Introduction to Data Science

Sign in at **tinyurl.com/hcsbc2-28-19**

Main Topics:

- APIs
- K Means Clustering
- K Nearest Neighbors

What is an API?

APIs:

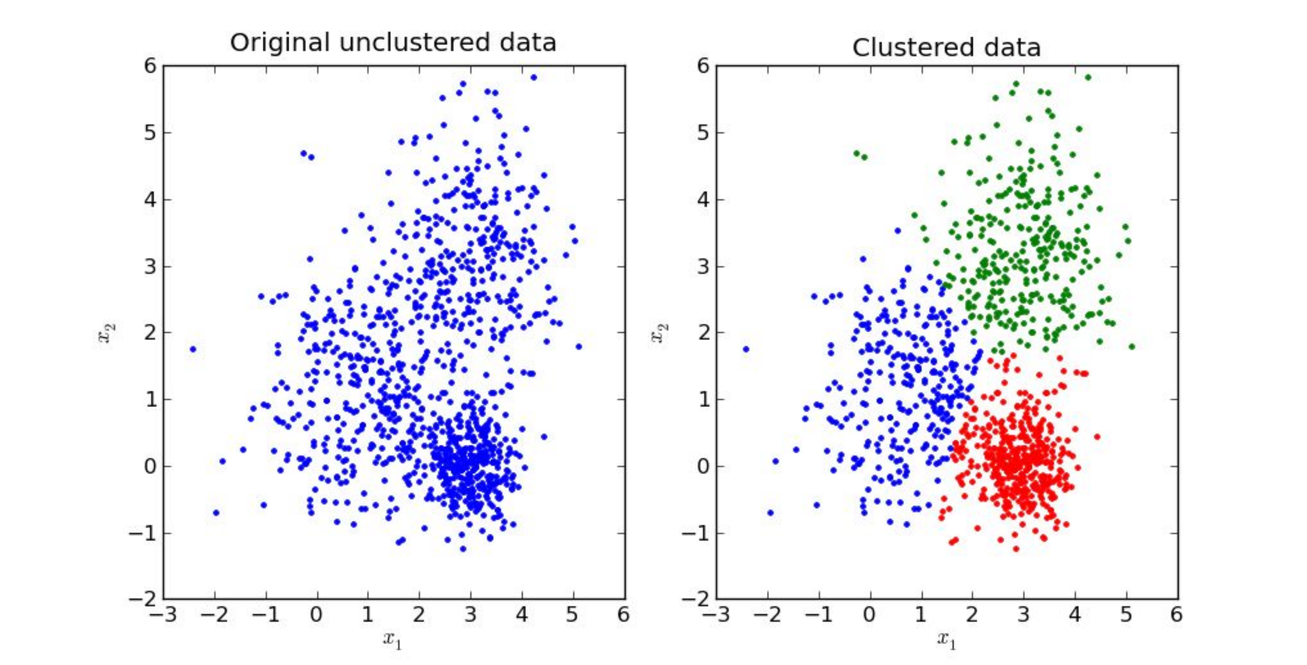
- Consist of a set of classes and functions
- Allow you to use functional code someone else has written.
- Always take the same input and produce the same output (allowing the API creator to edit code without affecting API users)
- Note: Some APIs require accounts/payment

API Examples:

- Google Maps (paid)
- jservice.io (http://jservice.io/)
- lyrics.ovh (https://lyricsovh.docs.apiary.io/#reference/0/lyrics-of-a-song/search)
- LOTS MORE: (https://github.com/toddmotto/public-apis#dictionaries)

Machine Learning

K Means Clustering (Unsupervised)

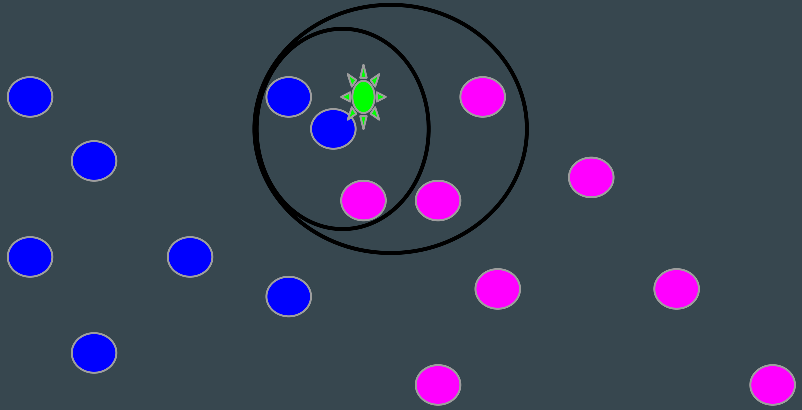


Algorithm

1. Choose k random points (p1, p2, ... pk)
2. Assign each data point to the point (pn) that is nearest to it
3. Recalculate points (p1, p2 ... pk) based on the mean position of the points in each group.
4. Repeat steps 2 and 3 until no points move between iterations

Visual: http://www.bytemuse.com/post/k-means-clustering-visualization/

K Nearest Neighbors (Supervised)



Algorithm

1. Use dataset already split into clusters (This is why we call it ‘supervised’)
2. Find the k nearest data points to the new point (these are the nearest neighbors)
3. Classify the new point as whichever cluster contains the most nearest neighbors

Assignment (Groups of 1-3)

- Apply 1 or both of the machine learning techniques we talked about today to a dataset of your choosing.
- This dataset will ideally be pulled from some API (check out the link on slide 5), but if you are having trouble, it's alright to download and use a csv file as we did with the tulip dataset.
- Include a README file with a SHORT, INFORMAL explanation of what data you used and why you thought it was interesting.
- Ideas:
    - Find data on test scores and GPAs, and try to sort students into colleges
    - Find weather data and try to sort days into seasons based on data
    - Anything you think is interesting!

Assignment Tips

- With Using APIs:
  - Make sure the API you want is available for free!
  - Be prepared to manipulate data in your ipython file to fit pandas database
- Machine learning algorithms can work in many dimensions, but think about what you feel comfortable working with and visualizing before getting started.
- NORMALIZE YOUR DATA!!!

Helpful Links:

- Pandas documentation: https://pandas.pydata.org/pandas-docs/stable/index.html
- Overview of K Means CLustering: http://benalexkeen.com/k-means-clustering-in-python/
- Overview of KNN Classification: https://www.kaggle.com/skalskip/iris-data-visualization-and-knn-classification
- Some APIs: https://github.com/toddmotto/public-apis#test-data
- US Gov. Data: https://catalog.data.gov/dataset?res_format=CSV
- More csvs: https://www.kaggle.com/datasets
- Example code from bootcamp:

Submission Instructions:

- Submit at: tinyurl.com/hcsbc2-sub
- If you're working with a group, only 1 of you has to submit.
- Deadline: March 14, 11:59 pm