BIOE 210, SPRING 2020

HOMEWORK 3

Due Wednesday, 4/8/2020 before 5:00pm CDT.
Upload a single PDF with your answers to Gradescope.

Part I (20 points)

(1) Consider the vectors $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$.

(a) Construct an orthonormal set of basis vectors from these vectors.

(b) Decompose the vector $\begin{pmatrix} -2 \\ 1 \\ 3 \end{pmatrix}$ onto your orthonormal basis.

You can use Matlab or a calculator to answer the following questions.

(2) Consider the matrix

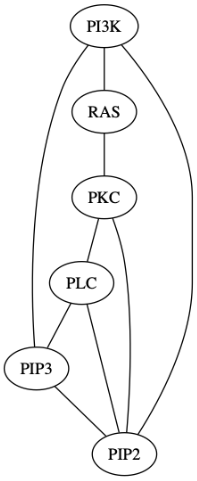
$$\mathbf{A} = \begin{pmatrix} 5 & 3 & 1 \\ 3 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

(a) Find the eigenvectors and the corresponding eigenvalues for the matrix.

(b) Decompose the vector $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ on the eigenvectors.

(c) Without performing a matrix multiplication, what is $\mathbf{Ax}$?

(3) The following figure depicts protein-protein interactions in a human signal transduction network. Your goal is to find the most central and least central proteins in this network.



(a) Construct an adjacency matrix to represent the connections in the network.
- An adjacency matrix is a square matrix **A** such that each element a_{ij} equals 1 if node i is directly connected to node j.
- Since the above graph is undirected, your adjacency matrix should be symmetric. If $a_{ij} = 1$, the $a_{ji} = 1$.
- The diagonal elements (a_{ii}) must be left zero, since no node is "connected" to itself.

(b) Calculate the leading eigenvector for the adjacency matrix. The leading eigenvector is associated with the eigenvalue with the largest magnitude.

(c) Using the magnitude of the entries in the leading eigenvector, report the most central and least central proteins in the network. How does the centrality of these proteins compare with the number of connections involving these proteins? Is the most central protein always the protein with the largest number of direct connections?

Part II: Machine Problem (30 points)

A team of researchers used DNA microarrays to measure gene expression in a large set of breast cancer cell lines (Kao, et. al, PLOS ONE 4(7): e6146. doi:10.1371/ journal.pone.0006146). In this exercise, you will use gene expression profiles from this study to build a classifier that differentiates between invasive and regular ductal carcinoma (IDC and DC).

(1) Load the mat file `HW3_data.mat` , which contains the following variables:

- `training_lines` is a Matlab table containing gene expression data for the IDC and DC cell lines. Each of the 8750 rows corresponds to a gene with variable expression across the cell lines. Each of the 28 columns represents a cell line. The following cell lines were classified as invasive (IDC) by a pathologist: BT474, BT483, BT549, EFM19, MDA134, MDA175, SUM102, T47D, UACC812, UACC893, ZR75_1, and ZR75_30. The remaining cell lines are noninvasive ductal carcinoma (DC).
- `patient_samples` is a Matlab table containing gene expression values for the same 8750 genes from the training data. Each column corresponds to a different patient biopsy.

(2) Build an SVM classifier that separates IDC from DC samples.
- The Matlab command `fitcsvm` accepts numerical arrays, not tables, so convert your table with the function `table2array`.
- Pay attention to the dimensions of your inputs, especially what rows and columns correspond to in your data and for `fitcsvm`.

(3) Perform both k-fold (with 4 folds) and leave-one-out cross validations using the command `crossval`. Using the function `kfoldLoss`, report the accuracy of your model using each validation method.

(4) Repeat the cross validation five times for both the k-fold and leave-one-out methods. Does the accuracy change for either method? Why or why not?

(5) Using the Matlab `predict` function, determine if each biopsy in the patient data set is invasive (IDC) or regular (DC) ductal carcinoma.

Remember to submit all code, outputs, and explanations for these problems.