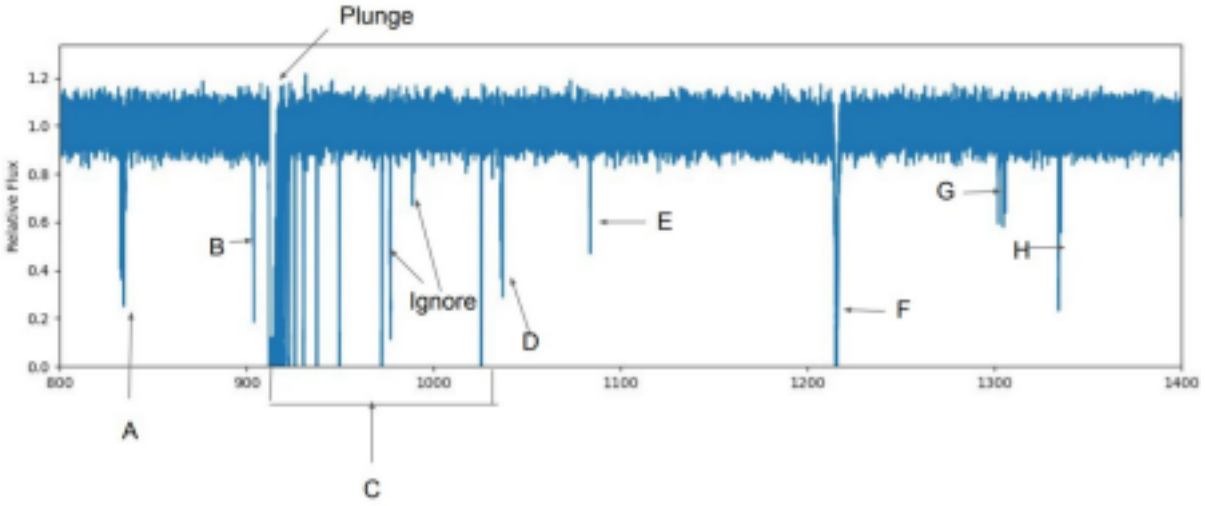
Dimensionality Reduction for Studying Diffuse Circumgalactic Medium Evaluation Test

Please work in a Python (Jupyter) notebook and submit it as part of your solution together with a pdf that shows all the outputs. Please find more submission instructions at the end.

Task 1: Ion Identification

Below is the normalized spectrum (plot of light intensity vs wavelength in Angstrom) of a gas cloud containing hydrogen, carbon, nitrogen, and oxygen. Each prominent line represents an absorption by one of these.



A. For features A-H, identify the atom responsible, and its ionic state. Assume there is no close grouping of lines from different elements. Use this source for spectral line data: https://physics.nist.gov/PhysRefData/Handbook/Tables/hydrogentable2.htm Just replace "hydrogen" with whatever element you want to check.

B. How are the lines in feature C related? And what causes the plunge near 900 Angstroms? You can explain in words, but reference an equation as well. (Hint: see https://en.wikipedia.org/wiki/Lyman_series)

Task 2: Astrophysical Absorption Line Exercise

Next task is to generate a spectrum of light after it has passed through a slab of gas, and investigate the Lyman α absorption line of hydrogen. The part below shows how you can calculate a spectrum and the necessary information to complete the 3 subtasks

Your task is to generate a **spectrum** (plot of light intensity vs wavelength) of light after passing through a slab of gas that absorbs some of the light. In particular, we are going to look at the Lyman α absorption line that is caused by hydrogen atoms in their ground state absorbing light that gives them energy jump up to their first excited state.

The intensity I of light of a given wavelength λ is given by this equation:

$$I(\lambda) = \exp\left[-\alpha(\nu)\, d\right] \tag{1}$$

where the frequency ν is related to λ by $\nu = c/\lambda$. The thickness of the slab is d. The absorption coefficient $\alpha(\nu)$ is given by the following equation:

$$\alpha(\nu) = \frac{e^2\, f\, n_{\rm H}}{4\pi m_e c} \frac{(1-x) g_0}{\mathcal{Z}} \frac{\Gamma}{(\nu-\nu_0)^2 + (\Gamma/4\pi)^2} \tag{2}$$

You can assume that the gas slab has the following properties:

$$n_{\rm H} = 0.1 \text{ cm}^{-3}$$
$$x = 0.1$$

You will need the following properties of hydrogen and the Lyman α transition:

$$\nu_0 = 2.46607 \times 10^{15} \text{ Hz}$$
$$\Gamma = 6.265 \times 10^{8} \text{ s}^{-1}$$
$$f = 0.4164$$
$$g_0 = 2$$
$$\mathcal{Z} = 2.00$$

Other physical constants:

$$m_e = 9.11 \times 10^{-28} \text{ g}$$
$$c = 3.00 \times 10^{10} \text{ cm/s}$$
$$e = 4.80 \times 10^{-10} \text{ cm}^{3/2}\ \text{g}^{1/2}\ \text{s}^{-1}$$

1. Generate the spectrum for the following different thicknesses of the slab (this is d in equation (1)):
   - 10^{14} cm
   - 10^{18} cm
   - 10^{21} cm

2. If the gas is moving towards or away from us, the central frequency ν_0 will be different. Generate all of those spectra from question 1 again but using $\nu_0 = 2.46632 \times 10^{15}$ Hz

3. Describe in words the differences between all of the 6 spectra you generated. Comment specifically on what properties look different, what properties look the same, and how different they are.

Task 3: Dimensionality Reduction Exercise

Here is the dataset:
https://drive.google.com/file/d/1Nl-RU5HggCSWWqINN4gbxHAVzER6T2Po/view?usp=sharing

Dataset Description: The dataset consists of 6 million labeled samples of two categories (classes) produced with Monte Carlo simulations. Each sample consists of 28 features. The first 21 features are basic features related to the degrees of freedom of the problem. The last seven are functions of the first 21. You can build as many additional features as you wish, but they need to be built from the original 21. For the purpose of this exercise, you can treat it as a "blackbox" classification problem with two classes.

A. Your first task is to build one or more **machine learning classifiers** to separate the two classes of events. Please provide the ROC curve that shows the classifier performance. We suggest that you evaluate the final performance of your classifier on the last 500k samples. **Be careful to not overfit the test dataset, as we will evaluate your model on another holdout dataset.**

B. Your next task is to apply one (or several) **machine learning dimensionality reduction** techniques of your choice to reduce the dimensionality of the problem as much as possible, without losing significant classification performance of your classifier. Please provide the ROC curve showing the performance of the reduced model and a set of features used by the model. **Be careful to not overfit the test dataset, as we will again evaluate your model on another holdout dataset.**

Test Submission Instructions: Please send us a link to all your completed work (github repo), Jupyter notebook + pdf of Jupyter notebook with output to ml4-sci@cern.ch with Evaluation Test: CGM in the title.