

HW7: Clustering

Due Nov 10, 2020 by 11:05am **Points** 100 **Submitting** a file upload
File Types zip **Available** Nov 3, 2020 at 8pm - Nov 10, 2020 at 11:05am 7 days

This assignment was locked Nov 10, 2020 at 11:05am.


Assignment Goals

- Implement hierarchical clustering
- Process fun real-world data

Summary

Using the publicly available Pokemon stats, you'll be performing clustering on these stats. Each Pokemon is defined by a row in the data set. Because there are various ways to characterize how strong a Pokemon is, it is often desirable to convert a raw stats sheet into a shorter feature vector. For this assignment, you will represent a Pokemon's strength by two numbers: "x" and "y". "x" will represent the Pokemon's total offensive strength, which is defined by Attack + Sp. Atk + Speed. Similarly, "y" will represent the Pokemon's total defensive strength, which is defined by Defense + Sp. Def + HP. After each Pokemon becomes that two-dimensional feature vector, you will cluster the first 20 Pokemon with hierarchical agglomerative clustering (HAC).

Program Specification

Download the data in CSV format: [Pokemon.csv](https://canvas.wisc.edu/courses/205182/files/16048268/download?download_frd=1) 
(https://canvas.wisc.edu/courses/205182/files/16048268/download?download_frd=1)

Write the following Python functions:

1. **load_data(filepath)** — takes in a string with a path to a CSV file formatted as in the link above, and returns the first 20 data points (**without** the Generation and Legendary columns but retaining all other columns) in a single structure.
2. **calculate_x_y(stats)** — takes in one row from the data loaded from the previous function, calculates the corresponding x, y values for that Pokemon as specified above, and returns them in a single structure.
3. **hac(dataset)** — performs single linkage hierarchical agglomerative clustering on the Pokemon with the (x,y) feature representation, and returns a data structure representing the clustering.

You may implement other helper functions as necessary, but these are the functions we will be testing.

Load Data

Read in the file specified in the argument (the DictReader from [Python's csv module](https://docs.python.org/3/library/csv.html#csv.DictReader) (<https://docs.python.org/3/library/csv.html#csv.DictReader>) will be of use) and return a list of dictionaries, where each row in the dataset is a dictionary with the column headers as keys and the row elements as values. These dictionaries should **not** include the Generation and Legendary columns, as we will not be using them in this program. Also, you only should have the first 20 Pokemon in this structure. Note the first 20 refers to row 1 through row 20. Lastly, make sure to convert columns with numerical data to int in this method (e.g. HP, Attack, etc).

You may assume the file exists and is a properly formatted CSV.

Calculate Feature Values

This function takes in the data from a single row of the raw dataset as read in the previous function (i.e. a single dictionary, **without** the Generation and Legendary values but retaining all other columns). This function should return the x, y values in a tuple, formatted as `(x, y)`.

Perform HAC

For this function, we would like you to mimic the behavior of [SciPy's HAC function](https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html) (<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>), `linkage()`. You may not use this function in your implementation, but we strongly recommend using it to verify your results!

Input: A collection of m observation vectors in n dimensions may be passed as an m by n array (for us, this will be a list of tuples, not a numpy array like for `linkage()`!). All elements of the condensed distance matrix must be finite, i.e. no NaNs or infs. In our case, m is the number of Pokemon (here 20) and n is 2: the x and y features for each Pokemon.

Using **single linkage**, perform the hierarchical agglomerative clustering algorithm as detailed on [slide 42 of this presentation](https://happyharrycn.github.io/CS540-Fall20/lectures/introML_hierarchical_clustering.pdf) (https://happyharrycn.github.io/CS540-Fall20/lectures/introML_hierarchical_clustering.pdf). Use a standard Euclidean distance function for calculating the distance between two points.

Output: An $(m-1)$ by 4 matrix `Z`. At the i -th iteration, clusters with indices `Z[i, 0]` and `Z[i, 1]` are combined to form cluster $m + i$. A cluster with an index less than m corresponds to one of the m original observations. The distance between clusters `Z[i, 0]` and `Z[i, 1]` is given by `Z[i, 2]`. The fourth value `Z[i, 3]` represents the number of original observations in the newly formed cluster.

That is:

- Number each of your starting data points from 0 to $m-1$. These are their original cluster numbers.
- Create an $(m-1) \times 4$ array or list. Iterate through the list row by row.
- For each row, determine which two clusters you will merge and put their numbers into the first and second elements of the row. The first point listed should be the smaller of the two cluster indexes.

The single-linkage distance between the two clusters goes into the third element of the row. The total number of points in the cluster goes into the fourth element.

- If you merge a cluster containing more than one data point, its number (for the first or second element of the row) is given by m + the row index in which the cluster was created.
- Before returning the data structure, convert it into a NumPy matrix.

If you follow these guidelines for input and output, your result should match the result of

`scipy.cluster.hierarchy.linkage()` and you can use that function to verify your results. Be aware that this function does **not** contain code to filter NaN values, so this filtering should be performed before calling the function.

Tie Breaking

In the event that there are multiple pairs of points with equal distance for the next cluster:

Given a set of pairs with equal distance $\{(x_i, x_j)\}$ where $i < j$, we prefer the pair with the smallest first cluster index i . If there are still ties $(x_i, x_j), \dots (x_i, x_k)$ where i is that smallest first index, we prefer the pair with the smallest second cluster index.

Be aware that this tie breaking strategy may not produce identical results to

`scipy.cluster.hierarchy.linkage()`.

Submission Notes

Please submit your files in a zip file named **hw7_<netid>.zip**, where you replace <netid> with your netID (your wisc.edu login). Inside your zip file, there should be **only** one file named: **pokemon_stats.py**. Do not submit a Jupyter notebook .ipynb file.

All code should be contained in functions or under a

```
if __name__=="__main__":
```

check so that it will not run if your code is imported to another program.

Be sure to **remove all debugging output** before submission. Failure to remove debugging output will be **penalized (10pts)**.

If a regrading request isn't justifiable (the initial grade is correct and clear, subject to the instructors' judgment), the request for regrading will be penalized (10 pts).

This assignment due at 11/10/2020 11:00am. Submitting right at 11:00am will result in a late submission. It is preferable to first submit a version well before the deadline (at least one hour before) and check the content/format of the submission to make sure it's the right version. Then, later update the submission until the deadline if needed.

Changelog

Updated to look at only the first 20 Pokemon

HW7

Criteria	Ratings		Pts
load_data() returns a list of dictionaries	10 pts Full Marks	0 pts No Marks	10 pts
load_data() result does not include Generation/Legendary columns	10 pts Full Marks	0 pts No Marks	10 pts
load_data() contains correct values from input csv	10 pts Full Marks	0 pts No Marks	10 pts
calculate_x_y() returns a single two-element tuple	10 pts Full Marks	0 pts No Marks	10 pts
calculate_x_y() return values match expected outputs	20 to >0.0 pts Full Marks	0 pts No Marks	20 pts
hac() returns an (m-1)*4 array, matrix, or list where m = the number of non-NaN Pokemon	10 pts Full Marks	0 pts No Marks	10 pts
the third column of the hac() return array is strictly increasing	10 pts Full Marks	0 pts No Marks	10 pts
hac() tie-breaking conforms to stated protocols	10 pts Full Marks	0 pts No Marks	10 pts
hac() return value matches expected outputs	10 to >0.0 pts Full Marks	0 pts No Marks	10 pts
Total Points: 100			