Application of Machine Learning Methods on Astronomical Databases

Apostolos Kiraleos

Abstract

Galaxy redshift is a crucial parameter in astronomy that provides information on the distance, age, and evolution of galaxies. This dissertation investigates the application of machine learning for predicting galaxy redshifts. It involves the development and training of a neural network to analyze galaxy spectra sourced from the European Space Agency's Gaia mission and showcases the practical implementation of machine learning in astronomy.

Table of Contents

List of Figures

1 Introduction

1.1 Object

This dissertation primarily addresses the critical issue of accurate galaxy redshift estimation. Galaxy redshifts are pivotal in modern astronomy, providing crucial insights into cosmic distances, universe evolution, and more. The topic is of significant interest and relevance within the field.

1.2 Purpose and Objectives

The main goal of this dissertation is to advance galaxy redshift estimation by applying advanced machine learning techniques, specifically Convolutional Neural Networks (CNNs). Our objectives include evaluating the accuracy of trained CNN models in predicting galaxy redshifts and comparing them with other methods like Support Vector Machines (SVMs). We aim to understand the practical applications of these models in astronomical research.

1.3 Methodology

To address these objectives, we employ a defined methodology. This involves designing and training a CNN model and evaluating its performance. Additionally, we compare the CNN approach with an SVM to identify their relative strengths and weaknesses in redshift estimation.

1.4 Innovation

The contribution of this work lies in advancing knowledge within the field of galaxy redshift estimation. By applying state-of-the-art machine learning techniques, we aim to provide an innovative approach to predicting galaxy redshifts, potentially improving accuracy and efficiency in astronomy.

1.5 Structure

The subsequent chapters of this dissertation are organized as follows:

1. **Introduction** (the current chapter): Provides an overview of the dissertation's focus, objectives, methodology, and innovation.
2. **Background**: Explores foundational concepts related to galaxy redshift estimation, including traditional methods and machine learning potential.
3. **Data Preparation**: Details the process of selecting and cleaning galaxy spectra data for model training.
4. **Methodology**: Elaborates on the methodologies used for CNN model training, evaluation, and comparison.
5. **Results and Discussion**: Presents the findings of our experiments, assesses model performance, and discusses implications.
6. **Conclusion**: Summarizes key takeaways and outlines potential areas for future research.

2 Background

2.1 Gaia space 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 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 galaxy, along with any alterations therein.

During its mission, Gaia conducts approximately 14 observations each year for its designated stars. Its primary goals include accurately mapping the positions, distances, motions, and brightness variations of stars. Gaia's mission is expected to uncover various celestial objects, including exoplanets and brown dwarfs, and thoroughly study hundreds of thousands of asteroids within our Solar System. Additionally, the mission involves studying over 1 million distant quasars and conducting new assessments of Albert Einstein's General Theory of Relativity.

2.2 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:

$$z = \frac{\lambda_{obsv} - \lambda_{emit}}{\lambda_{emit}}$$

where λ_{obsv} is the observed wavelength of light from the galaxy, and λ_{emit} 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 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.

2.3 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 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. In our case, instead of class probabilities, the output of the final fully connected layer will yield a single numeric value representing the predicted redshift of the observed galaxy.

2.4 Support Vector Machines

Support Vector Machines (SVMs) represent another powerful tool in the realm of machine learning, particularly suited for classification and regression tasks. Just as CNNs have found success in image and video analysis, SVMs have established themselves as effective tools in various domains, including text classification, image recognition, and bioinformatics.

SVMs draw inspiration from the concept of finding an optimal hyperplane to separate data points into different classes. This mathematical approach is rooted in maximizing the margin, or the distance between the hyperplane and the nearest

data points from each class. These closest data points are referred to as support vectors, forming the basis for the name "Support Vector Machines."

In contrast to CNNs, SVMs do not explicitly mimic the structure of biological neurons. Instead, they operate by transforming data into a high-dimensional feature space where the data becomes more linearly separable. This transformation is facilitated by kernel functions, which define the mapping of input data to the higher-dimensional space.