Practice Problems - Data Mining

Link Analysis

Problem 1 Given a graph in Figure 1.

(a) Construct the transition matrix of the graph.

(b) Star with the vector $\mathbf{v}_0 = [1/5, 1/5, 1/5, 1/5, 1/5]^T$, compute the Page rank vector $\mathbf{v}_3$ after three iterations with the taxation parameter $\beta = 0.8$.

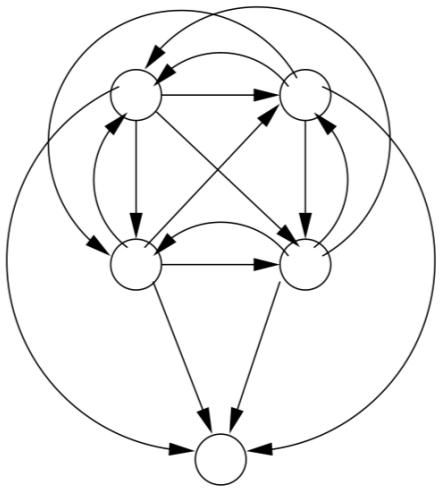

Figure 1: The graph for Problem 1.

Problem 2 [Exercise 5.1.5 MMDS Book] Suppose we recursively eliminate dead ends from the graph, solve the remaining graph, and estimate the PageRank for the dead-end by the algorithm we learn in Programming Assignment 3. Suppose the graph is a chain of dead ends, headed by a node with a self-loop, as suggested in Figure 2. What would be the Page- Rank assigned to each of the nodes, assuming that the graph has n nodes?

Problem 3 [Exercise 5.4.1 MMDS Book] In class, we analyzed the spam farm of Figure 3, where every supporting page links back to the target page. Repeat the analysis for a spam farm in which each supporting page links both to itself and to the target page.

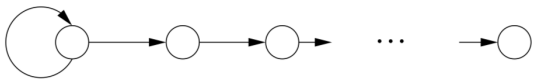

Figure 2: The graph for Problem 2.

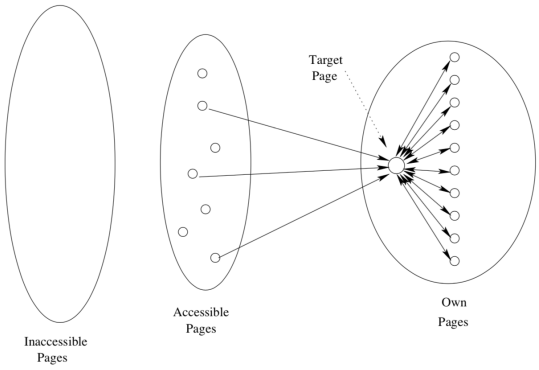

Figure 3: The graph for Problem 4.

Problem 4 [Exercise 5.4.3 MMDS Book] Suppose two spam farmers agree to link their spam farms. How would you link the pages in order to increase as much as possible the PageRank of each spam farm's target page? Is there an advantage to linking spam farms?

Problem 5 Start with a hubbiness vector $\mathbf{h}_0 = [1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}]$ and an authority vector $\mathbf{a}_0 = [1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}]$, compute the hubbiness and authority vectors $\mathbf{h}_3$, $\mathbf{a}_3$ after three iterations for the graph in Figure 1.

Clustering

Problem 6 Given a point set in Figure 4, compute the Euclidean distance from point $[4, 10]^T$ to all other points.

Problem 7 (Exercise 7.2.3 MMDS Book): Construct the cluster tree of the example point set in Figure 4 if we choose to merge the two clusters whose resulting cluster has the smallest diameter. The diameter of a cluster X is:

$$\text{Diam}(X) = \max_{x,y \in X} d(x, y) \tag{1}$$

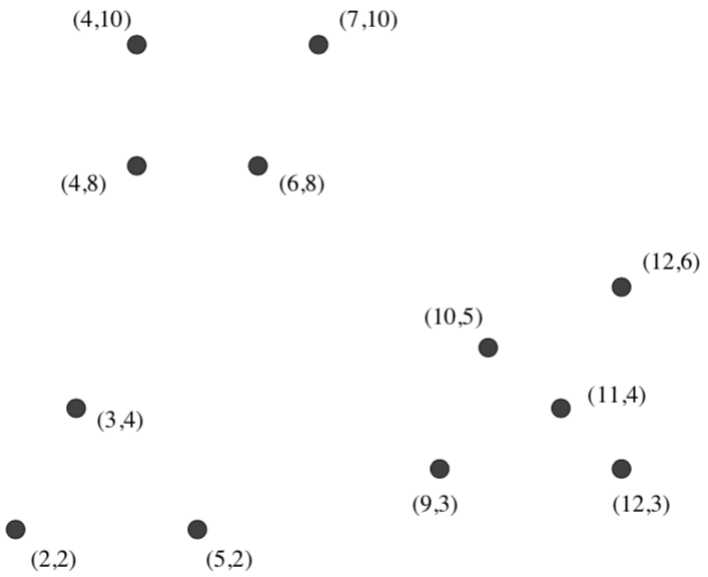

Figure 4: An example point set for Problem 6 and 7.

Problem 8 For the points of Figure 4, if we select four starting points using the method presented in class (Section 7.3.2 in the book), and the first point we choose is $(3, 4)$, which other points are selected.

Problem 9 Find four clusters after 2 iterations of K-means, using the four initial centroids found in Problem 8.

Advertising

Problem 10 Suppose that there are five advertisers, A,B,C, D, E. Each advertiser has a budget of 1. Each query cost 1 unit. Find a query sequence and a bidding scenario for all advertisers so that on that sequence, the greedy algorithm has competitive ratio $\frac{1}{2}$. A bidding scenario for an advertiser is a set of queries that the advertiser bids on.

Problem 11 Suppose that there are three advertisers, A,B, and C. There are three queries, x,y, and z. Each advertiser has a budget of 2. Each query cost 1 unit. Advertiser A bids only on x; B bids on x and y, while C bids on x,y, and z. Find a query sequence so that on that sequence, the balanced algorithm has competitive ratio at most $\frac{2}{3}$.

Problem 12 Sow that in the adword problem, where advertisers can bid by different amounts and can have different budgets, the balanced algorithm has competitive ratio arbitrarily close to 0.

Problem 13 Sow that in the adword problem, where advertisers can bid by different amounts and can have different budgets, the algorithm that always sells a query to the advertiser that places highest bid has competitive ratio arbitrarily close to 0.

Recommendation System

Problem 14 Given the utility matrix:

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} I_1 & I_2 & I_3 & I_4 \end{array} \\ \begin{pmatrix} 5 & & 4 & \\ 1 & & 3 & 2 \\ & 3 & 1 & 2 \\ 2 & & 4 & \end{pmatrix} \end{array}$$

We will factorize the matrix M into the product of two matrices U, V of dimensions 4×3 and 3×4, respectively. Suppose that we start with the following matrices:

$$U_0 = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 0 & 3 \end{bmatrix} \quad V_0 = \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \\ 2 & 2 & 0 & 0 \end{bmatrix} \tag{2}$$

(a) What is the RMSE of the factorization by U_0, V_0

(b) Find the first column of U, assuming other columns and V are given. That is, find x_1, x_3, x_3, x_4, so that the factorization

$$\begin{bmatrix} x_1 & 2 & 0 \\ x_2 & 1 & 1 \\ x_3 & 1 & 1 \\ x_4 & 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \\ 2 & 2 & 0 & 0 \end{bmatrix} \tag{3}$$

has minimum RMSE.

(c) Find the second row of V, assuming other rows and U are given. That is, find y_1, y_2, y_3, y_4, so that the factorization

$$\begin{bmatrix} x_1 & 2 & 0 \\ x_2 & 1 & 1 \\ x_3 & 1 & 1 \\ x_4 & 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 & 0 \\ y_1 & y_2 & y_3 & y_4 \\ 2 & 2 & 0 & 0 \end{bmatrix} \tag{4}$$

has minimum RMSE with x_1, x_2, x_3, x_4 from part (b).

Problem 15 We use the same utility matrix in Problem 14. Suppose that instead of optimizing for a whole column of U, we optimize one element at a time.

(a) Find u_{22}, so that the factorization

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & u_{22} & 1 \\ 1 & 1 & 1 \\ 2 & 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 \\ 2 & 2 & 0 & 0 \end{bmatrix} \tag{5}$$

has minimum RMSE.

(c) Find v_{23}, so that the factorization

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & u_{22} & 1 \\ 1 & 1 & 1 \\ 2 & 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 & 0 \\ 0 & 1 & v_{23} & 1 \\ 2 & 2 & 0 & 0 \end{bmatrix} \tag{6}$$

has minimum RMSE where u_{22} is from part (a).

Social Network Analysis

Problem 16 Given a graph in Figure 5. Using the Girvan-Newman algorithm:

(a) Compute the contribution of shortest paths from node (1), (2), (3), (4), (6) to betweeness of each edge.

(b) Calculate the betweeness of every edge.

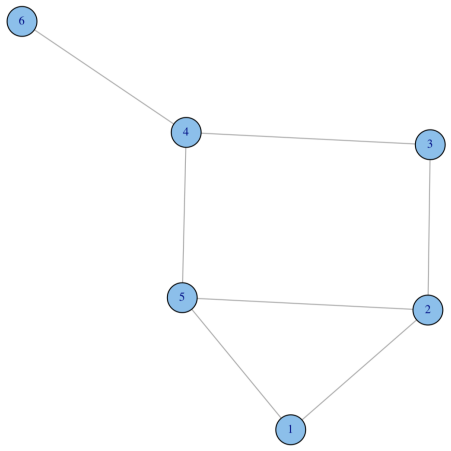

Figure 5: The graph for Problem 16. Imagine is taken from `http://www2.unb.ca/~ddu/6634/Lecture_notes/Lect10_community_R.pdf`

Problem 17

(a) Find the Laplacian matrix of the graph in Figure 5.

(b) Find the smallest eigenvalue and its corresponding eigenvector of the graph in Figure 5.

Problem 18 Given a graph in Figure 6.

(a) Given that Laplacian matrix has the second smallest eigenvalue $\lambda_2 = 2$, find the corresponding eigenvector.

(b) What partition of the nodes does the vector in part (a) suggest?

Problem 19 Given a graph in Figure 5.

(a) Suppose graphs are generated by picking a probability p and choosing each edge independently with probability p. What value of p gives the maximum likelihood of seeing that graph?

(b) What is the probability the graph in Figure 5 is generated with the value p in part (a)?

Problem 20 Compute the MLE for the graph in Figure 5 for the following guesses of the memberships of the two communities: $\{1, 2, 3, 4, 5\}, \{2, 3, 4, 5, 6\}$.

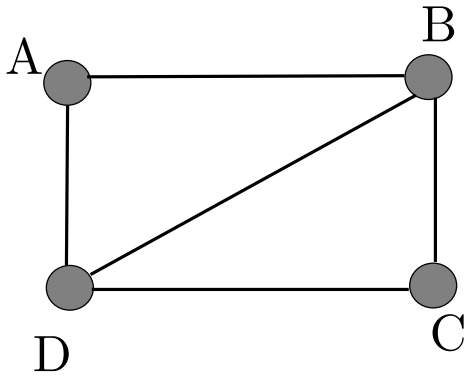

Figure 6: Graph for Problem 18

Problem 21 [Exercise 10.6.1] If, in Figure 7, you start the walk from Picture 2, what will be the similarity to Picture 2 of the other two pictures? Which do you expect to be more similar to Picture 2?

For a simple calculation, you can iterate the Sim Rank algorithm 5 times with $\beta = 0.8$

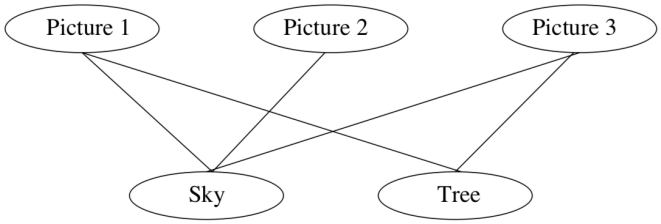

Figure 7: Graph for Problem 21

Problem 22 [Exercise 10.7.2 MMDS book] For the graph in Figure 8 determine:

(a) What is the minimum degree for a node to be considered a "heavy hitte"?

(b) Which nodes are heavy hitters?

(c) Which triangles are heavy-hitter triangles?

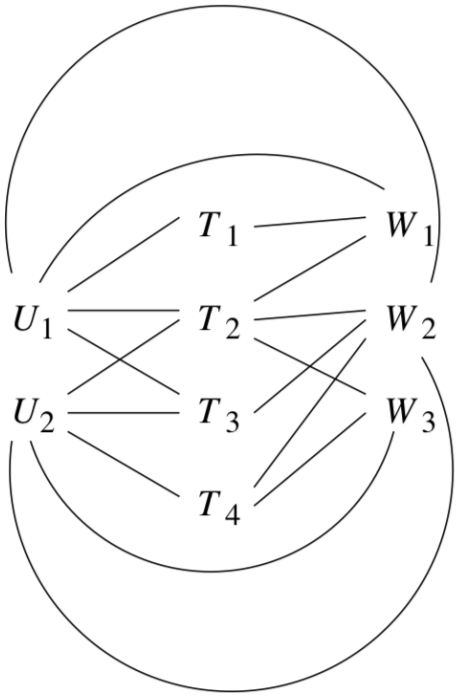

Figure 8: Graph for Problem 22

Dimensionality Reduction

Problem 23 Using Power Iteration algorithm to find:

(a) The largest eigenvalue and the corresponding eigenvector of the Laplacian matrix of the graph in Figure 6.

(b) The second largest eigenvalue and the corresponding eigenvector of the Laplacian matrix of the graph in Figure 6.

(c) The second smallest eigenvalue and the corresponding eigenvector of the Laplacian matrix of the graph in Figure 6.

Problem 24 [Exercise 11.1.3] For any symmetric 3×3 matrix

$$\begin{bmatrix} a-\lambda & b & c \\ b & d-\lambda & e \\ c & e & f-\lambda \end{bmatrix} \tag{7}$$

there is a cubic equation in λ that says the determinant of this matrix is 0. In terms of a through f, find this equation.

Problem 25 [Exercise 11.1.4] Find the eigenvectors and eigenvalues for the following matrix:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 5 \end{bmatrix} \tag{8}$$

using the solving equation approach.

Problem 26 [Exercise 11.2.1] Let M be the matrix of data points

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 3 & 9 \\ 4 & 16 \end{bmatrix} \tag{9}$$

(a) What are M^TM and MM^T?

(b) Compute eigenvectors and eigenvalues for M^TM.

(c) What do you expect to be the eigenvalues of MM^T?

(d) Find the eigenvectors of MM^T, using your eigenvalues from part (c).

Problem 27 [Exercise 11.2.2] Prove that if M is any matrix, then M^TM and MM^T are symmetric.

Problem 28 [Exercise 11.3.1] Let M be the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix} \tag{10}$$

It has rank 2, as you can see by observing that the first column plus the third column minus twice the second column equals 0.

(a) Compute the matrices M^TM and MM^T.

(b) Find the eigenvalues for your matrices of part (a).

(c) Find the eigenvectors for the matrices of part (a).

(d) Find the SVD for the original matrix M from parts (b) and (c). Note that there are only two nonzero eigenvalues, so your matrix Σ should have only two singular values, while U and V have only two columns.

(e) Set your smaller singular value to 0 and compute the one-dimensional approximation to the matrix M.

(f) How much of the energy of the original singular values is retained by the one-dimensional approximation?