CSE 255, Spring 2015: Homework 2

Instructions

Please submit your solution **at the beginning of the Monday week 5 lecture (October 26)** or outside of CSE 4102 beforehand. Please complete homework **individually**.

You will need the following files:

50,000 beer reviews : `http://jmcauley.ucsd.edu/cse255/data/beer/beer_50000.json`.

Facebook ego network : `http://jmcauley.ucsd.edu/cse255/data/facebook/egonet.txt`.

Code examples : `http://jmcauley.ucsd.edu/cse255/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. Suppose we wanted to 'compress' our data just by replacing each of the 50,000 points with their mean vector.[1] 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)?

2. Find the PCA components (i.e., the transform matrix) using the week 3 code. Suppose we want to compress the data using just three PCA dimensions. How large is the reconstruction error when doing so (1 mark)?[2]

Next we'll implement *hierarchical clustering* on the data. This will probably be slow, so you can just use the first 500 data points to implement your solution. We'll implement clustering as described in the lecture, by merging clusters that minimize the euclidean distance between centroids.

3. How large are the *last two* clusters to be merged, and what are their centroids (1 mark)?

4. Suppose you wanted to compress your data by replacing each point by one of the two centroids of the two clusters found above (i.e., every point belonging to cluster 1 gets replaced by the centroid of cluster 1, every point belonging to cluster 2 gets replaced by the centroid of cluster 2, etc.). How large is the reconstruction error when doing so (1 mark)?

Tasks (Community Detection):

Download the Facebook ego-network data.

1. How many connected components are in the graph, and how many nodes are in the largest connected component (1 mark)?

Next we'll implement a 'greedy' version of normalized cuts, using ***just the largest connected component*** found above. First, split it into two equal halves, just by taking the 50% of nodes with the lowest and 50% with the highest IDs.

2. What is the normalized-cut cost of the 50/50 split you found above (1 mark)?

[1] i.e., for each dimension, replace it by the mean of the 50,000 values for that dimension.

[2] Hint: See Lecture 5, slide 32.

Now we'll implement our greedy algorithm as follows: during each step, we'll move one node from one cluster to the other, choosing whichever move *minimizes the resulting normalized cut cost* (in case of a tie, pick the node with the lower ID). Repeat this until the cost can't be reduced any further.

3. What are the elements of the split, and what is its normalized cut cost (1 mark)?

4. Repeat this process, but this time using *four* communities (i.e., a 25%/25%/25%/25% split, and choosing the greediest move of a node among the four communities). Again, if there's a tie between nodes, pick the node with the lower ID, or if there's a tie between communities to move it to, then pick the community with the lowest ID node. Give the four communities (1 mark).