University of Waterloo
ECE 657A: Data and Knowledge Modeling and Analysis
Winter 2019
Homework 3: Eigenvector Decomposition
Due: January 25th, 2019 11:59pm

Overview

Collaboration: Do your work and report individually. You can collaborate on the right tools to use and setting up your programming environment.

Hand in: One report per person, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work. Report as a PDF or a python notebook.

General Objective: To study how to apply some of the methods discussed in class on two datasets. The emphasis is on analysis and presentation of results not on code implemented or used.

Specific Objectives:

- Establish your software stack to carry out data analysis homeworks, assignments and the project for the rest of the course.
- Load a simple dataset and perform some basic data preprocessing.

Tools: You can use libraries available in MATLAB, python, R or any other programs available to you. You need to mention which libraries you are using, any blogs or papers you used to figure out how to set carry out your calculations.

Data sets

For this homework you will use the Communities and Crime Data Set:

- `http://archive.ics.uci.edu/ml/datasets/communities+and+crime`
- Download from Data Folder link, read data set description.

Tasks

In the class we talked about how to reduce data dimensionality by extracting new set of features using PCA, LDA and other methods. The basis of these methods is the eigenvector decomposition of the data matrix.

- Load the crime dataset and store it as a matrix (The data is already normalized so you should not need to normalize the data further.)
- Compute the eigenvectors and eigenvalues (you can use the mathematical formulation or call a library in your chosen environment)
- Report a table with the top 20 eigenvalues, is there a clear point where you could cut off the dimensions?

- **For fun (on your own):** Plot all the data points in a 2D scatterplot by projecting data points to the two eigenvectors with the highest eigenvalues. Colour the points by some dimension of the original data (eg. PopDens or medIncome) to see what patterns arise.