Machine Learning Homework 6

Kernel K-means and Spectral Clustering

Due Date 23:59 2024/12/12.

I. **Homework Objective**

Use whatever your favorite language to code out kernel k-means, spectral clustering (both normalize cut and ratio cut). You should use the new kernel defined in **III. Kernel** , considering spatial similarity and color similarity upon the clustering.

Important: scikit-learn and SciPy package is NOT allowed in this assignment.

II. **Data**

Two 100*100 images are provided (image1.png & image2.png), and each pixel in the image should be treated as a data point, which means there are 10000 data points in each image.

III. **Kernel**

For both kernel k-means and spectral clustering, please use the new kernel defined below to compute the Gram matrix.

$$\mathrm{k}(x, x') = e^{-\gamma_s \|S(x)-S(x')\|^2} \times e^{-\gamma_c \|C(x)-C(x')\|^2}$$

This new defined kernel is basically multiplying two RBF kernels in order to consider spatial similarity and color similarity at the same time. $\mathrm{S}(x)$ is the spatial information (i.e. the coordinate of the pixel) of data x, and $\mathrm{C}(x)$ is the color information (i.e. the RGB values) of data x. Both γ_s and γ_c are hyper-parameters which you can tune in your own way.

IV. What should you do:

a. Part1: Clustering & Visualization

Please use the new defined kernel to implement kernel k-means, spectral clustering (both normalize cut and ratio cut) on the given two images. All the data points should be clustered into 2 clusters.

You need to make videos or GIF images to show the clustering procedure (visualize the cluster assignments of data points in each iteration, colorize each cluster with different colors) of your kernel k-means and spectral clustering (both normalize cut and ratio cut) programs.

(Hint : Numpy can help you to solve the eigenvalue problem.)

b. **Part2: Try more clusters.**

In addition to cluster data into 2 clusters, try more clusters (e.g. 3 or 4) and show your results.

You should make videos or GIF images to show the clustering process and discuss it in the report.

c. **Part3: Try different initializations.**

For the initialization of k-means clustering used in kernel k-means and spectral clustering (both normalize cut and ratio cut), try different ways and show corresponding results, e.g. k-means++.

You should make videos or GIF images to show the clustering process and discuss it in the report.

d. **Part4: Experiments on the coordinates in the eigenspace**

For spectral clustering (both normalize cut and ratio cut), you can try to examine whether the data points within the same cluster do have the same coordinates in the eigenspace of graph Laplacian or not.

You should plot the result and discuss it in the report.

V. Report:

- Submit a report in PDF format. The report should be written in English.
- You should explain everything you have done in this homework and show all your results in the report. (Since this assignment is mainly graded from a report, please spend more time on doing your report)
- You should follow the report format below and add titles for each part, otherwise your report grade may be impacted!

【Report format】

1. **Code with detailed explanations (30%)**
   - Paste the screenshot of your functions with comments and explain your code. For example, explain the process to clustering and show different initialization methods, etc.
   - Noted that if you don't explain your code, you cannot get any point even you show the code screenshots.
     i. Part1 (kernel k-means 5%, normalized cut 5%, ratio cut 5%)
     ii. Part2 (5%)
     iii. Part3 (5%)
     iv. Part4 (5%)

2. **Experiments settings and results (30%) & discussion (20%)**
   - Show the experiment setting (e.g. hyper-parameters) and results we asked you to show in **IV. What should you do.**
   - When referring to the result videos or GIF images, please paste some frames of them and specify the corresponding filenames of the videos or GIF images in the report.
   - Discuss the experimental results.
     i. Part1 (8%) & (5%)
     ii. Part2 (8%) & (5%)
     iii. Part3 (8%) & (5%)
     iv. Part4 (6%) & (5%)
3. **Observations and discussion (20%)**
   i. Compare the performance between different clustering methods. (8%)
   ii. Compare the execution time of different settings. (8%)
   iii. Anything you want to discuss. (4%)

VI. Turn in:

1. Report (.pdf)
2. Source code
3. Videos or GIF images of clustering procedure

You should put all above into a folder named **ML_HW6_yourstudentID_name** (e.g. ML_HW6_0856XXX_王小明).

And then zip the folder to **ML_HW6_yourstudentID_name.zip** (e.g. ML_HW6_0856XXX_王小明.zip).

- If the zip file name has format error or the report is not in pdf format, penalty will be imposed (-10).
- Submit your homework in time.
  - After the deadline, you can still submit in the following 7 days, you will get only 70% of the original score.
  - Starting from the seventh day after the deadline, you cannot submit your homework and you will get 0 score.
  - Whenever you submit your homework, the latest submission will be used for grading. (so **don't accidentally submit something after the deadline**, you will get 30% discount no matter what)

VII. Packages allowed in this assignment:

You are only allowed to use numpy, scipy.spatial.distance, package for reading image and visualizing results. Official introductions can be found online.
Important: scikit-learn and SciPy is not allowed.