Mohammed J. Zaki

Dmcourse /

Assign2

Assignment 2

Due Date: Tue 27th Sep, Before Midnight

Part I and II have to be done by both CSCI4390 and CSCI6390. There is an extra question for CSCI6390. The bonus question can be attempted by both CSCI4390 and CSCI6390.

(Part I) Kernel Principal Components Analysis (50 points)

You will implement the Kernel PCA (KPCA) algorithm as described in Algorithm 7.2 (Chapter 7, page 207). You need to compute the kernel and then center it, followed by extracting the dominant eigenvectors, which will give you the components of the directions in feature space. Next you will project and visualize the data.

Try KPCA on the UCI dataset Concrete_Data.txt. You can read the description of the data at the UCI repository. I have already converted the xls file into text, and removed the titles for easy import with Numpy. The data has 1030 points and 9 attributes.

To compute the principal components (PCs) of the kernel matrix, you may use the inbuilt numpy function **eigh**. Using the linear kernel, how many dimensions are required to capture 95% of the total variance? For the same linear kernel, compute the projected points along the first two kernel PCs, and create a scatter plot of the projected points.

Next, use the covariance matrix for the original data to compute the regular principal components, i.e., for the covariance matrix. Project the data onto the first two PCs. How do the eigenvalues and the projection compare with that obtained via Kernel PCA with linear kernel?

Finally, use the gaussian kernel and repeat the exercise for kernel PCA. Project the points onto the first two PCs and plot the scatter plot. For the variance or spread of the gaussian kernel, namely σ^2, read it from the command line. Try different values and submit your plot for the value that makes most sense to you (observe the projected plots for various spread values and then decide).

(Part II) Diagonals in High Dimensions (50 points)

Your goal is the compute the empirical probability mass function (EPMF) for the random variable X that represents the angle (in degrees) between any two diagonals in high dimensions.

Assume that there are d primary dimensions (the standard axes in cartesian coordinates), with each of them ranging from -1 to 1. There are 2^d additional half-diagonals in this space, one for each corner of the d-dimensional hypercube.

Write a python script that randomly generates $n = 100,000$ pairs of half-diagonals in the d-dimensional hypercube, and computes the angle between them (in degrees).

Plot the EPMF for three dif ferent values of d, as follows $d = \{10, 100, 1000\}$. What is the min, max, value range, mean and variance of X for each value of d?

Extra Question: CSCI6390 Only (20 points)

What would you have expected to have happened analytically? In other words, derive formulas for what should happen to angle between half-diagonals as $d \to \infty$. Does the EPMF conform to this trend? Explain why? or why not?

Bonus Question for both CSCI4390 and CSCI6390 (20 points)

What is the expected number of occurrences of a given angle θ between two half-diagonals, as a function of d (the dimensionality) and n (the sample size)?

What to submit

- Write two python scripts named as RCS ID-Assign2-part1.py and RCSID-Assign2-part2.py , one for each of the parts. For part1, read the filename from the command line, assume it is in the local directory. So, part1 will be run as **RCSID-Assign2-part1.py FILENAME SPREAD** . FILENAME is the datafile name, and SPREAD is the σ^2 value.
- Submit a PDF file named RCSID-Assign2.pdf that should include your solutions to each of the questions (just cut and paste the output from python). The figures should also be part of this file. Report in your output the spread value that made most sense to you.
- Submit the scripts and pdf file as an email attachment to: datamining.rpi@gmail.com . The subject of your email should be "RCSID-Assign2 Submission".

Page last modified on September 22, 2016, at 02:21 PM