Data Mining and Matrices
Assignment 4: Spectral Clustering

The archive provided to you contains this assignment description, the dataset, as well as the R code fragments for you to complete. Comments and documentation in the code provide further information.

It suffices to fill out the "holes" that are marked in the code fragments provided to you, but feel free to modify the code to your liking. You need to stick with R though.

Please adhere to the following guidelines in all the assignments. If you do not follow those guidelines, we may grade your solution as a FAIL.

Provide a single ZIP archive with name `dmm18-<assignment number>-<your ILIAS login>.zip`. The archive needs to contain:

- A single PDF report that contains answers to the tasks specified in the assignment, including helpful figures and a high-level description of your approach. Do not simply convert your R code to a PDF! Write a separate document, stay focused and brief. As a guideline, try to stay below 10 pages.
- All the code that you created and used in its original format.

Make sure that your report is self-explanatory and follows standard scientific practice. Use the tasks numbers of the assignments as your section and subsection numbers. Label all figures (and refer to figure labels in your write-up). Include references if you used additional sources or material.

Hand-in your solution via ILIAS until the date specified there. This is a hard deadline.

Introduction

This assignment focuses on spectral clustering. We provide implementations of all relevant methods; your task is to experiment with these methods and evaluate the results.

We will work with a dataset of handwritten digits (`http://ftp.ics.uci.edu/pub/machine-learning-databases/optdigits/optdigits.names`); there are 50 examples per digit. The dataset is a 500×64 matrix; each row corresponds to a digit, each digit is represented by an 8×8 "image", and each value consists of a gray value in $[0, 16]$.

The goal of this assignment is to cluster the digits using spectral clustering. In doing so, you will experiment with the wealth of parameters (and the wealth of their combinations) of spectral clustering.

The data is given in a CSV file named `digits.csv`. The R script `util.R` contains functions relevant to spectral clustering, the file `a04-sc.R` contains some examples on how to use these functions as well as prototype code for you to complete. Familiarize yourself with the dataset and the provided functions before working through the assignment.

1 Cuts

We start by computing the values of cut (`cut`), ratio cut (`rcut`) and normalized cut (`ncut`) for a given clustering of the $m = 500$ digits into $k = 10$ clusters. In what follows, we make use of the notation and definitions of Exercise 5.3.

As in the previous assignment, we represent a clustering by a vector `cluster` $\in \{ 0, \ldots, k-1 \}^m$ of cluster numbers; each element represents one data point (row) and its value indicates the corresponding cluster number.

a) Given a vector `cluster`, compute the corresponding *cluster assignment matrix* $C \in \{ 0, 1 \}^{m \times k}$. To do so, complete the function `cam` provided to you.

b) Given an adjacency matrix $\boldsymbol{W} \in \mathbb{R}_+^{m \times m}$ and a clustering `cluster`, compute the value of `cut`, `rcut`, and `ncut`. To do so, complete the function `cut` provided to you.

c) Run your `cut` function on the `cluster.test` clustering provided to you. Observe that `cut` $\gg$ `rcut` $\gg$ `ncut`. Why is this the case? Is this always true?

2 Similarity graphs

To construct a similarity graph, we make use of the Gaussian kernel (and vary parameter σ) as well as the various neighborhood graphs we discussed in the lecture.

a) Compute the full similarity graph using $\sigma = 50$. Study the distribution of the resulting similarities (e.g., using a histogram as described in the provided R script). Is $\sigma = 50$ a good choice? Try to find a good setting

for σ by trying both smaller and larger values (but do not run spectral clustering yet). Discuss!

b) For $\sigma = 50$, find (roughly) the smallest ϵ such that the ϵ-neighborhood graph is connected. Note that you can use the magnitudes of the smallest eigenvalues of the Laplacian to judge whether or not the graph is connected (as before, R may return a very small value instead of 0 eigenvalues). Now find the smallest k such that the symmetric kNN graph is connected, and the smallest k such that the mutual kNN graph is connected. Plot the resulting similarity matrices. Are they different? If so, why? Discuss!

c) For the symmetric k-nearest neighbor graph, manually determine values for σ and k that appear suitable to you. As before, do not run spectral clustering yet.

d) Consider any dataset in Euclidean space. Suppose that we use the Gaussian kernel with parameter σ to obtain similarities and subsequently construct a symmetric k-nearest neighbor graph. Describe (in simple terms) what changes to expect in the so-obtained graph when we increase or decrease σ. Is there anything that does not change?

3 Spectral clustering

In this experiment, we try to cluster the digit data into 10 different clusters; the "optimal" clustering assigns the same digits to the same cluster, and different digits to different clusters. As in the previous assignment, we know the "correct" clustering of the data (called `labels` in the R script); this is often not the case in practice. We study clustering quality in terms of accuracy and the confusion matrix (see provided R code).

a) First cluster the digits data using k-means on both the raw data and the first 10 principal component scores. Visualize the result and compute the accuracy. Are the results good? Which "errors" are made?

b) Use your parameter settings of task 2c) and run spectral clustering. Compare the result with the results obtained above. Which method worked best? Did your parameter settings produce good results? Which "errors" are made?

c) In practice, we may not know the optimal number of clusters. Use the eigengap heuristic to estimate a good choice for the number of clusters. Discuss!

d) Now "tune" the parameters of spectral clustering with 10 clusters to obtain an accuracy above 0.88. (This is something we can't do in practice!) Why do you think that the so-obtained parameters work well?

e) **Competition (optional).** Find a parameterization of spectral clustering to obtain the best possible clustering of the digits data with respect to accuracy. If you want, try any other clustering method that you know. We'll compare results in the tutorial group. Is there any hope to get better results on this data by any clustering method? Why or why not?

f) **DBSCAN (optional).** DBSCAN is a well-known method for graph-based clustering. Read up on DBSCAN and try it out on our data (package `fpc`, function `dbscan`). Do you get similar results as spectral clustering? What, in your opinion, are the differences between the clustering methods?

Handing in Your Solution

As before, your report should document what you did and which results you obtained. Report (briefly) the results with different parameter combinations and include plots of the similarity matrices for the parameter combinations you selected initially and ultimately. Include also all accuracy scores. Do not forget to list the parameter combinations you tried, and explain why you selected specific values for ϵ, k, and σ.

If you participate in the competition, explain how you derived the clustering (including all parameter settings) and give code that reproduces the clustering. Also describe your search process (e.g., How did you tune parameters? Why did you use a specific method?).