Automation of Biological Research: 02-450/02-750

Carnegie Mellon University

Homework 6
Due: 12/9/2016
Hand-in: Submit your code to Cody AND email your responses as a zipped file in the format $LastName_$FirstName_Homework6.zip to: ABR-instructors@googlegroups.com.

Keep the directory structure intact, with your written responses in the top level. So your zipped file should have the structure:

```
$LastName_$FirstName_Homework6.zip
|--$LastName_$FirstName_Homework6.pdf
|--Q1
    |---getAdjacencyMatrix.m
    |---getDegreeMatrix.m
    |---getNormalizedLaplacian.m
    |---SpectralClustering.m
    |---runExperimentsQ1.m
    |---simGaussian.m
    |---distEuclidean.m
    |---getDataForQ1.m
    |---twomoon_small.csv
|--Q2
    |---LandmarkSelection.m
    |---ExpandLandmarks.m
    |---LandmarkClustering.m
    |---FormatClustering.m
    |---getDataForQ2.m
    |---Heap.m
    |---MinHeap.m
    |---UpdateComponents.m
    |---runExperimentsQ2.m
|--Q3
    |---getDataForQ3.m
    |---runExperimentsQ3.m
    |---sequence_dataset.csv
```

Overview

We often want to cluster samples into subtypes. Typically, clustering is performed by expert analysis of phenotypes, but this method has a number of problems (limited throughput, costly, inaccurate phenotypes). To overcome such challenges, genomic markers are increasingly used for clustering. Passive clustering methods assume that a complete pairwise distance matrix is known before clustering; however, this distance matrix can be expensive to obtain. For instance, when clustering biological sequences, obtaining a distance matrix involves performing sequence alignment and scoring for n^2 sequences, which is computationally intensive. In this homework, we will investigate strategies for active clustering, in which we only query a small number of pairwise distances.

Our investigation will focus on the Landmark-Clustering algorithm, which you should read about in the paper here, or in the slides here. Additionally, in the first question we will use spectral clustering, which you should read about in the slides here.

We ARE using Cody for this assignment.

Question 1 - Passive Clustering - Synthetic Data (25 points)

In this question, we will explore passive strategies for clustering. We are comparing two methods of clustering discussed in class - spectral clustering and k-means clustering. `kmeans` is a built-in function in Matlab, so we will only be implementing the Spectral Clustering method. You will implement 2 functions - getDegreeMatrix and getNormalizedLaplacian. These methods are called by the SpectralClustering function. For each file, you have stub code and need to fill in the blanks. The blanks are short - each file can be completed with fewer than 3 additional lines of code. For each coding question, submit your answers to Cody and include the code in the zipped file you submit.

Tasks

A. Implement getDegreeMatrix (5 points) Submit your solution to Cody.

B. Implement getNormalizedLaplacian (5 points) Submit your solution to Cody.

C. Synthetic Data (5 points) Use the runExperimentsQ1.m file to plot all the data points, colored by their cluster label from the Spectral Clustering method. Include your plots in your handin. Qualitatively, what is the difference between the decision surfaces of the two methods? Where does this difference come from?

D. Number of Queries (10 points) How many pairwise distance queries does the spectral clustering method take? How many pairwise distance queries does k-means take? To calculate these numbers, count the number of iterations (which are printed out) and multiply by n^2.

What to hand in

Hand in your code for parts A-B (also submit to Cody), and plots and responses to parts C-D.

Question 2 - Landmark Clustering - Synthetic Data (45 points)

Often, getting the pairwise distances between points can be costly. For example, to get the pairwise distance between two DNA sequences, we need to first align the two sequences, which is computationally expensive. As discussed in class, active clustering reduces the number of pairwise distances needed to cluster points. Here, we will be implmenting and testing the Landmark-Clustering algorithm.

You will implement 2 functions: LandmarkSelection and ExpandLandmarks. The LandmarkClustering function calls these functions. runExperimentsQ2 parses the data in the data file, so you do not need to write code to manipulate the file. The files you need to edit are LandmarkSelection.m and ExpandLandmarks.m.

Tasks

A. LandmarkSelection (10 points) Submit your solution to Cody.

B. ExpandLandmarks (20 points) Submit your solution to Cody. NOTE: This is algorithm 4 in the paper, not algorithm 3. Algorithm 3 is a high-level pseudocode, algorithm 4 acutally tells you how to implement it.

C. Synthetic Data Decision Surfaces (10 points) Use the runExperimentsQ2.m file to plot all the synthetic data points, colored by their cluster. Which passive method does the decision surface of the Landmark-Clustering method most resemble? Why? Include your plots below.

D. Number of Queries (5 points) How many queries does the Landmark Clustering method make for this dataset?

What to hand in

Hand in your code for parts A-B (also submit on Cody), and your plots and answers to C-D.

Question 3 - Real Data and Analysis (30 points)

Now we will test the clustering methods on biological sequences, which are classified according to evolutionary relatedness. Sequence alignment and distance metrics are computationally costly for large sequences, so our goal is to reduce the number of pairwise distances that must be calculated. **There is no code to be submitted to Cody for this problem.**

A. Decision Surfaces (10 points) Use the runExperimentsQ3.m file to plot all the data points, colored by their clusters. Include your plots. Which method seems to work best?

B. Assumptions (10 points) What is the (c, ϵ)-property that the Landmark-Clustering method assumes? Should we expect it to hold for biological datasets? Why or why not?

C. Parameters (10 points) What is the effect of the s_{min} parameter in the Landmark-Clustering method?

What to hand in

Hand in your plots and answers to A-C.