CSCI4390-6390 Assign2

Assign2: High Dimensional Data and Dimensionality Reduction

Due Date: Sep 22, before midnight (11:59:59PM, Alofi Time; GMT-11)

Both Part I and II have to be done by all sections. Differences have been specified by **CSCI4390** and **CSCI6390** labels.

Part I: Principal Components Analysis (50 points)

You will implement the PCA algorithm as described in Algorithm 7.1 (Chapter 7). You need to compute the eigenvectors, and then project visualize the data. To compute the principal components (PCs), you may use the inbuilt numpy function **eigh**.

Run PCA on the Appliances energy prediction data set You should ignore the first attribute, which is a date-time variable, and you should a the last attribute, which is a duplicate of the previous one.

Next, determine and print how many dimensions are required to capture $\alpha = 0.975$ fraction of the total variance?

Also print the mean squared error in the approximation using the first three components.

Plot the PCs

CSCI4390 Only: Project the points along the first two PCs, and create a scatter plot of the projected points.

CSCI64390 Only: Project the points along the first three PCs, and create a 3D scatter plot of the projected points.

Part II: Diagonals in High Dimensions (50 points)

Your goal is the compute the probability mass function for the random variable X that represents the angle (in degrees) between any two 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 ar additional half-diagonals in this space, one for each corner of the d-dimensional hypercube.

Randomly generate $n = 100,000$ pairs of half-diagonals in the d-dimensional hypercube (random d-dimensional vectors with elements and compute the angle between them (in degrees).

Plot the probability mass function (PMF) for three different values of d, as follows $d = 10, 100, 1000$. Recall that PMF is simply the pl angle versus the probability of observing that angle in the sample of n points for a given value of d. What is the min, max, value range, m variance of X for each value of d?

CSCI6390 Only: What would expect analytically? In other words, derive formulas for what should happen to angle between half-diagona ∞. Does the PMF conform to this trend? Explain why? or why not?

What to submit

- Write two python scripts named as Assign2-part1.py and 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 **Assign2-part1.py FILENAME ALPHA**. FILENAME is the datafile name, and ALPHA is the approximation threshold α. In other your script must compute and return the correct number of components to capture α fraction of total variance.
- For part2, the script will be run as **Assign2-part2.py**.
- Submit a PDF file named Assign2.pdf that should include your solutions to each of the questions (just cut and paste the output from The figures should also be part of this file. **Failure the submit the PDF will result in lost points.**
- Submit the scripts and pdf file via submitty

Policy on Academic Honesty

You are free to discuss how to tackle the assignment, but all coding must be your own. Please do not copy or modify code from anyone els including code on the web. Any students caught violating the academic honesty principle will get an automatic F grade on the course and v referred to the dean of students for disciplinary action.