CS 229: Observing Dark Worlds

Anubhav Singla, Nandan Sawant
Guide: Prof. Andrew Ng

Introduction

Dark Matter Halo is a hypothetical component of a galaxy whose mass dominates the total mass of the galaxy. Dark matter is said to neither absorb nor emit light and therefore cannot be directly observed. However, the existence of dark matter halo can be inferred. It bends the space-time such that any light from a background galaxy will have its path altered making the galaxy appear elliptical. In this project, we investigate the problem [1] of predicting the halo locations based on ellipticities of galaxies in the sky.

Problem and Data

The data[1] consists of 300 training skies and 120 test skies. Each sky contains 400-800 galaxies. Locations (x & y co-ordinates) and ellipticities (e1 & e2) of the galaxies are known. A sky can contain 1, 2 or 3 halos. The number of halos in each sky is told to us a priori. Locations (x & y co-ordinates) of the halos in the training skies are provided to us.

Our goal is to accurately predict locations of the halos in the test skies. Evaluation metric[2] is the average distance between predicted and actual halo locations, with an additional penalty term for any angular bias.

Model

The ellipticity induced by the dark matter halo on the galaxy is tangential. We assume that the ellipticity induced is radially symmetric and a function of the distance from the halo. We denote the ellipticity induced by the halo on the galaxy at distance r by f(r).

The ellipticity of a galaxy at point (x,y) tangential to the halo at (x',y') is $e_{tangential} = -(e_1 \cos(2\emptyset) + e_2 \sin(2\emptyset))$, where $\emptyset$ is the angle of the galaxy with respect to the point (in our case, the dark matter center) given by $\emptyset = \arctan\left(\frac{y-y'}{x-x'}\right)$.

However, the galaxies are inherently elliptical (and not circular). Moreover, we are approximating 3 dimensional distances and ellipticities in 2 dimensions. We model the effect of all these factors as an additive white Gaussian noise.

For a galaxy at point (x, y), we can write:

$$e_{tangential}(x, y, x', y') = f(r) + AWGN$$

$where$, (x', y') is the halo location and $r = \sqrt{(x-x')^2 + (y-y')^2}$

Converting back to e_1 and e_2,

$$e_1(x, y) = f(r)\cos(2\phi) + AWGN$$
$$e_2(x, y) = f(r)\sin(2\phi) + AWGN$$

In case we have N halos in the sky,

$$e_1(x, y) = \sum_{i=1}^{N} f(r_i)\cos(2\phi_i) + N(0, \sigma^2)$$

$$e_2(x, y) = \sum_{i=1}^{N} f(r_i)\sin(2\phi_i) + N(0, \sigma^2)$$

In vector notation,

$$\boldsymbol{e}(x,y) = \sum_{i=1}^{N} \boldsymbol{f}(r_i) + N(0,\Sigma)$$

where, $\mathbf{e} = [e1 \quad e2]^T$ and $\boldsymbol{f}(r_i) = [f(r_i)\cos(2\phi_i) \;\; f(r_i)\sin(2\phi_i)]^T$.

For each sky, we are given **e**(x,y) for all the galaxies. We want to predict halo locations (x',y') that best explain these observations. We use the model described above to design the prediction algorithm.

Effect of Halo versus Distance

In this section, we learn f(r), the induced ellipticity as a function of distance from the halo.

The function f(r) can be derived using the theory of gravitational lensing but this approach would require a deep understanding of the subject matter. We tried several data driven approaches – linear regression, polynomial fit (of up to 5th order) and locally weighted linear regression (LWLR). LWLR seemed to give us the best results, which we describe below. This isn't surprising given the enormous amount of data (60,000 data points), which ensures that LWLR does not over-fit and at the same time, provides more degrees of freedom than a linear regression or a small order polynomial fit.

We first look at the training skies with only one halo. For each sky, we compute the tangential ellipticities of the galaxies with respect to the halo location. We run a **locally weighted linear regression** between the computed tangential ellipticities and the distance between the halo and the galaxy. We see the following results:

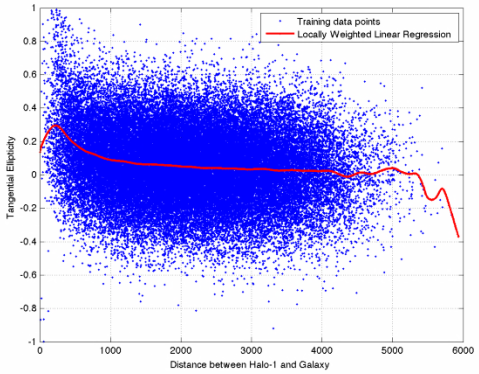

Figure 1: LWLR curve modeling induced ellipticity as a function of the distance between halo and the galaxy

Figure shows (the red curve) our estimate for f(r). We also plot the histogram of deviation from the LWLR estimate. This reaffirms our assumption that the noise is AWGN.

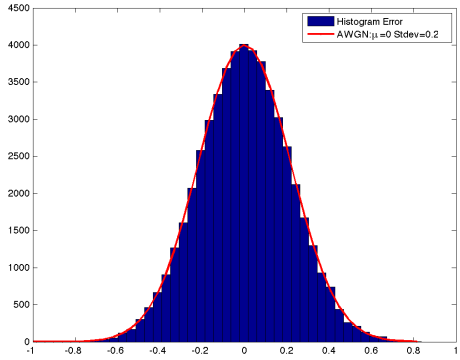

Figure 2: The deviation from the LWLR fit of induced ellipticity

Further, we conjecture that f(r) is not identical for all halos. In the physics context, we can attribute this to halos having different masses and sizes. Based on our crude data analysis, we assume that there are 2 types of halos and run an **unsupervised 2-means clustering algorithm**. The labels were assigned to the halos based on the minimum norm while the cluster centroids were computed using LWLR. Notice that this is slightly different from the traditional k-means clustering. Figure below shows the two corresponding LWLR cluster centroids.

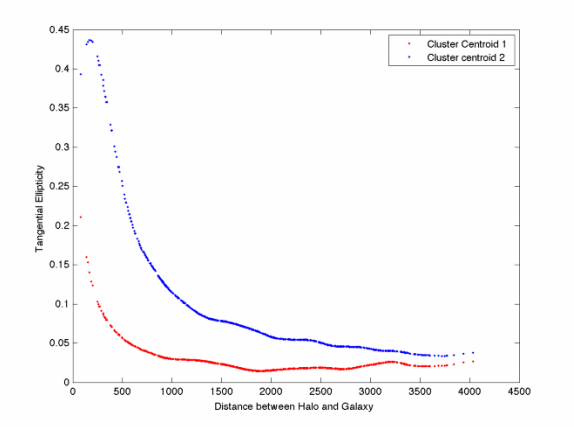

Figure 3: LWLR cluster centroids after 2-means clustering

We also observe that a significant fraction of halo 1s was mapped to the (blue) cluster centroid 1 while most of halo 2 and 3 were mapped to the (red) cluster centroid 2. Moreover, running clustering algorithm on halo 2 and 3 alone does not yield such separation. This leads us to the conclusion that halo 1 is much stronger than halo 2 and 3. In further discussion, we model halo 1 by $f_1(r)$ and halo 2 and 3 together by $f_{23}(r)$.

Matched Filter Approach

In the last section, we showed that the effect of halo 1 is much stronger than the effect of halo 2 and 3. Therefore, we can estimate location of halo 1 independent of halo 2 and halo3, treating interference from halo 2 and 3 as noise.

We discretize the problem by dividing the sky into a fine grid. Center of each grid square is a candidate for the halo 1 location. If the sky contains G galaxies, each candidate halo location can be represented by a G dimensional feature vector where g^{th} feature corresponds to tangential ellipticity at the galaxy g with respect to the candidate halo location.

With these feature vectors at disposal, we can potentially model the problem as a logistic regression and train it using a soft-max or SVM classifier. However, given the variable number of galaxies in each sky, small number of training skies (300) as compared to the typical size of such feature vector (400-800) and a priori knowledge of number of halos in the sky, we instead use a **Matched Filter**[3] technique known to be optimal for signal processing problems with Gaussian noise profiles.

Let $\mathbf{e^{ij}}$ be the observed ellipticity feature vector for the candidate halo location (i, j). According to Matched Filter technique, we pick, halo 1 location (x',y') as

$$\arg\max_{(i,j)} \frac{\boldsymbol{f}_1^{ij\,T} e^{ij}}{\|\boldsymbol{f}_1^{ij}\|}$$

Where, $\mathbf{f_1^{ij}}$ is the predicted ellipticity feature vector based on the distance based modeling described in Figure 3. Note that this is nothing but the weighted average of the components of the feature vector i.e. this is of the form $\theta^T x$. Figure below shows the matched filter output over the candidate halo locations.

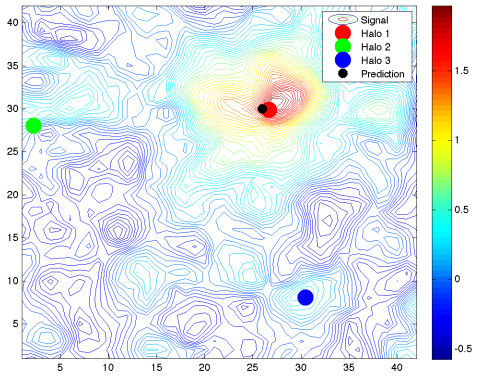

Figure 4: Matched filter output for Halo 1 Prediction. Red dot is the actual halo 1 location while black dot is the predicted location for halo 1.

After predicting halo 1, we subtract the effect of halo 1 for each galaxy and re-compute the feature vectors. We use similar matched filter technique for halo 2 and 3 and choose halo locations (x',y') such as

$$\arg\max_{(i,j)} \frac{\boldsymbol{f}_{23}^{ij\ T} e^{ij}}{\|\boldsymbol{f}_{23}^{ij}\|}$$

where, $\mathbf{f_{23}^{ij}}$ is the predicted ellipticity feature vector based on the distance based modeling described in the previous section. Figure below shows the matched filter output of the sky after subtracting the effect of halo 1.

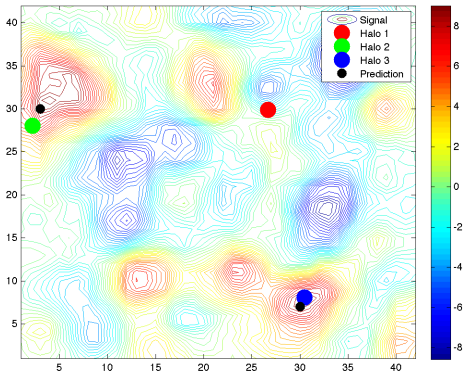

Figure 5: Matched filter output after removing Halo 1. Notice that the red cluster indicating high values around halo 1 has disappeared. Black dot indicates the prediction while green and blue dots indicate actual Halo 2 and 3 locations.

The process above is repeated for each sky.

Additional Heuristics

We also implement some additional heuristics to improve the performance of our system.

Zero Padding

We observe that the locations of the two halos in the sky are sufficiently separated. Therefore, after detecting a particular halo, we zero pad the area around the predicted location making the likelihood of finding the next halo 0 in the close-by grid squares.

Low Pass Filtering

Nearby grid squares in the sky would have a similar likelihood of being the halo locations. Therefore, we pass the output of the matched filter through a low pass filter to average out any abrupt variations.

Further classifying Halo 1

We further classify the halo 1 into two categories – 'strong' and 'weak' and use a separate distance dependence of induced ellipticity for each class, the detail that was omitted in earlier sections for simplicity.

Results

Our algorithm does an excellent job of predicting halo 1 locations. The scatter plot below testifies for the accuracy of our predictions.

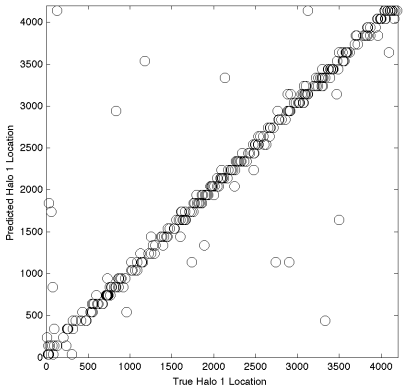

Figure 6: Predicted locations vs. actual locations for Halo 1. Note that most predictions fall around the slope 1 line

We achieve the average halo1 distance error as low as 50 units. The evaluation metric[2] of 0.17 can be achieved for halo 1.

However, the performance for halo 2 and 3 is not as great. Figure below shows the scatter

plot of predicted halo locations versus actual locations for halo 2.

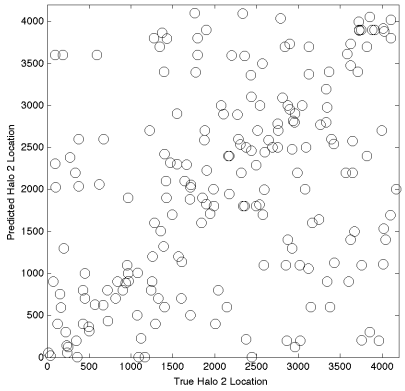

Figure 7: Predicted locations vs actual locations for Halo 2.

Notice that, although the points are close to the slope 1 line to an extent, there are a large number of misplaced halos increasing the average distance error to around 700-1000 units (for randomly chosen 120 skies).

We have a **theoretical explanation** for this performance. From Figure 2, the standard deviation of AWGN noise σ is 0.2. The matched filter computes the weighted sum over all the galaxies, which brings down the noise to $\hat{\sigma} = \frac{\sigma}{\sqrt{G}} \sim 0.0075 - 0.01$, where G is the number of galaxies. Based on the LWLR curve found in Figure 3, the magnitude of average tangential ellipticity induced by Halo 1 is 0.035. Halo 1 is therefore accurately estimated. On the other hand, the average induced ellipticity for Halo 2 and 3 is close to 0.01. This value being close to the magnitude of the noise leads to poor prediction.

To provide a better measure of performance, we randomly chose 120 skies out of the 300 training skies provided to us, as our test data. We compute the evaluation metric for this test set. We repeat this process for 1000 such test sets. The **histogram** of the performance is shown below. We also use this histogram to fine tune the parameters of our algorithm.

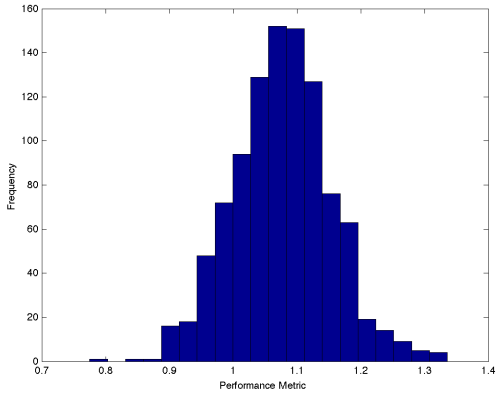

Figure 8: Histogram of performance metric for randomly chosen test sets of 120 skies each

One can notice that the performance averages around the score of 1.07 with standard deviation of 0.07 and the best score of 0.78. Our score on Kaggle's public test set is 1.13. Here is a summary of the performance:

Benchmark	**Score**
Randomly Placed Halos	1.95
Gridded Signal Benchmark	1.58
Lenstool*	1.02
Our algorithm**	1.07

* Based on theory of gravitational lensing

** Averaged over the training set

Acknowledgements

We'd like to thank Prof. Ng and the entire CS229 teaching staff for their able guidance. We'd like to thank Kaggle.com for the dataset.

References

[1] http://www.kaggle.com/c/DarkWorlds/

[2] http://www.kaggle.com/c/DarkWorlds/details/evaluation

[3] Wireless Communications by Andrea Goldsmith, Cambridge Univ. Press, 2005