Applications of unsupervised machine learning techniques for data exploration and discovery in ISM science

Dalya BARON

(Stanford University, USA)

Hands-on application

After reviewing all the important aspects of applying unsupervised machine learning techniques in the lecture, we will now put this knowledge into practice, following the guidelines we discussed.

- You can work individually or in groups of up to four people, depending on how you learn best.
- You will choose a dataset to work with—this may include multi-wavelength data from PHANGS, PDRs4All, or another dataset to be determined. These datasets will be provided to you prior to the hands-on session.
- You will select one class of methods to explore: clustering, dimensionality reduction, or outlier detection. Within that class, you will identify up to two algorithms you'd like to try. A list of techniques for each class will be provided ahead of time, along with links to online resources and documentation for the functions you'll be using.
- It is highly recommended to work in Python, as all of the techniques have been implemented in it, and most use a fairly standard input-output structure. The easiest way to ensure that all the relevant packages are available is to install Anaconda. We recommend working in a new environment specifically defined for the school. Below are some online resources that explain the importance of using separate Anaconda environments, along with instructions for installing Anaconda and setting up your first environment (the first link includes all the steps you need):
  - https://www.geeksforgeeks.org/machine-learning/set-up-virtual-environment-for-python-using-anaconda/
  - https://www.anaconda.com/docs/getting-started/anaconda/install
  - https://www.anaconda.com/docs/tools/working-with-conda/environments
- You will examine two different approaches to representing the data: working with a set of physically derived features, and working directly with the raw data. By applying your selected techniques to both representations, you will compare their outputs and use the guidelines from the lecture to interpret the results.

- In addition to online documentation and Jupyter notebooks demonstrating the various algorithms—both of which will be provided—you will also have the option to interact with a Large Language Model (LLM), such as ChatGPT, Claude, Gemini, etc., to help you learn about the techniques and set up the initial code structure. LLMs are becoming a powerful resource that can significantly shorten the learning curve and reduce the time it takes to get a first version of the code running. If you choose to work with LLMs, we will discuss how to use them effectively and how to critically evaluate their outputs.