Homework 4
MATH 191 Topics in Data Science
Posted: Wednesday, November 18, 2015
Due in class, Monday, November 30, 2015

Problem 1: The Combinatorial Laplacian (20p)
Let L denote the Combinatorial Laplacian associated to a graph G. Show that the multiplicity of the zero eigenvalue is at least the number of connected components of the graph. (In fact, equality holds, but you are not being asked to prove this.)

Problem 2: Ranking (20 points)
(a) The HW-4 folder contain two CSV files related to this ranking problem. The first CSV file *England_2009_2010_TeamNames.csv* contains the names of all teams in the English Premier League 2009-2010 season. The second file *England_2009_2010_Scores.csv* contains the net outcomes of all the two matches played by any team (home and away). In other words, $C_{ij} = -1$ if team i won both matches, $C_{ij} = 1$ if team i lost both matches, and $C_{ij} = 0$ otherwise. Implement the Serial-Rank algorithm we discussed in class. Please submit your code and the obtained ranking.

Problem 3: Diffusion Maps (20 points)
The same HW-4 folder also contains an R object file *twoCircles.data*, with the 2-dimensional coordinates of $n = 500$ points. Denote the points corresponding to the first 250 rows by *inner* points, and the points corresponding to the last 250 rows by *outer* points.

(a) Plot the point cloud, and mark the two different kind of plots with two different symbols (and, if you wish, with two different colors, but you do not have to submit your HW in color). Here is how you can do this in R `http://www.endmemo.com/program/R/pchsymbols.php`

(b) Perform PCA on this data set, and show a 2-D plot of your data points (using the 2 principal component scores). Again plot the inner and outer points with distinct symbols.

(c) Implement the diffusion maps algorithm we covered in class. (Note that a simple way to compute the matrix of all pairwise distances in one shot is

DIST = as.matrix(dist(A, method = "euclidean", diag = FALSE, upper = TRUE, p = 2));

Experiment with different values of the parameter ϵ in diffusion maps, and include the embeddings you obtained via the first two non-trivial eigenvectors, for two different values of $\epsilon = \{0.75, 1\}$, and plot the inner and outer points with distinct symbols. Which of these two representations is easier to cluster/separate? Compare the results with those from (b).

Problem 4: A minimization problem over $\mathbb{Z}_2$ (20 points)
If A is a matrix of size $n \times n$ with entries $\{-1, 0, 1\}$, and $\boldsymbol{x} \in \mathbb{Z}_2^n$ denotes a vector of length n with elements $\{-1, +1\}$, prove that the following holds true

$$\min_{\boldsymbol{x} \in \mathbb{Z}_2^n} \sum_{(i,j) \in E} (x_i - A_{ij} x_j)^2 = \min_{\boldsymbol{x} \in \mathbb{Z}_2^n} \boldsymbol{x}^T (D - A) \boldsymbol{x}, \tag{1}$$

where D is a diagonal matrix with D_{ii} being the degree of node i, that is, $D_{ii} = \sum_{j=1}^{n} |A_{ij}|$.