Simple machine learning techniques for galaxy classification

MSc - II Practical

I. Types of galaxies and their classification

The most common basis for classification of galaxies is their morphology. Accordingly, galaxies can be classified into ellipticals, spirals, lenticulars and irregular structures. There are also galaxies which show clear signs of interactions and mergers. Visual inspection to do manual classification however, has its own drawbacks. It brings subjectivity into the process, and is a time-consuming activity. In this exercise, we learn to use simple machine learning methods which can be applied to classify galaxies in a more efficient way.

II. Machine learning

Machine learning is an approach by which computers and algorithms can accomplish acts of learning similar to how humans learn. This can be very useful in trying to identify patterns across very large sets of data, and are particularly handy exploration tools in arriving at empirical relations which may be probed further for the underlying science.

The key first step in machine learning is to split the entire data set into a training fraction and a test fraction. The training fraction is used by the algorithm to learn to identify patterns in the data set, which are then applied to the test data. If the learning during training has been effective, the results will be more accuracte on application to test data. In order to make the learning efficient, a large fraction of the data set must be allotted to the training fraction.

Features are input data for machine learning models, which then maps them to a target variable. The accuracy of the learning can be represented using a confusion matrix which compares the actual classification value with that provided by the algorithm.

Decision trees work by learning simple decision rules, while random forests employ several decision trees to improve the accuracy and fitting. Both algorithms may be applied to regression as well as classification-based problems.

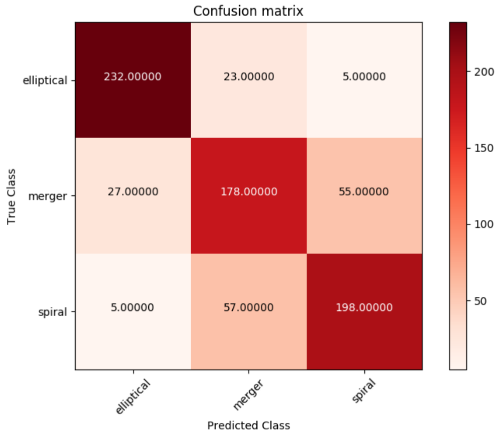

Figure 1: Confusion matrix; Credit: Data-driven Astronomy course (Coursera)

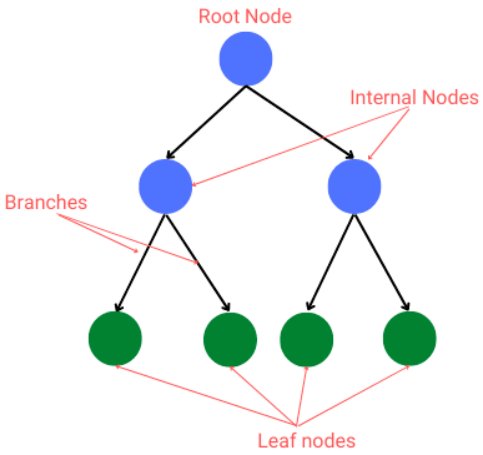

Figure 2: Decision tree; Credit: `https://www.machinelearningnuggets.com/decision-trees-and-random-forests/`

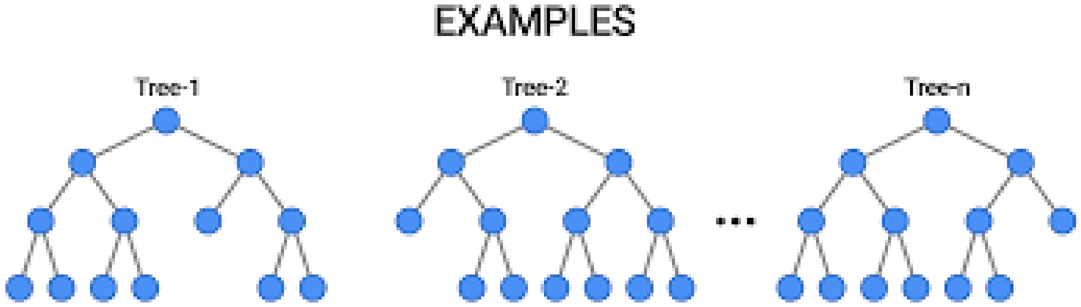

Figure 3: Random forest; Credit: `https://towardsdatascience.com/how-do-random-forests-decision-trees-decide-simply-explained-with-an-example-in-python-6737eb183`
`gi=aed878eaf5d7`

III. Practical exercise

- Install *scikit-learn*.
- Read in the Numpy binary (.npy) file provided to you using the *numpy.load()* function. The data set is made up of various colours u-g, g-r, r-i and i-z, eccentricity, adaptive moments in each filter and the Petrosian flux values for several galaxies.
- Split the data into training and test sets.
- Generate features and targets.
- Set up a decision tree classifier, train it and then apply it to the test data.
- Determine the accuracy of the results by making a confusion matrix.
- Repeat the training and testing process using a random forest classifier. Again, plot a confusion matrix.

IV. References

https://www.ibm.com/topics/machine-learning
https://scikit-learn.org/stable/
https://scikit-learn.org/stable/modules/tree.html
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/classification.html#star-quasar-classification-naive-bayes

***