[3] to simulate astronomical images. We create single or blended objects using a number of flux densities and sampling methods and can simulate a catalog of postage stamps under the real conditions a future survey is likely to encounter. The Point-Spread Function (PSF), the model of the effects of atmospheric turbulence on images and sky-noise, the uncertainty in the brightness of the sky at a given moment during exposure times are modeled in Galsim and give it the power to approximate real conditions; Fig 1. provides such an example. For any single profile, we specify the following parameters:

- Flux (Photon Counts)
- Half-light Radius (Size)
- Ellipticity Component 1 (e1)
- Ellipticity Component 2 (e2)
- x_0 of Centroid
- y_0 of Centroid

with e1 and e2 such that $\sqrt{e_1^2 + e_2^2} \leq 1$

Using Galsim gives us the distinct advantage of creating postage stamps of blended and non-blended objects while knowing the class label. We begin our study in the **simplest scenario**: no sky noise, no PSF, and maintain each profile as a Gaussian such that blends are sums of Gaussians. Object parameters are distributed uniformly such that the majority of flux does not lie outside the postage stamp. We generate 1000 postage stamps with

[*] Thank you to Patricia Burchat, Mandeep S.S. Gill, and Joshua Evan Meyers for providing me with the wonderful opportunity to learn about weak lensing and for providing a rich and engaging learning environment.

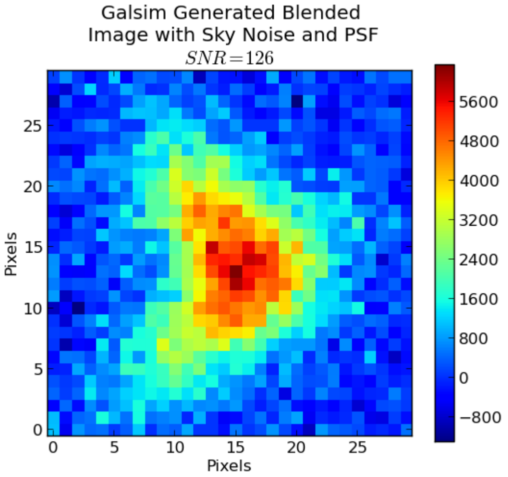

FIG. 1. Example Image

roughly half blends and half non-blends with up to 3 objects in one postage stamp.

C. Features

We consider a number of non-parametric statistics that identify different properties of the object in a given postage stamp.

1. *The Gini Coefficent*

The Gini Coefficent is a metric used more commonly in economics to describe the distribution of wealth in a society[1]. It was adapted for galaxy morphology classification in 2003 to quantify the relative distribution of flux within the pixels associated with the galaxy. Generally speaking, the Gini Coefficient is zero in an egalitarian society and unity in one individual has all the wealth. Analogously, the Gini Coefficient measures the spread of light in the postage stamp.

For a discrete distribution, the Gini Coefficient is defined as[1]:

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_i^n \sum_j^n |X_i - X_j| \tag{1}$$

where $\bar{X}$ is the mean over all pixel flux values, n is the number of pixels in the postage stamp and X_i is the flux in the ith pixel. We test to see if the Gini Coefficient is sensitive to overlapping profiles.

2. *Asymmetry*

Highly overlapping galaxies exhibit large asymmetry due luminosity profiles that are rotationally variant[2]. If a postage stamp of unique overlapping galaxies is rotated 180°and then subtracted from the original non-rotated postage stamp, many pixels from the residual image may be non-zero. This simple procedure produces a statistic that identifies whether the object in a given postage stamp is likely to be a blend. We use the following definition of asymmetry[2]:

$$A_{abs} = \frac{1}{2} \sum_i^n \frac{|X_{org,i} - X_{rot,i}|}{|X_{org,i}|} \tag{2}$$

Where $X_{org,i}$ is the flux count in the ith pixel of the original image, $X_{rot,i}$ is the flux count in the ith pixel of the rotated image, and n is the number of pixels in the postage stamp. We choose the center of rotation such that A_{abs} is minimized by computing A_{abs} in all pixels and choosing the minimal value of A_{abs} corresponding to the center of rotation.

3. *The Moment of Light*

The total second-order moment, M_{tot}, is defined as the flux in each pixel f_i multiplied by the squared distance to the center of the objects summed over all the galaxy pixels. We use the following definition[1]:

$$M_{tot} = \sum_i^n M_i = \sum_i^n f_i((x_i - x_c)^2 + (y_i - y_c)^2)) \tag{3}$$

Where x_i and y_i are the coordinates of the ith pixel and x_c and y_c are chosen such that M_{tot} is minimized. We then use following definition for M_{20} as the normalized second order moment of the brightest 20% of the flux of the galaxy[1]. To compute M_{20}, we sort the galaxy pixels by flux in decreasing order, sum M_i over the brightest pixels until the sum of the brightest pixels equals 20% of the total galaxy flux, and then normalize by M_{tot}:

$$M_{20} = \log \frac{\sum_i^n M_i}{M_{tot}} \quad \text{while} \quad \sum_i^n f_i < 0.2 f_{tot} \tag{4}$$

where f_{tot} is the total flux count of all pixels. By setting x_c and y_c to be minimal for the entire postage stamp, Lotz et al. state that M_{20} is senstitive to overlapping galaxies. We go on to test this by generating M_{20} over all pixels for visual insight and choose the minimum value.

4. *Centroid Estimation*

Since blended objects are sums of individual profiles with peaked distributions, a subset of blended profiles are likely to exhibit multiple peaks. Using a computer vision and image processing library called Mahotas[4], we are able to identify regions in a postage stamp that are disjoint. For this process, we run through a range of pixel flux values set by the max pixel flux observed. For each value in our range, we obtain the threshold pixel value, zero all pixel flux values lower than the threshold, smooth the image with a gaussian filter to minimize noisy pixels and identify if disjoint regions are present. We then choose the max number of centroids observed in the iteration.

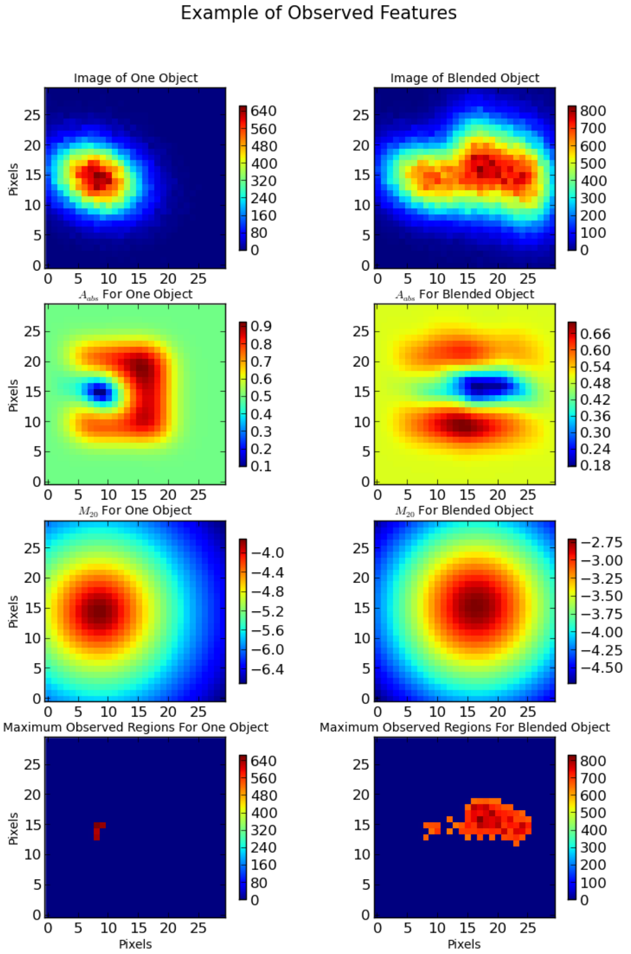

FIG. 2. Example Feature Computation For a Blend and Non-Blend

D. Modeling

We choose the following supervised learning methods from the Python sklearn library[5] for binary classification:

- Multinomial Naive Bayes (MNB)
- Linear Logistic Regression (LR)
- Support Vector Machine (SVM)
- Random Forests (RF)

Although Random Forests were not discussed in class, current literature ranks RF above SVM's for their non-parametric nature, ability to deal with non-linear-separability, and operational ease.
We first begin optimizing parameters for MNB. Because MNB can be bucketed differently, we ask what is the generalization error for differently discretized MNB models.

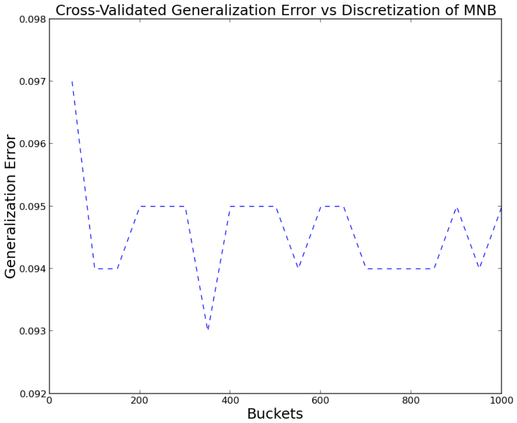

FIG. 3. Generalization Error vs Bucketing for MNB

Analysis of FIG. 3. returns 350 buckets for the lowest estimate of the generalization error; we go on to use that discretization for MNB.

For LR and SVM, we ask what regularization constant provides the lowest cross-validated generalization error translating to soft vs hard margins.

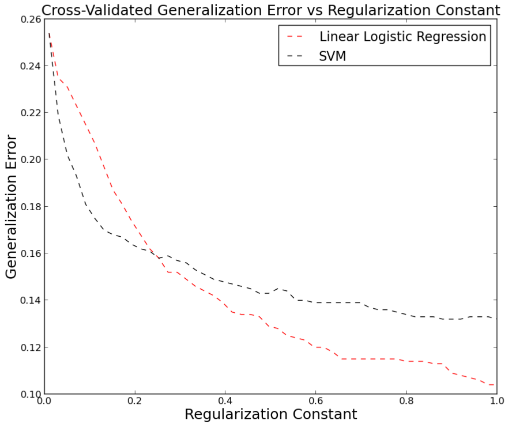

FIG. 4. Generalization Error vs Regularization Constant For LR, SVM

We conclude that the soft margin corresponding to $C = 1$ provides the lowest generalization error from FIG 4.

Finally, we ask how many decision trees will provide the lowest cross-validated generalization error for RF estimation. Analysis of FIG. 5 returns 45 decision trees is optimal. We now use the parameters set by the above analysis to create optimal learning curves.

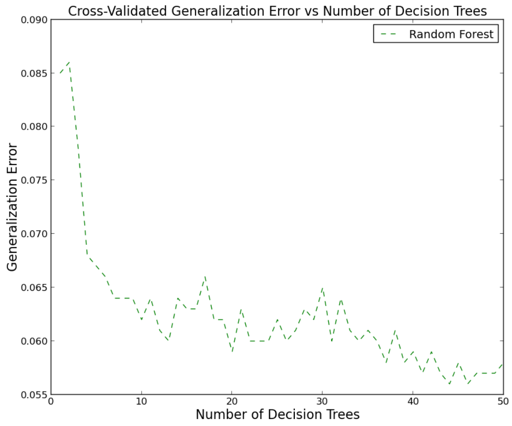

FIG. 5. Generalization Error vs Number of Decision Trees For RF

III. RESULTS

1. *Supervised Learning*

To compute learning curves, we run through a different number of training examples, increasing the number through each iteration, compute the cross-validated training error for each model and plot. The results are as follows:

According to FIG. 6. MNB and RF perform well attesting to the strength of MNB in the small training set regime as a high-bias/low-variance classifier and the strength of RF as an ensemble method with high estimation power in any training set regime. As the training set size increases, separability of data may become reduced, influencing MNB to perform worse than at the smaller training set regime. LR and SVM converge asymptotically and RF provides the lowest generalization error. We now perform a similar analysis on the test set.

Using the parameters defined in our previous analysis, we use our models on our test set to compute learning curves.

Comparing the learning curves from our training to the test set returns a similar pattern (FIG. 7.). RF returns the lowest asymptotic generalization error with about 0.05 generalization error while MNB performs well in the low training set regime with 0.08 generalization error.

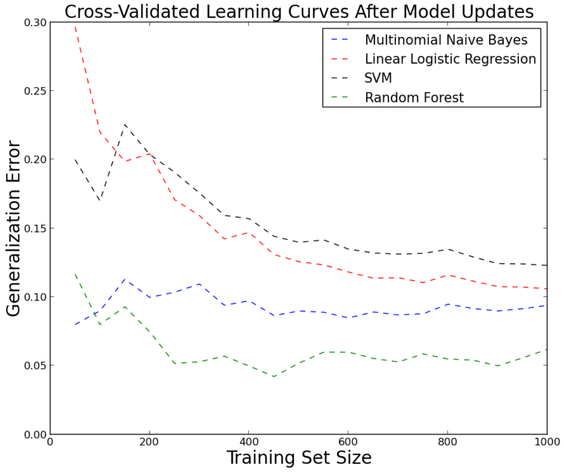

FIG. 6. CV Learning Curves For Training Set

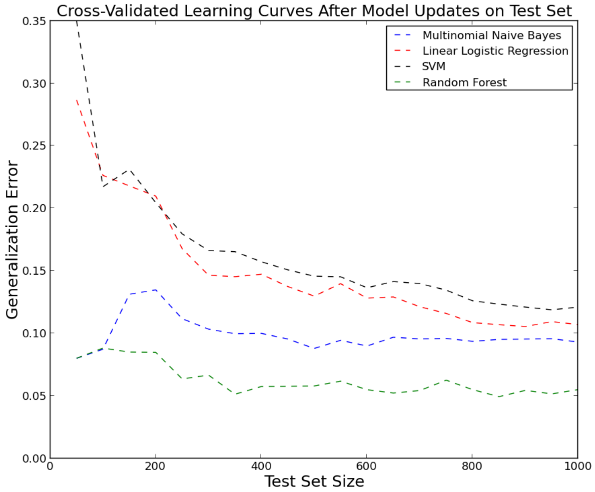

FIG. 7. CV Learning Curves For Test Set

2. *Unsupervised Learning*

Now we ask, for binary classification, does the data show structure such that there exist different modes corresponding to the two class labels.

For each feature that is defined continuously, we use a gaussian mixture model to compute the modes in the data. We input 2 modes corresponding to the binary classes. We overlay the fitted gaussians above the true data corresponding to the blended and non-blended data.

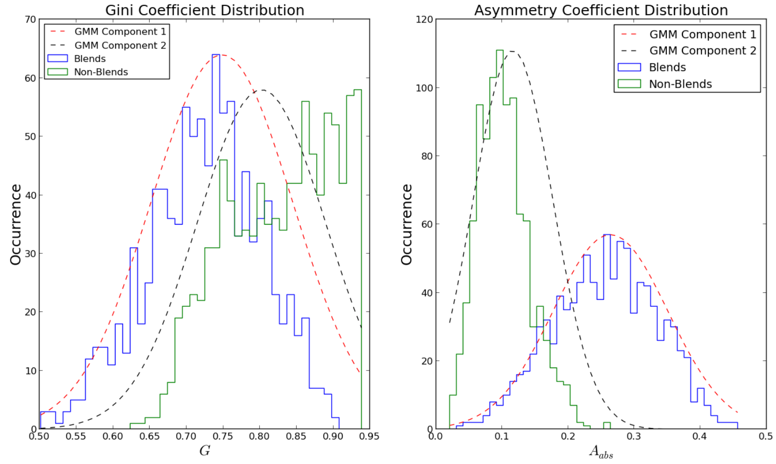

FIG. 8. GMM Models for G and A_{abs}

On Fig 8. and 9, we see that for A_{abs} and M_{20}, the data separates quite cleanly. For blended objects, A_{abs} and M_{20} have much more variation while generally taking on greater values than those of non-blended objects. A_{abs} and M_{20} for non-blended objects have low variation with a high peak. This translates to blends generally being more asymmetric and blended objects are generally more spatially extended.

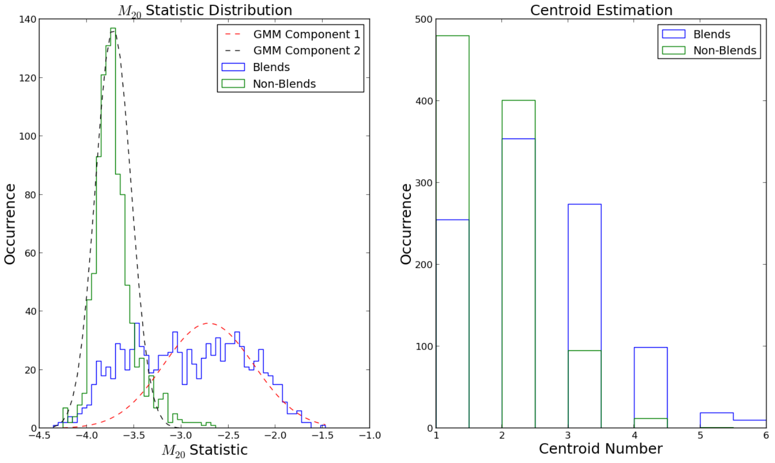

FIG. 9. GMM Model for M_{20} and Centroid Estimate Histogram

The same conclusion as before does not necessarily apply for G; the GMM returns two closely overlapping distributions mimicking the large overlap between many of the observed values. While the separation is not as clean as for A_{abs} and M_{20}, G roughly takes on larger values for non-blended objects. This translates to blended objects generally spreading light more in a given postage stamp while non-blended objects contain more light in smaller region.

On centroid estimation, we note that the centroid estimates for blended objects favored number of centroids greater than 1 with mode at 2 and conversely for non-blended objects, the mode was 1 estimated centroid. The shape of the distribution bolsters the accuracy of the centroid estimation although further tuning of parameters should be considered for much higher accuracy.

IV. CONCLUSION AND FUTURE DIRECTION

Using Galsim, we have simulated a number of postage stamps, 1000 each in our training set and test set with blended and non-blended gaussian profiles. For each profile, we selected random parameters in order to collect non-parametric statistics on each postage stamp. After obtaining the features of each postage stamp, we used cross-validation techniques to identify the parameters of each model that provided the minimum cross-validation generalization error on our training set. Using those models: MNB, LR, SVM, and RF, we constructed learning curves on both the training and test set and conclude that RF does the best asymptotically with 0.05 generalization error. In the low training set regime, RF and MNB perform well with under 0.10 generalization error.

In terms of feature analysis, we have shown that for the continuous features, the modes in the data corresponding to our two class labels are fully realized and can be cleanly separated, giving insight to the success of MNB.

Future work consists of moving towards more realistic modeling. Galaxies are generally modeled as a combination of two distinct profiles that approximate the luminous bulge and the galaxy disk. Inclusion of sky-noise and the PSF at signal-to-noise ratio regimes similar to what LSST may encounter and researching more features that may be sensitive to overlapping galaxies are also necessary.

[1] Lotz, J., Primack, I., Madau, P.: A New Non-Parametric Approach to Galaxy Morphological Classification Astron.J.128:163-182, 2004

[2] Conselice, C., Bershady, M., Jangren, A.: The Asymmetry of Galaxies: Physical Morphology for Nearby and High-Redshift Galaxies Astrophys,J.529:886, 2000 ApJ

[3] Http://www.great3challenge.info/?q=galsim.

[4] Mahotas.readthedocs.org.

[5] Scikit-learn.org