CS 481, Homework 1

Out: January 22, 2017, Due: Feb 04, 2017, Total: 75

Note:

- This homework will carry 7.5 points towards your final score.
- Please answer all the questions below.
- If a question asks you to write a code, you need to submit a working code through canvas submission site. It is also your responsibility to test that the code runs in pegasus.cs.iupui.edu machine.
- Homeworks are individual work, please do not collaborate with others inside or outside of the class. Software will be used to determine code similarity, so do not take a chance.
- Start early and if you need help, post your questions on piazza or use instructor/TA's office hour.
- To make it uniform throughout the class, for covariance matrix computation divide by n instead of $n-1$ in all your works.

Questions

Download the following file from `http://archive.ics.uci.edu/ml/machine-learning-databases/magic/magic04.data`. More information on this dataset can be obtained from UCI ML Repository. The dataset has 10 real attributes, and the last one is simply the class label, which is categorical, and which you will ignore for this assignment. Assume that attributes are numbered starting from 0. You should use Python and the NumPy scientific computing package for answering the following questions.

1: Write a function to Compute the sample covariance matrix as inner products between the columns of the centered data matrix (see equation 2.31). Show that the result from your function matches the one using `numpy.cov` function. (10)

2: Use `linalg.eig` to find the first two dominant eigenvectors, and compute the projection of data points on the subspace spanned by these two eigenvectors. Now, compute the variance of the datapoints in the projected subspace using the subroutine that you wrote for Question 1 (Do not print the projected datapoints on stdout, only print the value of the variance). (15)

3: Use `linalg.eig` to find all the eigenvectors, and print the covariance matrix Σ in its eigen-decomposition form ($U\Lambda U^T$) (10)

4: Write a subroutine to implement PCA Algorithm (Algorithm 7.1, Page 198). (30)

5: Use the program above and find the principle vectors that we need to preserve 90% of variance? Print the co-ordinate of the first 10 data points by using the above set of vectors as the new basis vector. (10)

Deliverables

Submit a scriptfile and an output file. Write a script named assign1-iuUserName.py that takes as input the data filename, and prints the answers to the above question to stdout. Do not hard code the filename inside the script, but rather you should read the file name from the command line. Save your output to a textfile and name it assign1-iuUserName.txt, Submit your script and output file in canvas.