Joint Stellar Mass - Redshift PDFs using Random Forest

SUNIL MUCESH

SUPERVISORS: PROF. OFER

Random Forest: Introduction

!

Random Forest: Algorithm

A random forest consists of many decision trees with a few tweaks.

1. Sample randomly from data with replacement.
2. Choose only a subset of features.
3. Create a decision tree from the bootstrapped sample.
4. Repeat to create a random forest.

To make a prediction:

- Classification - Majority Vote
- Regression - Average

Incorporating Errors

- We can incorporate errors in the data into the algorithm.
- One way of doing this is to scatter the magnitudes of a galaxy according to the errors multiple times.
- Draw errors randomly from a gaussian distribution centred about the magnitude and with standard deviation given by the error.

Results: Point Estimates

- Trained two RFs to predict redshift and stellar mass with 80% data.
- Input features: magnitudes in *griz* bands + colours (*g*-

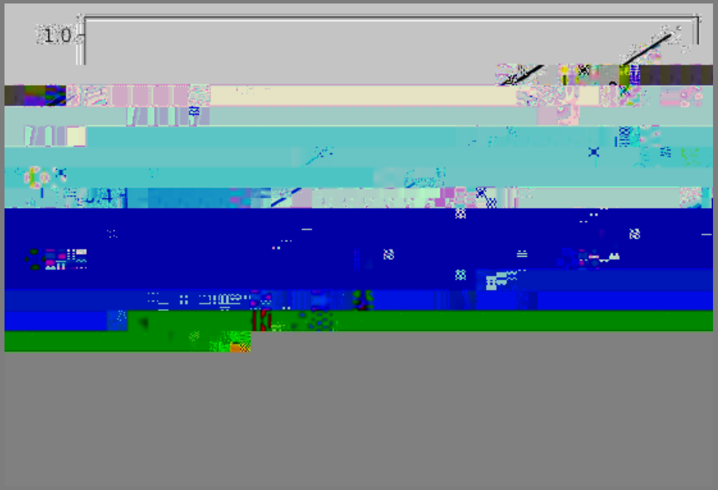
Probability distributions

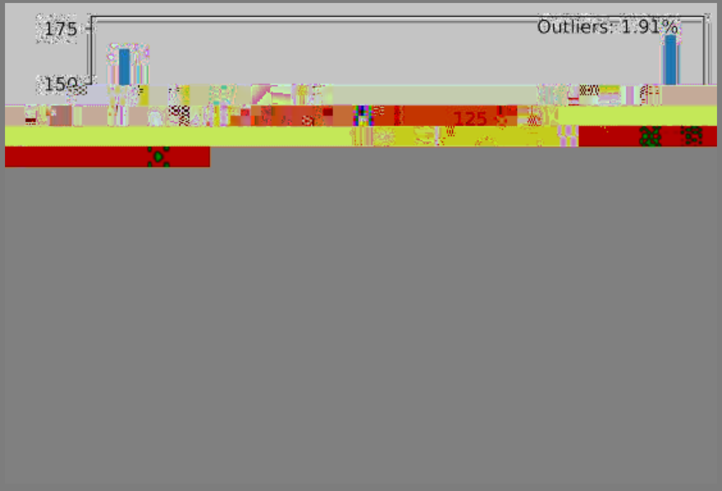
- Random forest can be described as a clustering algorithm.
- It aims to group together similar galaxies and these end up in the same leaf nodes of the decision trees.
- For a point estimate, we averaged the redshift or stellar mass values of the galaxies in leaf nodes.
- To extract a probability distribution, we can simply gather all the values in the leaf nodes in all the decision trees.

How accurate are the extracted PDFs?

- Unlike point estimates, the ‘true’ PDFs are not available for comparison.
- To get started, we can compare the true value to the extracted

Results: Redshift & Stellar Mass PIT distributions





Theory: Joint PDFs

Simultaneous Method:

- Build one model which predicts redshift and stellar mass simultaneously.
- Extract the joint distribution.

Separate Method:

- Build two separate models, one which predicts redshift and another which predicts stellar mass given a redshift.
-

Steps: Separate Method

1. Train the first model to predict redshift.
2. Train the second model to predict stellar mass but include redshift + all features used in the first model.
3. For a test galaxy, extract the marginal pdf of redshift from the first model.
4. For each value of redshift, run the second model to extract conditional pdf of stellar mass | redshift. All the other features are kept the same.
5. Bin each conditional probability distribution into fixed redshift and stellar mass bins.
6. Finally, multiply the binned conditional probability distributions by the marginal pdf of redshift to get the joint pdf.

$$f(M,z) = f(M|z) * f(z)$$

Results: Joint PDFs

Simultaneous

Separate

Thank you. Any questions?