MLD 2019 - Problem set 4

Introduction

The goal here is simply to get some experience with deep learning and optionally random forests.

1. Deep Learning on stellar spectra

This is quite straightforward.

Use the jupyter notebook (or convert it to a python code) in the Lecture 5/Notebook directory called Deep Learning of stellar spectra. This currently has a training MSE of 0.079 and a test MSE of 0.582.

Do any changes you want to the network and see how low a test MSE you can reach. Keep track of your results and make a plot of how they change - if you manage to always get a decreasing test score you do way better than I.

2. Photometric redshifts

The distance to a galaxy can be a challenging measurement to make. For most extra-galactic objects, we can not get true distances, and rely on redshifts. As long as the galaxy is sufficiently far away, this is very close to the true distance and we will here focus on redshifts.

Redshifts can best be determined using spectroscopy, but for large samples this requires too much observing time to be practical. For that reason many current and future surveys depend on the possibility to determine distances using photometry. These are known as photometric redshifts and they are the focus of this problem.

For our purposes this can be considered as a process where we learn a function that approximately predicts the redshifts of a galaxy when given parameters .

There are two files that come with the project in the directory Datafiles. These are memorably called PhotoZFileA.vot and PhotoZFileB.vot, I will refer to these as file A and file B respectively below.. These are VOtables and contain the following information:

```
Counter:  Just a running counter, starting at 1
mag_r: The r-band total magnitude of the galaxy.
u-g: the u-g colour of the galaxy
g-r: the g-r colour of the galaxy
r-i: the r-i colour of the galaxy
i-z: the i-z colour of the galaxy
z_spec: the spectroscopic redshift of the galaxy. We will take this as the
  true redshift.
```

The requirements of large future cosmological surveys are that $\left\langle \left| (z_{\text{phot}} - z_{\text{spec}})/(1 + z_{\text{spec}}) \right| \right\rangle < 0.01$ or even smaller. To be consistent with this, use

$$E(\theta) = \text{median}(\left| (z_{\text{spec}} - f(\theta))/(1 + z_{\text{spec}}) \right|)$$

as the measure of the discrepancy between photometric and spectroscopic redshifts.

a) The early work on photometric redshifts was done using linear regression. Design a regression estimator using either ridge, LASSO or linear regression to predict photometric redshifts, make sure to justify all your choices you make. Use the whole of file A for training and aim to obtain $E(\theta) < 0.01$ as training error.

b) Try next to create a random forest and/or neural network to predict the photometric redshifts of the objects. It should not be difficult to get an error below 0.01 but see how much you can get it below 0.001 and with a median absolute deviation (MAD) <0.02.

where MAD = median $\left(|x - \text{median}(x)| \right)$

Hint: to set up a RandomForest regression you can do for the basics (this is not a good one):

```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor(bootstrap=True, criterion='mse',
  n_estimators=10)
```

see the online documentation of RandomForestRegressor to modify this.