Application of Machine Learning Methods on Astronomical Databases

Predicting Galaxy Redshifts Using Convolutional Neural Networks: A Comparative Study with the Gaia Unresolved Galaxy Classifier

Apostolos Kiraleos

Abstract

Galaxy redshift is a crucial parameter in astronomy that provides information on the distance, age, and evolution of galaxies. In this thesis, we explore the use of convolutional neural networks (CNNs) for predicting galaxy redshifts from spectra, using data from the Gaia mission. We compare the performance of our CNN model with that of the Gaia Unresolved Galaxy Classifier (UGC), which uses support vector machines (SVMs) for classification.

Contents

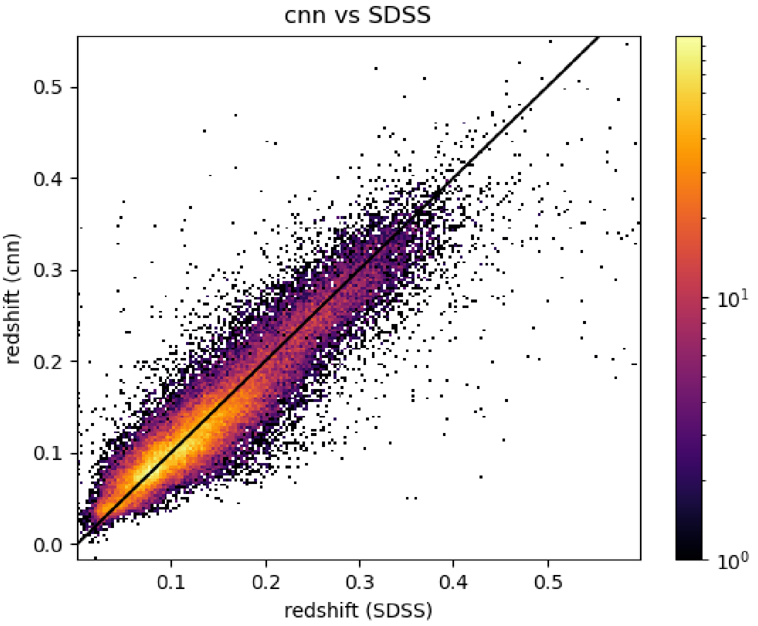

Figure 1: Pandoc Logo

Chapter 1: Introduction

1.1 Gaia obvservatory

The Gaia mission is a European Space Agency (ESA) space observatory that has been in operation since 2013. The primary goal of the mission is to create a three-dimensional map of our galaxy, the Milky Way, by measuring the positions, distances, and motions of over two billion stars.

At its core, Gaia is equipped with two optical telescopes accompanied by three scientific instruments, which collaborate to accurately ascertain the positions and velocities of stars. Additionally, these instruments disperse the starlight into spectra, facilitating detailed analysis.

Throughout its mission, the spacecraft executes a deliberate rotation, systematically scanning the entire celestial sphere with its two telescopes. As the detectors continuously record the positions of celestial objects, they also capture the objects' movements within the cosmos, along with any alterations therein.

Over the duration of its mission, Gaia conducts approximately 14 observations annually for each of its designated stars. Its primary objectives include the precise mapping of stellar positions, distances, motion patterns, and variations in luminosity. Gaia's mission anticipates the revelation of an extensive array of novel celestial entities, encompassing exoplanets and brown dwarfs, alongside the thorough examination of hundreds of thousands of asteroids located within our own Solar System. Furthermore, the mission encompasses an investigation into more than 1 million distant quasars while subjecting Albert Einstein's General Theory of Relativity to rigorous new assessments.

1.2 Galaxy redshift

Galaxy redshift is a fundamental astronomical property that describes the relative motion of a galaxy with respect to Earth. The redshift of a galaxy is measured by analyzing the spectrum of light emitted by the galaxy, which appears to be shifted towards longer wavelengths due to the Doppler effect. Redshift is a crucial parameter in astronomy, as it provides information about the distance, velocity, and evolution of galaxies.

The redshift of a galaxy is measured in units of "z", which is defined as the fractional shift in the wavelength of light emitted by the galaxy. Specifically, the redshift "z" is defined as:

$$z = (\lambda_o - \lambda_e)/\lambda_e$$

where λ_o is the observed wavelength of light from the galaxy, and λ_e is the wavelength of that same light as emitted by the galaxy. A redshift of $z = 0$ corresponds to no shift in the wavelength (i.e., the observed and emitted wavelengths are the same), while a redshift of $z = 1$ corresponds to a shift of 100% in the wavelength (i.e., the observed wavelength is twice as long as the emitted wavelength).

Accurate and efficient estimation of galaxy redshift is therefore essential for a wide range of astronomical studies, including galaxy formation and evolution, large-scale structure of the universe, and dark matter distribution. However, measuring galaxy redshifts can be a challenging task due to various factors such as observational noise, instrumental effects, and variations in galaxy spectra.

In recent years, machine learning techniques have emerged as powerful tools for galaxy redshift estimation, leveraging large datasets and complex algorithms to improve the accuracy and efficiency of the estimation process. In this thesis, we focus on the application of convolutional neural networks (CNNs) to galaxy redshift estimation, and compare their performance with the Gaia Unresolved Galaxy Classifier (UGC).

1.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of artificial neural network (ANN) that has proven to be highly effective for tasks involving image and video analysis, such as object detection, segmentation, and classification. CNNs are inspired by the structure and function of the visual cortex in animals, which contains specialized cells called neurons that are tuned to detect specific visual features. In a similar way, CNNs are designed to learn and extract meaningful visual features from raw image data.

At the core of a CNN are individual processing units called neurons, which are organized into layers. Each neuron receives input from other neurons in the previous layer, applies a mathematical operation to that input, and passes the output to the next layer. The output of the final layer of neurons is the predicted output of the network for a given input.

The key innovation of CNNs is their use of convolutional layers, which enable the network to automatically learn and extract local spatial features from raw input data. In a convolutional layer, each neuron is connected only to a small, localized region of the input data, known as the receptive field. By sharing weights across all neurons within a receptive field, the network can efficiently learn to detect local patterns and features, regardless of their location within the input image.

CNNs typically also include pooling layers, which downsample the output of the previous layer by taking the maximum or average value within small local regions. This helps to reduce the dimensionality of the input and extract higher-level features from the local features learned in the previous convolutional layer.

The final layers of a CNN are typically fully connected layers, which take the outputs of the previous convolutional and pooling layers and use them to make a prediction. In the case of image classification, for example, the output of the final fully connected layer might be a vector of probabilities indicating the likelihood of each possible class.

Overall, CNNs are a powerful and flexible tool for image analysis tasks, and have achieved state-of-the-art performance on many benchmark datasets. In this thesis, we explore the application of CNNs to galaxy redshift estimation, and compare their performance to traditional methods like the UGC classifier. We also investigate the use of transfer learning, data augmentation, and other techniques to improve the performance of CNNs on this task.

1.4 Support Vector Machines

Support Vector Machines (SVMs) are a class of supervised machine learning algorithms that excel in classification and regression tasks. Introduced by Vladimir Vapnik and his colleagues in the 1960s, SVMs have gained widespread popularity and are widely used in various fields, including image classification, text analysis, and bioinformatics.

At their core, SVMs are based on the idea of finding a hyperplane that best separates data points belonging to different classes in a high-dimensional feature space. This hyperplane is chosen in such a way that it maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class. These nearest data points are known as support vectors, hence the name “Support Vector Machines.”

SVMs are particularly well-suited for situations where the data is not linearly separable, meaning that a simple straight line (hyperplane) cannot cleanly divide the data into different classes. To address this, SVMs use a mathematical technique called the kernel trick. This allows SVMs to implicitly map the data into a higher-dimensional space where it becomes linearly separable. Common kernel functions include the linear, polynomial, radial basis function (RBF), and sigmoid kernels.

Chapter 2: Our problem