MSC-BDT5002 Knowledge Discovery and Data Mining, Fall 2018

Assignment 3

Deadline: Nov. 14th, 11:59 pm ,2018

Task Description

Unsupervised learning is a branch of machine learning that learns from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. In this assignment, you need to cluster a certain amount of image data. **We will not tell you an exact number of cluster you should do, you need to find the most appropriate number of clusters by yourself.**

File Description

Data URL：https://www.dropbox.com/sh/gr2istwq2qrnjy8/AAD2dP8T57hQnDvh1UW-3wZUa?dl=0

In total there are 5,011 images.

Sample_submission.csv: The sample submission format you should follow.

Notes

1. Your assignment will be graded by the clustering accuracy and clarification for your feature engineering and model details (in readme.pdf).
2. TA will check your source code carefully, so your code must be runnable. Keep your code clean and comment it clearly.
3. You can use any programming language. In principle, python is preferred.
4. Cheating is not allowed. Your result MUST be reproducible.
5. Plagiarism will lead to zero mark.
6. You can use any clustering methods.

Submission Guidelines

1. Assignment should be submitted to mscbdt5002fall18@gmail.com as attachment.
2. You need to zip the following three files together:

   a . **A3_itsc_stuid_readme.pdf**. Write your feature engineering in it

   b . **A3_itsc_stuid_code.zip**: The zip file contains all your source codes.

   c . **A3_itsc_stuid_prediction.csv**: The clustering result. Each column stands for one cluster. For example, if you get 8 clusters, your .csv file should contains 8 columns. The number in the column indicates the name of the image without a suffix. For example, if the name of an image is ‘01234.jpg’, you only need to write ‘01234’ into the result.

3. Attachment should be named in the format of: A3_itsc_stuid.zip. E.g. A3_lliny_20181314.zip

4. Submissions after the deadline or not following the rules above are NOT accepted.