Machine Learning Homework 6

Kernel K-means & Spectral clustering

`Due Date 23:55 2022/5/26`

> ⚠️ You are only allowed to use `numpy`, `scipy.spatial.distance`, and *package for reading image and visualizing results*
> ⚠️ Important: `scikit-learn` and `Scipy` is not allowed.

1. Homework objective

Use whatever your favorite language to implement:

- **kernel k-means**
- **spectral clustering** (both `normalized cut` and `ratio cut`).

You should consider spatial similarity and color similarity upon the clustering.

2. Data

Two 100×100 images are provided, and each pixel in the image should be treated as a data point, which means there are 10000 data points in each image.

3. Kernel

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

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

This new defined kernel is basically multiplying two RBF kernels in order to consider spatial similarity and color similarity at the same time. $S(x)$ is the spatial information (i.e. the coordinate of the pixel) of data x, and $C(x)$ 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.

4. Requirements

Part 1.

- **You need to make videos or GIF images to show the clustering procedure of your kernel k-means and spectral clustering.**
- 👁️ Visualize the cluster assignments of data points in each iteration,
- 🎨 Colorize each cluster with different colors
- ✅ Do both **normalized cut** and **ratio cut**.
- 💡 Hint : Numpy can help you to solve the eigenvalue problem.

Part 2.

- **In addition to cluster data into 2 clusters, try more clusters (e.g. 3 or 4 ...) and show your results.**
- 👁️🎨You also need to make videos or GIF images as in Part1
- ✅ Do both **normalized cut** and **ratio cut**.
- 🤔 Discussion: What is the best cluster number? Why do you think so?

Part 3.

- **Try different ways to initialize kernel k-means, (e.g. k-means++) and spectral clustering (both normalized cut and ratio cut).**
- 👁️🎨You also need to make videos or GIF images as in Part1
- 🤔 Compare the result for each initialization.

Part 4.

- **For spectral clustering (both normalized cut and ratio cut), 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.
- 🤔 **Bonus**: Compare the result for different cluster number (This part is optional, you may not get any bonus point, please do it for fun only).

5. Report

Important Rules

- 🚨Submit a report in *pdf format*. The report should be written in *English*.
- 🚨Please follow the **report format**. If you skip some sections in the report format, you score will be affected. Additional content outside the format is welcome (but we may not be able to give you extra points).
- 🚨Please don't explain the code line by line. You need to explain it clearly and well structured. For example, explain which part you done in the function.
- 🚨Since this homework is mainly graded by report, please spend more time on it. (e.g. well organized) We won't give you any points if you just finish the code.

Report Format

1. **Code (40%) :**
   - 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.
   - Part 1. (kernel k-means 5%, normalized cut 5%, ratio cut 5%)
   - Part 2. (5%)
   - Part 3. (10%)
   - Part 4. (10%)
   - ☠️ *Note that if you don't explain your code, you cannot get any points in section 2 and 3 either*.
2. **Experiments(30%) & Discussion(20%)**
   - Please show the results for both images in this section.
   - Part 1. (8%)&(5%)
   - Part 2. (8%)&(5%)
   - Part 3. (8%)&(5%)
   - Part 4. (6%)&(5%)
3. **Observations and Discussion (10%)**
   - Anything you want to discuss, such as comparing the performance between different kernels or the execute time of different settings.
4. **Any Suggestion for this class (0%)**

6. Submision

To Submit the Homework, you should:

- Zip your contents in one file, including
  - Report (.pdf)
  - Source Code
  - Video of GIF images of clustering procedure
- Name the zip file as `ML_HW6_{your student id}_{your 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, there will be a penalty (we are considering -10).
- Submit your homework *in time*
  - After deadline, you can still sumit in the following 7 days, you will get only 70% of original score.
  - Starting from the seven'th day after the deadline, you can not 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 accidently submit something after deadline, you will get 70% discount no matter what)