

This is the second project for this class. For this project, you will work in groups of 3 to synthesize the ideas we have discussed in the past few weeks in the context of the movie recommendation problem. You will make an iPython Notebook and submit it to Canvas. In this notebook, you will answer the following questions. **You should elaborate on each question as much as you can using markdown (text) cells!** You will be graded on your presentation and your responses.

You should use the notebooks `Classification_Week1.ipynb` and `Classification_Week2.ipynb` as guidance. You will need to copy over the cells in which `our_data` and `our_normalized_features` are defined. You may also find it useful to copy over some of the functions we defined in the notebook over the course of our week of work.

1. Make the names of every group member clearly indicated in the notebook, either in the title, filename, or a markdown cell at the start of the document.
2. Use the **k-nearest neighbors algorithm to predict the scores every student would give to every film that they haven't seen**. The final result of this problem should be an array similar to `our_data` that does not have *any* zero entries in it. (Note: in the case of a row with all zeros, feel free to replace one or two of the zeros with a random ranking. Or, you could ask a friend, family member, etc. for their rankings if you like!)

~~(a) You **must** comment upon the choice of k . What value of k did you use and why?~~ k-nearest ↑

~~(b) You may use the k-nearest neighbors algorithm to find the k most similar people to a given person and use that to make your predictions, or you may find the k most similar movies to a particular movie and use that for your predictions. If you do both and comment upon the differences, you will receive extra credit.~~ Movie prediction?

Hint: Suppose person 3 has not seen movie 5. We could update their score for that movie with this line:

```
our_data[3,5] = ...
```

where the right hand side calculates the predicted score person 3 will give to movie 5.

3. Use the k-means algorithm to mathematically identify some *clusters* of data. **You could find clusters of students with similar movie tastes, clusters of movies with similar features, or clusters of movies that have similar scores across the classroom.**
 - (a) You **must** comment upon the choice of k . What value of k did you use and why?
 - (b) Once you have found your clusters, look at which people or which movies are in each cluster. Explain in words what aspects are shared by objects in each cluster.
 - (c) Produce at least **one 2D plot and one 3D plot** in which the **different clusters** are represented by different color dots and the center of each cluster is represented by a colored star. Comment upon these plots.