CSE 190, Spring 2015: Homework 2

Instructions

Please submit your solution **at the beginning of the week 5 lecture (April 28)** or outside of CSE 4102 beforehand. Please complete homework **individually**.

You will need the following files:

50,000 beer reviews : `http://jmcauley.ucsd.edu/cse190/data/beer/beer_50000.json`.

Facebook ego network : `http://jmcauley.ucsd.edu/cse190/data/facebook/egonet.txt`.

Code examples : `http://jmcauley.ucsd.edu/cse190/code/week3.py`

Executing the code requires a working install of Python 2.7 with the scipy/sklearn packages installed.

Tasks (PCA & Clustering):

From the 50,000 beer reviews data, construct features as shown in the code example from week 3, i.e., `X = [[x['review/overall'], x['review/taste'], x['review/aroma'], x['review/appearance'], x['review/palate']] for x in data]`

1. What is the 5-d mean ($\bar{x}$) of the 50,000 ratings (1 mark)?
2. Suppose we wanted to 'compress' our data just by replacing each of the 50,000 points with the mean found above. What is the 'reconstruction error', here defined as

$$\sum_{x \in X} \|\bar{x} - x\|_2^2$$

   for the compressed data (recall that $\|y\|_2^2 = \sum_i y_i^2$) (1 mark)?

Perform k-means clustering on the same data, with two clusters whose centroids are initialized to `centroids = [[0,0,0,0,1],[0,0,0,1,0]]` (you may use whatever language you like!)

3. What are the centroids of the two clusters after convergence (1 mark)?
4. Suppose we ran k-means and the centroids we obtained were:

$$\begin{aligned} c_1 &= [4.17993, 4.23675, 4.14107, 4.08866, 4.12518] \\ c_2 &= [3.09862, 3.06899, 3.14020, 3.38222, 3.11332] \end{aligned}$$

   (note these that these should be *similar to* the solution you obtained for the previous question, but not quite the same!)
5. How many of the 50,000 points are closest to each of the two centroids c_1 and c_2 above (1 mark)?
6. Suppose you wanted to compress your data by replacing each point by one of the two centroids c_1 and c_2 above (i.e., whichever one is nearest each point). What would be the reconstruction error in this case (1 mark)?

Tasks (community detection):

Download the Facebook ego-network data.

1. How many nodes and edges are in the graph (1 mark)?
2. How many connected components are in the graph, and how many nodes are in the largest connected component (1 mark)?
3. Implement *clique percolation* and apply it to the ego-network graph. How many communities are discovered after running clique percolation with with 4-cliques, and which nodes are their members (2 mark)? *Hint: The code from lecture 3 includes a snippet to extract all 3- and 4-cliques from the ego-network, which may be useful.*