Comparing photometric redshift to other methods of measuring extragalactic distances

How does photometric redshift work and how does it quantitively compare to measuring distance through the Cepheid period-luminosity relation and type Ia supernovae?

IB Physics Extended Essay

3997 words

Contents

Introduction

From ancient times, human have looked up the skies and wondered if one day we can reach the stars. This has motivated humankind to measure distances to moons and stars. However, modern astronomers are motivated by another reason to measure cosmic distances: to understand the fundamental properties of the universe. In 1929, Edwin Hubble discovered the universe is expanding by observing Cepheid variables, leading to the Hubble constant. By measuring vast extragalactic distances more and more accurately, astronomers are striving for more accurate value of the Hubble constant. This allows us to discover more about the past, present and future of the whole universe.

Finally, machine learning is a rapidly growing field because of the ever-growing amount of data and the ever-growing need of efficiently processing such data. In astronomy, data collected from surveys and telescopes has increased rapidly in the past few decades, and astronomers have set their eyes on machine learning to drive their data-driven quest on understanding the universe.

In this essay, I will train a neural network that performs photometric redshift, and quantitively compare it with other methods of measuring vast distances, including the period-luminosity relation of Cepheid variable stars, and the magnitude of type Ia supernovae by using a statistical metric. My research question is:

> How does photometric redshift work and how does it quantitively compare to measuring distance through the Cepheid period-luminosity relation and type Ia supernovae?

Background

Magnitude system

Firstly, let's introduce the magnitude system, which is a logarithmic measure of brightness. There are two kinds of magnitudes used for stars and galaxies: the apparent magnitude m is the brightness as it appears in the sky as observed on Earth, while the absolute magnitude M is the intrinsic brightness of an object, defined as the apparent magnitude of the object if it was placed 10 parsecs away from Earth. Note that the word 'magnitude' already refers to the brightness in the field of astronomy. The magnitude system is defined as follows: an increase in 5 magnitudes correspond to 100 times the brightness.

$$m_1 - m_2 = -5\log_{100}\frac{F_1}{F_2} = -5 \times \left(\log_{100}\frac{F_1}{F_2} \div \log_{100} 10\right) \times \log_{100} 10 = -2.5\log_{10}\frac{F_1}{F_2}$$

where m_i are the magnitudes, and F_i are the fluxes of the objects, which is the total amount of light energy intercepted by the detector divided by the area of the detector, measured in the unit $W\ m^{-2}$. This is not to be confused with luminosity, which is the total amount of light energy produced by the object per second, in unit W.

Traditionally, the zero-point of the magnitude scale is calibrated as the brightness of the star Vega, but other systems may have different zero-points. Moreover, the scale works 'in reverse', so brighter objects have a smaller magnitude. As the brightness scale has no upper or lower limit, negative magnitude are also possible, such as Sirius, which has an apparent magnitude of -1.46. Finally, if there is a subscript, for example M_v, that indicates the absolute magnitude measure in the visual V broadband filter.

Redshift

Another concept that has to be introduced is cosmological redshift. This is a measure of distance that deals with intergalactic distances involving galaxies, supernovae and quasars. Cosmological redshift is the phenomenon in which light emitted from extremely faraway objects are stretched while travelling through the expanding universe, like a drawn arrow on a stretching balloon. Light shifts to the red side of the spectrum as it is stretched, hence the name. The cosmological redshift z is defined mathematically as:

$$z = \frac{\lambda_{obs} - \lambda_{em}}{\lambda_{em}} = \frac{\lambda_{obs}}{\lambda_{em}} - 1$$

where λ_{obs} is the wavelength of light received by the observer and λ_{em} is the wavelength of light emitted by the source. Since it is based on a ratio between wavelengths, it is dimensionless.

The 2 common methods for comparison (Cepheid period-luminosity relation, type Ia supernovae) both measure the distance by observing the apparent magnitude of some celestial objects, obtaining the absolute magnitude of it through some properties, then comparing the 2 quantities to obtain the distance or redshift.

If we have both the apparent and absolute magnitude of an object, we can calculate its luminosity distance D_L by the flux-luminosity relationship (Gabrielli, P.377):

$$F = \frac{L}{4\pi D_L^2}$$

$$M - m = -2.5\log_{10}\frac{F_M}{F_m} = -2.5\log_{10}\left(\frac{L}{4\pi \times 10^2} \div \frac{L}{4\pi \times D_L^2}\right) = -2.5\log_{10}\left(\frac{D_L}{10}\right)^2 = -5\log_{10}\frac{D_L}{10}$$

$$\therefore D_L = 10^{\frac{m-M}{5}+1}$$

However, D_L is only approximate and fails at large distances due to the apparent magnitude being affected by the enlargement of space over the travel time of the electromagnetic radiation. To correct for this, the luminosity distance D_L is converted to redshift z.

However, only the equation to convert z into D_L was found (Pettini p.3):

$$D_L = \frac{c(1+z)}{H_0} \times \int_0^z \frac{dz}{\sqrt{\Omega_{m,0} \times (1+z)^3 + \Omega_{\Lambda,0}}}$$

where c is the speed of light, H_0 is the Hubble constant at current time in km s^{-1} Mpc^{-1}, The unitless Ω terms are the density parameters of our universe, and they sum up to 1 if our universe is flat (C.R. Nave). $\Omega_{m,0}$ is the mass density including baryonic mass and dark matter at current time, and $\Omega_{\Lambda,0}$ is the effective mass density of dark energy at current time. This equation assumes a flat spacetime, which observations closely matches, and neglects the radiation density at current time, which is around the magnitude of 10^{-5} (C.R. Nave). The values of $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$ are 0.3089 ± 0.0062 and 0.6911 ± 0.0062 calculated by cosmic microwave background observations from the Planck satellite (Ade, P.A. et al., p.32 last column).

To convert luminosity distance D_L into to redshift z, a technique called binary search was used, which is explained in appendix A. See figure 1 for a plot of redshift against luminosity distance.

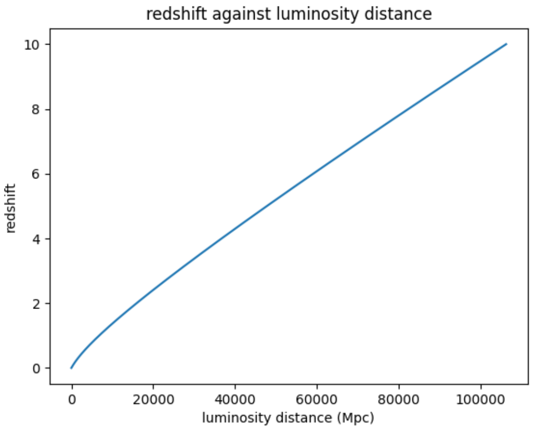


Figure 1

Photometric redshift

Spectroscopy

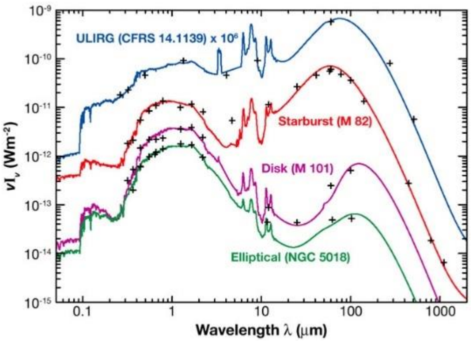


Figure 2: Examples of spectral energy distributions of galaxies from Frédéric Galliano. Étude Multi-Longueurs d'Onde de Galaxies Naines Proches: Propriétés des Milieux Interstellaires de Faible Métallicité. Astrophysique [astro-ph]. Université Paris Sud - Paris XI, 2004.

Before discussing photometric redshift, redshift obtained from spectroscopy (spectroscopic redshift) must be addressed as both methods are similar and spectroscopic redshift is also needed for calibration of photometric redshift. Spectroscopy is the study of electromagnetic radiation in terms of its constituent wavelengths. In a spectrograph, light is passed through a slit mask and dispersed by diffraction gratings, obtaining the spectral energy distribution (SED) of celestial objects. SEDs of 4 different types of galaxies are shown in figure 2, each with different unique features.

The x-axis corresponds to the wavelength, which is usually measured in μm or Angstrom (Å), where 1 Angstrom is equal to 10^{-10} meters. The y-axis correspond to the flux in different wavelengths of light emitted by an object. In this essay, data from SDSS (Sloan Digital Sky Survey) is used, in which flux is measured in nanomaggies. A star of 1 nanomaggie has an apparent brightness magnitude of 22.5, and is related to the apparent brightness as follows (*Measures of flux and magnitude*):

$$m = 22.5 - 2.5 \log_{10} f$$

where m is the apparent brightness and f is flux in nanomaggies.

Redshift (z) is the elongation of the wavelength of electromagnetic radiation of a faraway object due to the expansion of universe as mentioned previously. Therefore, the measured SED is shifted to the right compared to the actual SED (see figure 3).

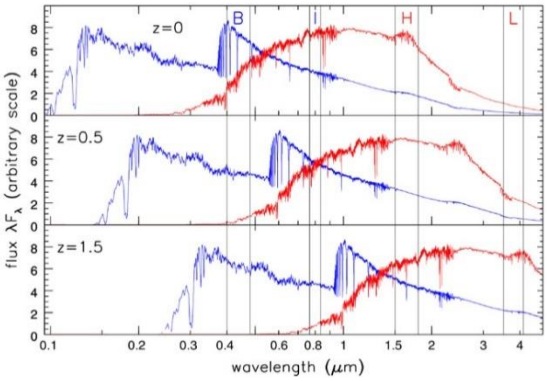


Figure 3: different redshifts of 2 SEDs, S. Charlot 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

To find out how much the spectrum is shifted, astronomers look for characteristic peaks and dips corresponding to emission or absorption lines in the spectrum. Once the corresponding features in the redshifted spectrum is identified, the redshift can be easily calculated from the previously mentioned formula $z = \frac{\lambda_{obs}-\lambda_{em}}{\lambda_{em}}$.

In galaxies, emission lines are often found in star-forming regions, where photons were emitted due to the excited atoms going from a high energy state to a low energy state. Each element has different energy levels, in which electrons that goes to a lower energy state emit photons of different wavelength ($E = hf$), so they can be used as characteristic peaks. On the other hand, absorption lines (dips in the SED) are formed when radiation from the center of a galaxy is absorbed by the gas clouds.

Photometry

Spectroscopic redshifts of individual galaxies are important, but it is severely insufficient for modern astronomy because spectroscopy is a very time-consuming and expensive process as the spectrograph must be precisely positioned to obtain a SED in 1 single point in the sky. In order to produce redshifts for ideally large amount of objects more efficiently in our universe, the concept of photometric redshifts was born.

In photometry, the flux from large chunks of the sky is captured at once, making it much more efficient than spectroscopy. CCD (charge-coupled devices) cameras are used, which is essentially a grid of CCD photometers (devices that count incoming photons). Typically, different light filters are used in combination to provide more information. In SDSS, 5 broadband filters u', g', r', i', z' are used, capturing the total flux in different sections of the SED, producing 5 images for each chunk of the sky. There are also other filter systems, such as the traditional UBV system.

The 5 numbers of flux can be thought of as a SED with extremely low resolution, and the method of photometric redshift is trying to estimate the redshift from these values.

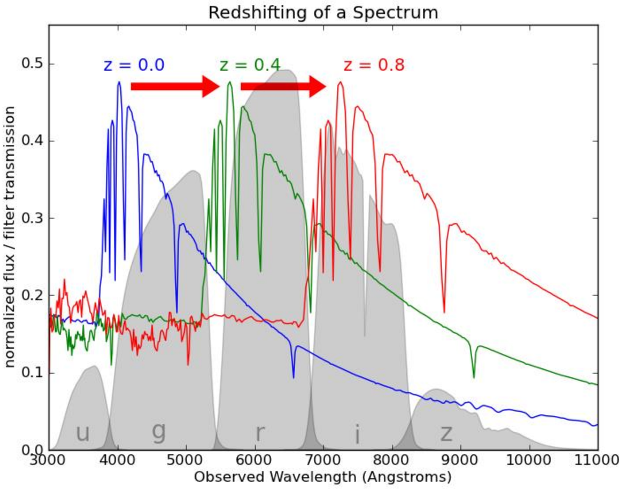


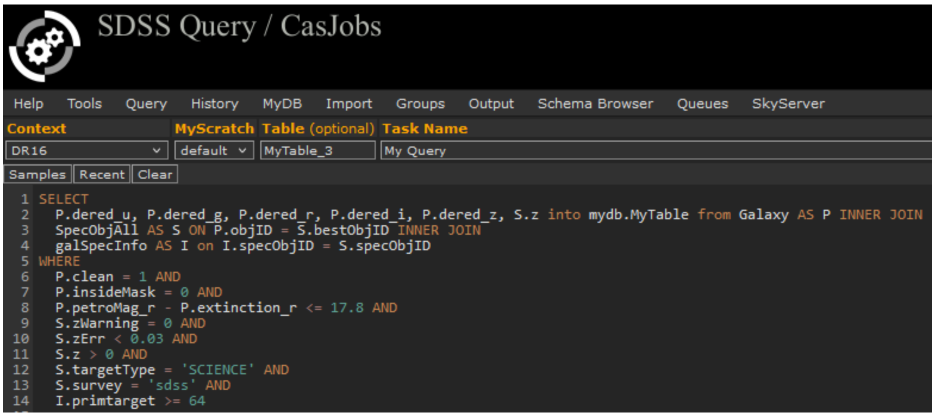
Figure 4: redshifting with ugriz filters, 2.3.5. Regression: Photometric Redshifts of Galaxies https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/regression.html

When photometric redshift were first investigated, statistical methods such as template fitting (Loh and Spillar) and quadratic regression (Connolly et al.) were used to empirically find the redshift. Nowadays, astronomers are using the new technique of machine learning to effectively let the machine 'learn' the pattern between the input (flux from broadband filters) and the output (redshift), which is also efficient for processing large amount of data.

Galaxy selection

The machine learning technique of neural network will be used as it can learn non-linear relationship between the input and output. However, a good dataset must first be chosen to train a neural network with good results. `dered_u`, `dered_g`, `dered_r`, `dered_i`, `dered_z` from the photometry database are chosen as inputs, which are the model fit magnitudes subtracted by extinction, which is the absorption or scattering of light by dust and gas in the interstellar medium between the celestial object observed and the observer. On the other hand, `z` from the spectroscopy database is chosen as the true output, as spectroscopic redshift is more accurate. The galaxies are chosen from SDSS DR16 (Data Release 16) based on the following criteria:

- the photometry and spectroscopy are clean and has no errors
- the dereddened Petrosian magnitude is ≤ 17.8, which reduces fluctuations of the dataset (Strauss et al., p.2)
- the galaxy is observed by SDSS (other surveys are also included in the database)



Context: DR16
MyScratch: default
Table (optional): MyTable_3
Task Name: My Query

```sql
SELECT
  P.dered_u, P.dered_g, P.dered_r, P.dered_i, P.dered_z, S.z into mydb.MyTable from Galaxy AS P INNER JOIN
  SpecObjAll AS S ON P.objID = S.bestObjID INNER JOIN
  galSpecInfo AS I on I.specObjID = S.specObjID
WHERE
  P.clean = 1 AND
  P.insideMask = 0 AND
  P.petroMag_r - P.extinction_r <= 17.8 AND
  S.zWarning = 0 AND
  S.zErr < 0.03 AND
  S.z > 0 AND
  S.targetType = 'SCIENCE' AND
  S.survey = 'sdss' AND
  I.primtarget >= 64
```

Figure 5

All datasets, including those for the other two methods, are in Appendix B.

Neural network

Generally, a machine learning method is first trained by a training dataset, where it tries to improve itself each iteration by comparing its result with the model output of the training dataset. After the learning process, it is evaluated by testing it on a completely different dataset, called the testing dataset, to actually grade if it learned the relationship between input and output or it just learned the training dataset. The downloaded dataset of 551617 galaxies are divided into the training set and test set by the ratio 3：1. The training set is scaled and shifted such that its mean is 0 and its standard deviation is 1, as the learning model employed assumes so. The test set is transformed the same way as the training set, but its mean might not be 0.

After trying different machine learning algorithms including decision trees and k-nearest neighbours, neural network was chosen for its superior performance. A neural network consists of an input layer, an output layer and 1 or more hidden layers. Each edge represents a weight, and a layer is transformed to another by summing up the value of node i in the previous layer times the weight of edge from i to the current node, similar to vector-matrix multiplication. The hidden layers allow the network to learn non-linear relationships as it is a combination of multiple vector-matrix multiplication. The neural network tries to minimize the loss function, which measures how different its output and the model output is. In each iteration, all the weights are slightly changed according to the neural network to gradually minimize the loss (Ross).

A neural network also have a lot of hyperparameters to tune, which are parameters of the network itself, such as its learning rate, hidden layer sizes, activation function etc. If the hyperparameters are set incorrectly, the neural network may overfit, where it 'learns' too much about the training dataset and is unable to generalize to other data, or it may underfit, where it does not learn the relationship enough. I tried many different values and combinations of the hyperparameters to produce the best possible result. The below diagram demonstrates underfitting and overfitting in a simple 1-input (x-axis) 1-output (y-axis) machine learning method.

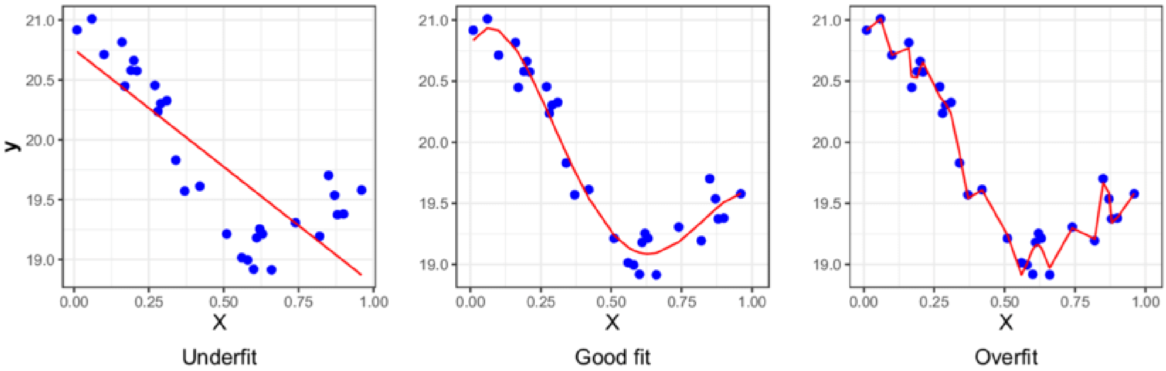


Figure 6: demonstration of underfit and overfit, Badillo, Solveig, et al. "An Introduction to Machine Learning." Clinical Pharmacology & Therapeutics, vol. 107, no. 4, 2020, pp. 871–885., https://doi.org/10.1002/cpt.1796.

The resulting neural network is implemented by `scikit-learn` (Pedregosa et al.), a Python library of machine learning algorithms. It uses the loss function of mean squared error, the activation function of the hyperbolic tangent function, and has a structure shown in figure 8.

The increased number of nodes in the first hidden layer allows the neural network to extrapolate relationships between the input, and the second hidden layer condenses the results into the redshift. Also, all the code of the neural network and other methods will be in Appendix C.

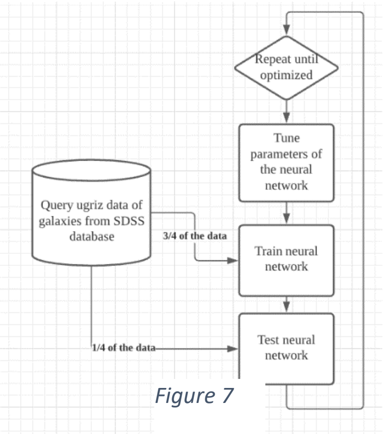


Figure 7

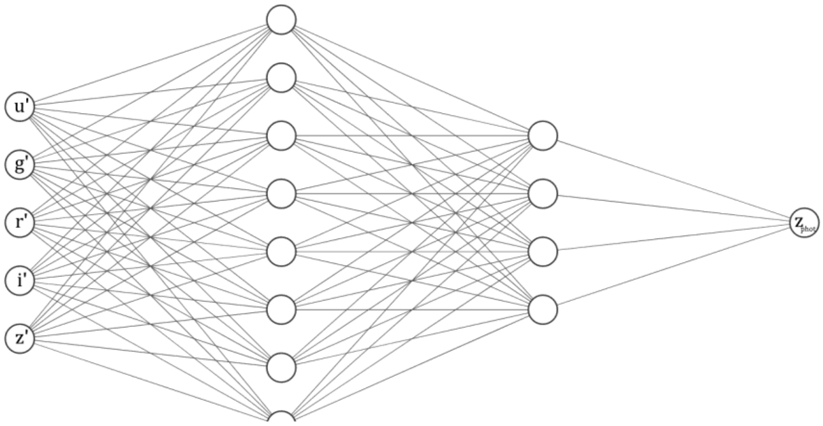


Figure 8: schematic of the neural network, generated from http://alexlenail.me/NN-SVG/index.html

It achieved mean squared loss of 0.000323 in the training dataset, and a mean squared loss of 0.000567 in the testing dataset. The scatterplot roughly follows the goal red line of predicted redshift = actual redshift with a few outliers. However, there is a significant failure in galaxies with redshift < 0.05, and this is most likely because there is the relation between the input flux and redshift is too small in nearby galaxies.

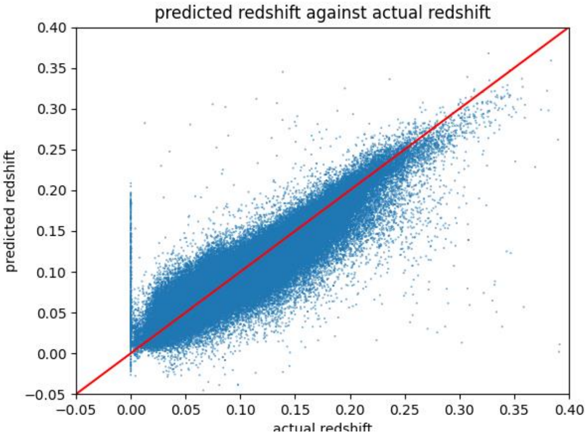


Figure 9: result of neural network on testing data

Comparison with other methods

Metric for comparison

For comparisons between photometric redshift by the neural network I created and each of the other methods, I am going to use the below statistics (Cohen et al., p.13):

For each data point $(z_{spec,i}, z_{pred,i})$ where $z_{spec,i}$ is the spectroscopic redshift of the i-th galaxy and assumed to be the ‘true’ value, and $z_{pred,i}$ is the predicted redshift of the i-th galaxy either by photometric redshift or other methods: the residuals $\Delta_{pred,i}$ is defined as the following:

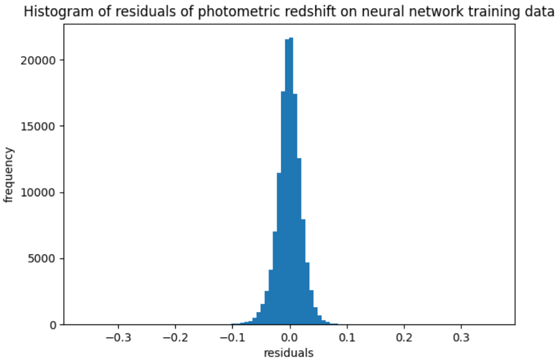
$$\Delta_{pred,i} = \frac{z_{spec,i} - z_{pred,i}}{1 + z_{spec,i}}$$

It is basically the difference between predicted and reference redshifts scaled by $(1 + z_{spec})$. This is to avoid errors in small redshifts artificially inflating the results

(Cohen et al., p.8). The following statistics derived from $\Delta_{pred,i}$ will be used for comparison:

- μ_{pred}, the mean of all $\Delta_{pred,i}$ with the same method used
- σ_{pred}, the standard deviation of all $\Delta_{pred,i}$ with the same method used
- η_{pred}, the fraction of galaxies with $\Delta_{pred,i} > 4\sigma_{pred}$, or the fraction of outliers

Applying this method on the photometric redshift neural network with all of its test samples (training data are excluded), the following statistics are obtained:



$$\mu_{phot} = -0.00083$$
$$\sigma_{phot} = 0.022$$
$$\eta_{phot} = 0.0018$$

Figure 10

According to this, the mean μ_{phot} is very close to 0, which indicates that there are little systematic error. Moreover, it is in the same order of magnitude with the results obtained from (Pasquet et al., p.7), which was 0.00010. The failure in small redshifts is probably the reason why it is negative (in that region z_{phot} is significantly larger than z_{pred}). The standard deviation and fraction of outliers also seems pretty low, but it must be compared with other methods to really convey significant meaning. Note that in the following comparisons, different datasets will be used as not all galaxies in the neural network testing data are compatible with the methods.

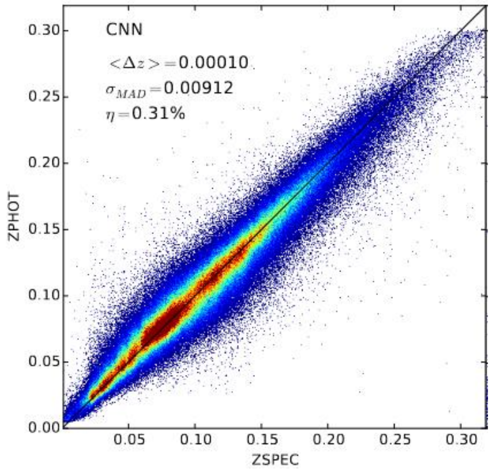


Figure 11: result of a sophisticated convolutional neural network (Pasquet et al., p.7)

Cepheid variable stars

Cepheid variable stars are a type of star that pulsates physically, changing in both diameter and brightness. In 1908, astronomer Henrietta Swan Leavitt discovered a relationship between the period and luminosity of Cepheid variables when investigating and cataloguing thousands of variable stars in the Small Magellanic Cloud and the Large Magallanic Cloud (Leavitt). This is one of the most earliest and historically important method to measure cosmological distances. In 1924, Edwin Hubble showed that the Cepheid variables in the Andromeda Galaxy were not in our own, establishing that there are other galaxies present in the universe (Hubble). Here is a typical light curve:

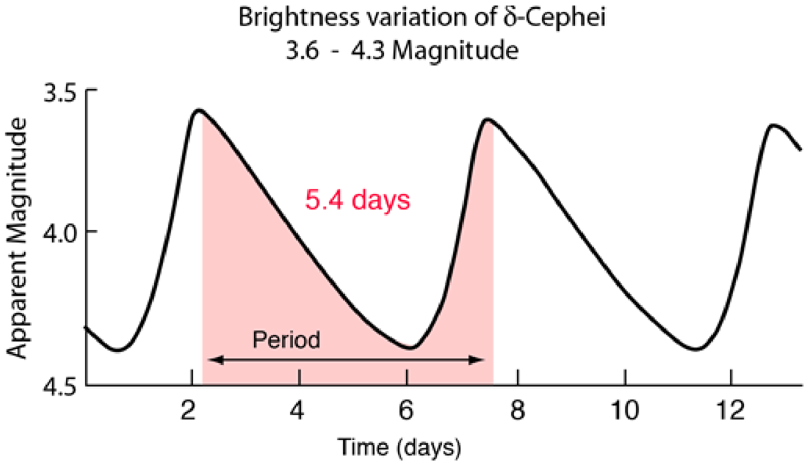


Figure 12: luminosity cycle of a cepheid star, C.R., Nave. "Delta Cephei." Cepheid Variable Stars, Georgia State University Department of Physics and Astronomy, http://hyperphysics.phy-astr.gsu.edu/hbase/Astro/cepheid.html.

Cepheid variables pulsates due to the kappa mechanism, where kappa represents the radiative opacity at a particular depth of the stellar atmosphere. In a normal star, opacity decreases upon compression, allowing radiation to escape more rapidly and maintaining hydrostatic equilibrium. However, opacity increases upon compression in Cepheid variable stars, so heat flow is blocked and causes a build-up of pressure that pushes the layer back out again (Baker). This results in a cyclic process in which the layer repeatedly moves inward and then is forced back out again, like steam leaking from a boiling kettle.

The period-luminosity relation is calibrated as follows (Freedman et al., p.6):

$$M_V = (-2.760 \pm 0.03)(\log_{10} P - 1) - (4.218 \pm 0.02)$$

$$M_I = (-2.962 \pm 0.02)(\log_{10} P - 1) - (4.904 \pm 0.01)$$

where P is the period in days, M_V is the absolute magnitude in visual V filter and M_I is the absolute magnitude in infrared I filter.

To account for interstellar extinction, the true distance modulus ($\mu = m - M$) is calculated (Freedman et al., p.6):

$$\mu_0 = \mu_V - R(\mu_V - \mu_I) = m_V - M_V - R(m_V - M_V - m_I + M_I)$$

where R is the extinction factor. $R(\mu_V - \mu_I)$ is also known as colour excess. This can be seen as correcting the observed magnitude to also account for light absorbed by the interstellar, so μ_0 still represents the difference between absolute and apparent magnitudes. Referring back to the background section,

$$D_L = 10^{\frac{m-M}{5}+1} = 10^{\frac{\mu_0}{5}+1}$$

Finally, to obtain the corresponding redshift, the method explained in the background section is used.

19 galaxies which had corresponding SDSS spectrographic redshift data were chosen from the 31 galaxies in the paper. The spectroscopic distances ('actual redshift' in the diagrams) were obtained from the SIMBAD astronomical database.

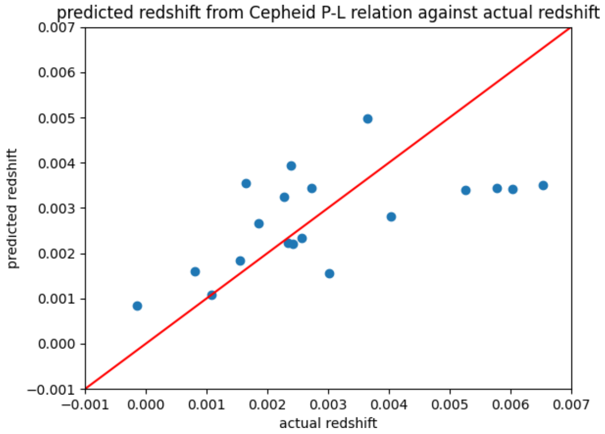


Figure 14

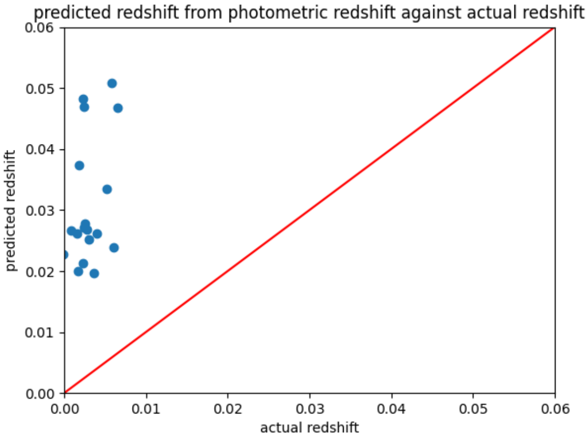


Figure 13

The distance derived from Cepheid variables are accurate, while those produced by the neural network is completely off the mark. This is expected as the neural network performed poorly in small redshifts, and Cepheids can only be observed from close distances. Applying the metric of residuals, the following data are obtained:

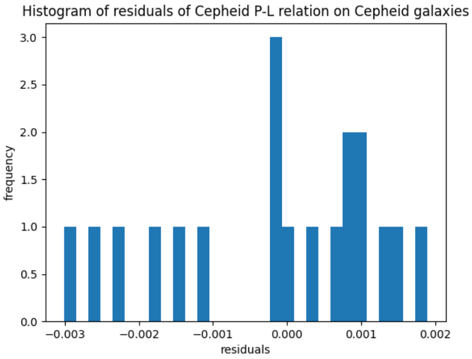


Figure 16

$$\mu_{ceph} = -0.00020$$
$$\sigma_{ceph} = 0.014$$
$$\eta_{ceph} = 0$$

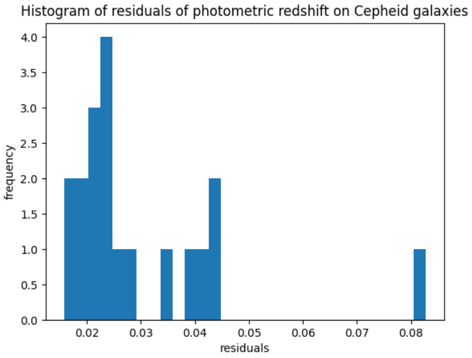


Figure 15

$$\mu_{phot} = 0.030$$
$$\sigma_{phot} = 0.015$$
$$\eta_{phot} = 0.053$$

μ_{phot} is 2 orders of magnitude larger than μ_{ceph}, indicating the large error. Interestingly, their standard deviations are similar, so they offer the same level of precision, or random error.

Type Ia supernovae

Supernovae are cataclysmic nuclear explosions occurring at the death of stars, and they are very luminous objects able to outshine an entire galaxy at its peak. A type Ia supernova is a subclass of supernovae which does not contain hydrogen, and presents a singly ionized silicon (Si ii) line at 615 nm near peak light, so it can be identified easily in spectroscopy and its redshift can be found (Sasdelli, p.2). Type Ia supernovae are formed in binary star-white dwarf systems, where the gas from the star is transferred to the white dwarf due to gravity, the white dwarf accretes too much material and explodes as it reaches the Chandrashekar limit of 1.4 solar masses, the maximum stable mass of a white dwarf star (*Introduction to supernova remnants*).

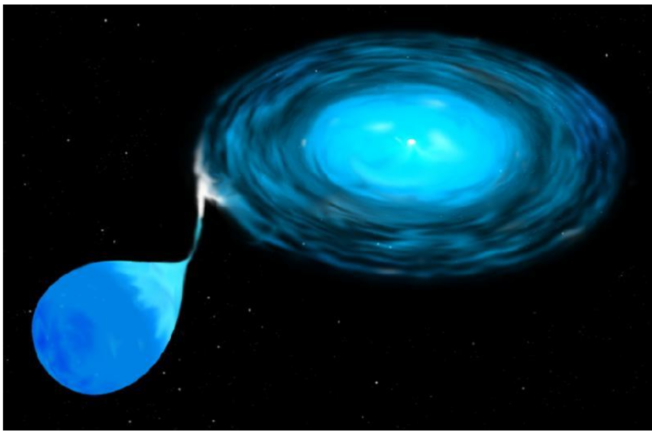


Figure 17: artistic depiction of white dwarf formation, "Hubble Probes the Workings of a Stellar Hydrogen Bomb." HubbleSite.org, STSci, https://hubblesite.org/contents/news-releases/1995/news-1995-23.html.

Since all type Ia supernovae explode at almost the same conditions, it is supposed that the differences in peak luminosities of type Ia supernovae are correlated with how quickly their light curves decline after maximum light. The peak of the light curve of all type Ia supernova reaches a consistent absolute magnitude, therefore they are said to be standard candles (a known standard luminosity).

Since the absolute luminosity of type Ia supernovae is known, its distance can be calculated by measuring its apparent brightness using the method discussed in the background section. The absolute magnitude of type Ia supernovae was chosen to be -19.3 according to the most updated literature value (Hillebrandt and Niemeyer, p.4).

The type Ia supernovae data (apparent and absolute magnitudes, colour excess) are from the Open Supernovae Catalog. It is then filtered such that only supernovae whose host galaxy have SDSS data (u' g' r' i' z' flux, spectroscopic redshift) are left. The final dataset comprises of 268 data points.

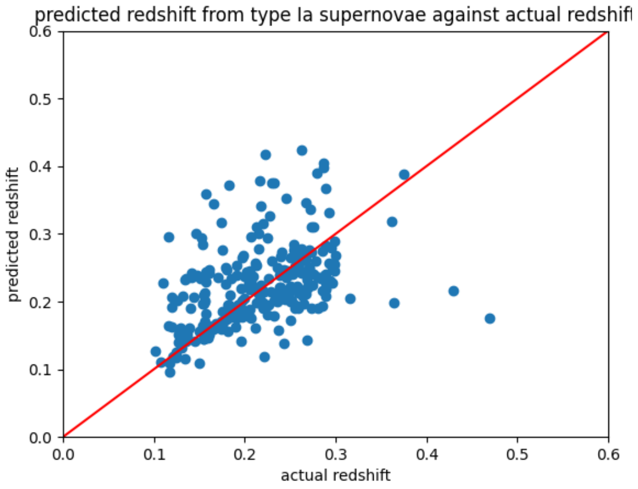


Figure 19

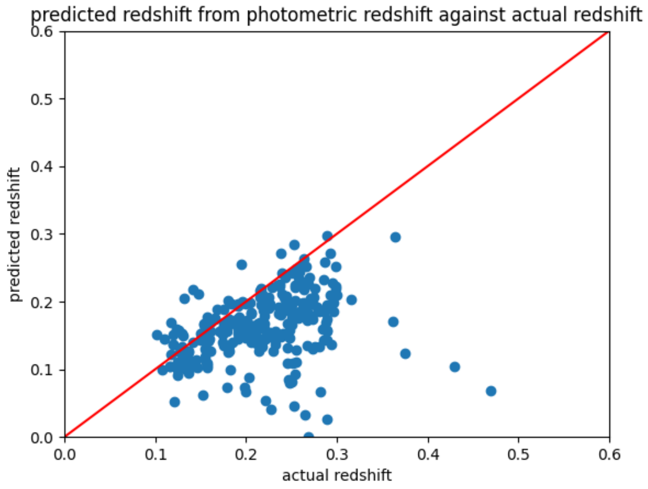


Figure 18

Examining the scatter plots, it is immediately obvious that the photometric redshift performed much better than when comparing with Cepheid method. It is also interesting how the supernovae plot seems skewed towards over-predicting the redshift, while the photometric redshift plot is skewed towards under-predicting the redshift. Both method also seem to struggle with the 2 supernovae with the highest spectroscopic ('actual') redshift, so these might be unreliable data. It is not immediately obvious which method performs better, so the previously discussed metric with residuals are used:

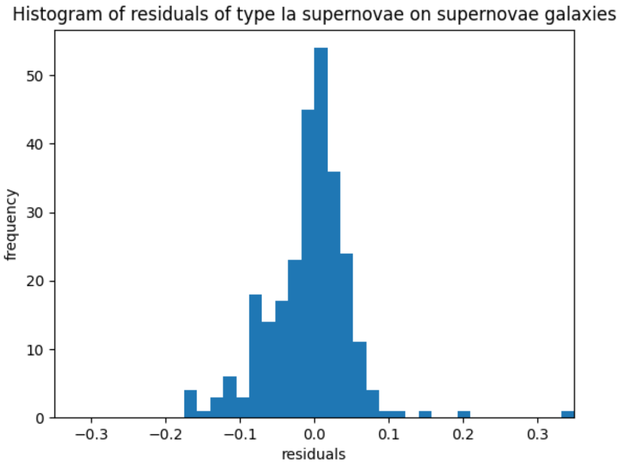


Figure 21

$$\mu_{sn} = -0.0067$$
$$\sigma_{sn} = 0.056$$
$$\eta_{sn} = 0.037$$

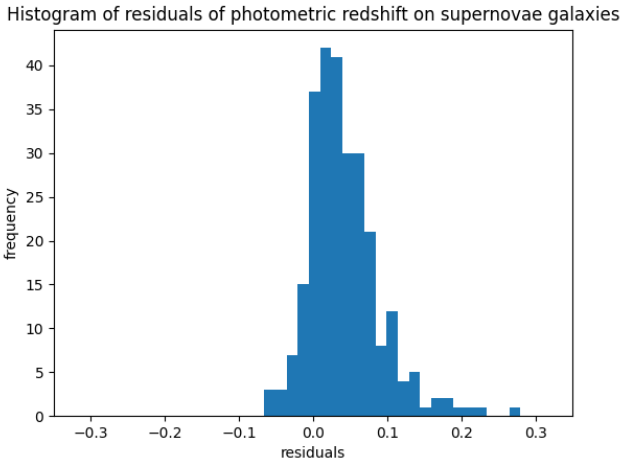


Figure 20

$$\mu_{phot} = 0.043$$
$$\sigma_{phot} = 0.052$$
$$\eta_{phot} = 0.015$$

Again, the histograms show the opposite skewness in the predicted data. This is also reflected in the means: μ_{sn} is negative while μ_{phot} is positive. Also, the magnitude of μ_{sn} is smaller than that of μ_{phot}, indicating that photometric redshift has a higher prediction bias. Surprisingly, the standard deviation σ_{phot} is lower than σ_{sn}, so my neural network outperformed type Ia supernovae in the sense that the predicted value are more clustered around the real value (spectroscopic redshift). My neural network also has a lower fraction of outliers than the supernovae method, which means there are less data points that the network predicts very incorrectly in the same 268 data points.

Conclusion

In this essay I have trained my own neural network that produces photometric redshift. I then compared it to other methods of measuring distances to other galaxies by picking a galaxy set for each comparison and using that set to test the neural network and the other one.

For galaxies with small redshift, the period-luminosity relation of Cepheid variables performs considerably better than the photometric redshift. However, photometric redshift still has a few advantages compared to Cepheid variables: it can apply to all types of galaxies and it can find the redshift of faraway galaxies. Comparing photometric redshift to type Ia supernovae, it performed slightly better in random error, but worse in systematic error. However, this still shows the immense potential of machine learning in that a neural network setup by an amateur can rival techniques that were cutting-edge in the past

My neural network is far from perfect, and many optimizations can be made, such as considering galaxy morphology as inputs, or even using images in the 5 filters as inputs, as shown on the right. However, I am satisfied as it demonstrates the basic concept of photometric redshift, and successfully captures the trend. Another extension is to compare more methods of measuring extragalatic distances such as the Tully-Fisher relation in spiral galaxies and the Faber-Jackson Relation in elliptical galaxies. Finally, the metric can be expanded upon, such as with MAD (median absolute deviation) and cross-validation for a deeper statistical comparison.

One major evaluation point is the data used in comparisons. While preparing the datasets for Cepheid and supernovae comparison, I overlooked an important detail: whether if the data overlap with the training data of the photometric redshift network. It is possible that some of the data for the comparisons, which is the testing data for the network, overlaps with the training data. Testing and training data should never overlap as that means the network already 'knows' the answer and have a slight bias towards the correct value. This may be a reason why my network performed well in the supernovae dataset: it may have already 'seen' supernovae galaxies in training, which undermines the reliability of the result. However, it was too late when I realized this and it was very difficult to filter the comparison data. Filtering it to check if it overlaps with the network training data will definitely improve the reliability of the results.

Finally, there is a lot of uncertainty I was not able to explore in this essay. A major reason is that I decided to focus on a machine learning method that was basically a black box, so errors couldn't be propagated from input to output, and only the distribution of the output could be analyzed directly. However, I still included uncertainty whenever possible to reflect the precisions of measurements / estimations.

In reality, all these methods come together to form the cosmic distance ladder, ones suitable for measuring shorter distances calibrating other methods for longer ranges. Through this investigation, I hope that I have shown the promising future of machine learning in photometric redshift as another concrete method in the cosmic distance ladder, advancing astrophysics with more and more precise measurements.

Bibliography

Ade, P. A., et al. “PLANCK2015 Results.” *Astronomy & Astrophysics*, vol. 594, 2016, https://doi.org/10.1051/0004-6361/201525830.

Baker, N., and R. Kippenhahn. “The Pulsations of Models of δ Cephei Stars.” *The Astronomical Journal*, vol. 66, 1961, p. 278., https://doi.org/10.1086/108542.

C.R., Nave. “Density Parameter, Ω.” *Density Parameter, Omega*, Georgia State University Department of Physics and Astronomy, http://hyperphysics.phy-astr.gsu.edu/hbase/Astro/denpar.html.

Cohen, Judith G., et al. “Caltech Faint Galaxy Redshift Survey. x. A Redshift Survey in the Region of the Hubble Deep Field North.” *The Astrophysical Journal*, vol. 538, no. 1, 2000, pp. 29–52., https://doi.org/10.1086/309096.

Connolly, A. J., et al. “Slicing through Multicolor Space: Galaxy Redshifts from Broadband Photometry.” *The Astronomical Journal*, vol. 110, 1995, p. 2655., https://doi.org/10.1086/117720.

Freedman, Wendy L., et al. “Final Results from Thehubble Space Telescopekey Project to Measure the Hubble Constant.” *The Astrophysical Journal*, vol. 553, no. 1, 2001, pp. 47–72., https://doi.org/10.1086/320638.

Gabrielli, A. “C.1 Cosmological Parameters.” *Statistical Physics for Cosmic Structures*, Springer, Berlin, 2005.

Hillebrandt, Wolfgang, and Jens C. Niemeyer. “Type Ia Supernova Explosion Models.” *Annual Review of Astronomy and Astrophysics*, vol. 38, no. 1, 2000, pp. 191–230., https://doi.org/10.1146/annurev.astro.38.1.191.

Hubble, E. P. "Cepheids in spiral nebulae". *The Observatory*. 48: 139. Bibcode:1925Obs....48..139H

“Introduction to Supernova Remnants.” *NASA*, NASA, https://heasarc.gsfc.nasa.gov/docs/objects/snrs/snrstext.html.

Leavitt, Henrietta S.; Pickering, Edward C. "Periods of 25 variable stars in the Small Magellanic Cloud". *Harvard College Observatory Circular*. 173: 1–3. Bibcode:1912HarCi.173....1L

Loh, E. D., and E. J. Spillar. “Photometric Redshifts of Galaxies.” *The Astrophysical Journal*, vol. 303, 1986, p. 154., https://doi.org/10.1086/164062.

“Measures of Flux and Magnitude.” *SDSS*, SDSS-III, http://www.sdss3.org/dr8/algorithms/magnitudes.php#nmgy.

Pasquet, Johanna, et al. “Photometric Redshifts from SDSS Images Using a Convolutional Neural Network.” *Astronomy & Astrophysics*, vol. 621, 2018, https://doi.org/10.1051/0004-6361/201833617.

Pettini, Max. *6.1 Applications of the Luminosity Distance*. University of Cambridge Institute of Astronomy, https://people.ast.cam.ac.uk/~pettini/Intro%20Cosmology/Lecture06.pdf.

Ross, Matt. “Under the Hood of Neural Network Forward Propagation - the Dreaded Matrix Multiplication.” *Medium*, Towards Data Science, 21 Oct. 2017, https://towardsdatascience.com/under-the-hood-of-neural-network-forward-propagation-the-dreaded-matrix-multiplication-a5360b33426.

Sasdelli, Michele, et al. “Abundance Stratification in Type Ia Supernovae – IV. the Luminous, Peculiar SN 1991T.” *Monthly Notices of the Royal Astronomical Society,* vol. 445, no. 1, 2014, pp. 711–725., https://doi.org/10.1093/mnras/stu1777.

Strauss, Michael A., et al. “Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample.” *The Astronomical Journal*, vol. 124, no. 3, 2002, pp. 1810–1824., https://doi.org/10.1086/342343.

Appendix

Appendix A

We have a method f to convert z into D_L. Given a specific value of D_L, z can be found using binary search because the function is monotonic (as seen from figure 1).

Set $z_l \coloneqq 0$ and $z_r \coloneqq 10$. (It covers all the possible range our z can be in)

```
Repeat the following 20 times:
    $z_{mid} \coloneqq (z_l + z_r) \div 2$
    $D_{L,mid} \coloneqq f(z_{mid})$
    if $D_L > D_{L,mid}$, set $z_l = z_{mid}$; else set $z_r = z_{mid}$
```

At each iteration, the range (z_l, z_r) is halved. After 20 iterations, it is precise enough for our data.

Appendix B

All the data used:

https://www.dropbox.com/sh/ej931tkecnkearw/AAA2iYf2aw02RFozdOjob8Qaa?dl=0

The files are in different formats since they are all downloaded from different databases or processed by me. They can be viewed by simple text editors (Notepad, for example)

Databases used:

https://skyserver.sdss.org/casjobs/

http://simbad.u-strasbg.fr/simbad/

https://vizier.u-strasbg.fr/viz-bin/VizieR-3

https://sne.space/

Appendix C

Below is the Python code used. It requires `numpy`, `scikit-learn`, `matplotlib` and `scipy` libraries.

```
import numpy
data = numpy.genfromtxt("SDSS_galaxies.csv", delimiter = ",")
data = numpy.delete(data, (0), axis = 0)
x = data[:, 30:35]
z = data[:, 35]
n = len(x)
ntrain = int(3 * n / 4)
xtrain = x[:ntrain, :]
ztrain = z[:ntrain]
xtest = x[ntrain:n, :]
ztest = z[ntrain:n]
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(xtrain)
xtrain = scaler.transform(xtrain)
xtest = scaler.transform(xtest)
from sklearn.neural_network import MLPRegressor
from matplotlib import pyplot

rgs = MLPRegressor(hidden_layer_sizes = (8,4), activation = 'tanh', solver = 'adam', tol =
0.0000000001, learning_rate_init = 0.01, learning_rate = 'adaptive', random_state = 1, max_iter =
5000, verbose = 1).fit(xtrain, ztrain)

zpred = rgs.predict(xtrain)
sum((ztest-zpred)**2)/(n-ntrain)
pyplot.title("predicted redshift against actual redshift")
pyplot.xlabel("actual redshift")
pyplot.ylabel("predicted redshift")
s = [0.1] * len(ztrain)
pyplot.scatter(ztrain, zpred, s=s)
pyplot.plot(numpy.linspace(-1,1,1000),numpy.linspace(-1,1,1000),color='red')
pyplot.show()

data = numpy.genfromtxt("cepheid_galaxies.csv", delimiter = ",")
brh = data[:, :5]
brh = scaler.transform(brh);
```

```
zz = rgs.predict(brh)
pyplot.scatter(data[:,5], act)

act = numpy.genfromtxt("D:\G11 & G12\EE\cepzreal.csv", delimiter = ",")


pyplot.scatter(act,data[:,5])
pyplot.plot(numpy.linspace(-1,1,1000),numpy.linspace(-1,1,1000),color='red')
pyplot.xlim([-0.001, 0.007])
pyplot.ylim([-0.001, 0.007])
pyplot.title("predicted redshift from Cepheid P-L relation against actual redshift")
pyplot.xlabel("actual redshift")
pyplot.ylabel("predicted redshift")
pyplot.show()

d = (data[:,5]-act)/(1+data[:,5])
m = numpy.mean(d)
s = numpy.std(d)
count = 0
for i in range(len(act)):
  count += (d[i] > (4 * s))

o = count / len(d)
print(m,s,o)
pyplot.hist(d, 30)
pyplot.title('Histogram of residuals of Cepheid P-L relation on Cepheid galaxies')
pyplot.ylabel('frequency')
pyplot.xlabel('residuals')
pyplot.show()

pyplot.scatter(act,zz)
pyplot.plot(numpy.linspace(-1,1,1000),numpy.linspace(-1,1,1000),color='red')
pyplot.title("predicted redshift from photometric redshift against actual redshift")
pyplot.xlim([-0.00, 0.06])
pyplot.ylim([-0.00, 0.06])
pyplot.xlabel("actual redshift")
pyplot.ylabel("predicted redshift")
pyplot.show()

d = (zz-act)/(1+zz)
m = numpy.mean(d)
s = numpy.std(d)
count = 0
for i in range(len(act)):
  count += (d[i] > (4 * s))

o = count / len(d)
print(m,s,o)
pyplot.hist(d, 30)
pyplot.title('Histogram of residuals of photometric redshift on Cepheid galaxies')
pyplot.ylabel('frequency')
pyplot.xlabel('residuals')
pyplot.show()

asu = numpy.genfromtxt("asu2.tsv", delimiter = "|")
asu = numpy.delete(asu, (0), axis = 0)
asu = numpy.delete(asu, (0), axis = 0)
#asu = numpy.array(sorted(asu, key = lambda x: x[4]))
```

```
snx = numpy.zeros((len(asu), 5))
snx[:, 0] = asu[:, 7]
snx[:, 1] = asu[:, 9]
snx[:, 2] = asu[:, 11]
snx[:, 3] = asu[:, 13]
snx[:, 4] = asu[:, 15]

snx = scaler.transform(snx)
sny = asu[:, 17]

snpred = rgs.predict(snx)
pyplot.title("predicted redshift from photometric redshift against actual redshift")
pyplot.xlabel("actual redshift")
pyplot.ylabel("predicted redshift")
pyplot.xlim([0, 0.6])
pyplot.ylim([0, 0.6])
s = [0.1] * len(zpred)
pyplot.scatter(sny, snpred)
pyplot.plot(numpy.linspace(-1,1,1000),numpy.linspace(-1,1,1000),color='red')
pyplot.show()


d = (sny-snpred)/(1+sny)
m = numpy.mean(d)
s = numpy.std(d)
count = 0
for i in range(len(sny)):
  count += (d[i] > (4 * s))

o = count / len(sny)
print(m,s,o)
pyplot.hist(d, 30)
pyplot.title('Histogram of residuals of photometric redshift on supernovae galaxies')
pyplot.ylabel('frequency')
pyplot.xlabel('residuals')
pyplot.xlim([-0.35, 0.35])
pyplot.show()

asu = numpy.genfromtxt("asu2.tsv", delimiter = "|")
asu = numpy.delete(asu, (0), axis = 0)
asu = numpy.delete(asu, (0), axis = 0)
sny = asu[:, 17]
def calc_dl(z):
  H0 = 67.74
  WM = 0.3089
  WV = 0.6911
  WR = 0.
  WK = 0.
  c = 299792.458
  a = 1.0
  az = 0.5
  h = H0 / 100.
  WR = 2.47E-5 / (h * h)
  WK = 1 – WM – WR - WV
  az = 1.0 / (1 + 1.0 * z)
  age = 0.
  n=1000
    a = az + (1 - az) * (i + 0.5)/n
    adot = numpy.sqrt(WK + (WM / a)+(WR / (a * a))+(WV * a * a))
    DCMR = DCMR + 1 ./ (a*adot)
  DCMR = (1 .- az) * DCMR / n
  return (c/H0)* DCMR/az

def bs(dl):
  l = 0.0
  r = 10.0
  for i in range(20):
    mid = (l + r) / 2.0
    if calc_dl(mid) < dl:
      l = mid
    else:
      r = mid
  return l

osc = numpy.genfromtxt("sndata_w_sdss_data.txt", delimiter = " ")
z = numpy.zeros(len(osc))
for i in range(len(osc)):
  z[i] = bs(10 ** ((osc[i, 0] + 19.3 - osc[i, 2]) / 5.0 - 5.0))

pyplot.scatter(sny, z)
pyplot.plot(numpy.linspace(-1,1,1000),numpy.linspace(-1,1,1000),color='red')
pyplot.title("predicted redshift from type Ia supernovae against actual redshift")
pyplot.xlabel("actual redshift")
pyplot.ylabel("predicted redshift")
pyplot.xlim([0, 0.6])
pyplot.ylim([0, 0.6])
pyplot.show()

d = (sny-z)/(1+sny)
m = numpy.mean(d)
s = numpy.std(d)
count = 0
for i in range(len(sny)):
  count += (d[i] > (4 * s))

o = count / len(d)
print(m,s,o)
pyplot.hist(d, 30)
pyplot.title('Histogram of residuals of type Ia supernovae on supernovae galaxies')
pyplot.ylabel('frequency')
pyplot.xlabel('residuals')
pyplot.xlim([-0.35, 0.35])
pyplot.show()
```