The University of Texas at Austin
Department of Electrical and Computer Engineering

EE379K: Data Science Lab — Fall 2018

Lab Four

Caramanis/Dimakis Due: Tuesday, Oct 30, 10:00am 2018.

Problem 1: PCA.

1. Generate 20 random points in $d = 3$, from a Gaussian multivariate distribution with mean $[0, 0, 0]$ and covariance matrix

$$\Sigma_1 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.7 \end{bmatrix}.$$

   Let's call this data with label 1. Also generate 20 random points in $d = 3$ from another Gaussian with mean $[1, 1, 1]$ and covariance

$$\Sigma_2 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.01 \end{bmatrix}.$$

   Let's call that data with label 2. Create a three dimensional plot of the clouds of data points, labeled with the two labels.

2. What do the points look like ?

3. Concatenate all the points and ignore the labels for now. You have created an X matrix with 40 data points in $d = 3$ dimensions. Find the covariance matrix of this dataset. **Do not simply use np.cov, build the covariance matrix from the definition using linear algebra operations in python.**

4. Let's do PCA on this dataset using $k = 2$ dimensions: Find the two eigenvectors of the covariance matrix with the largest eigenvalues. Project the data points on these two vectors and show the two dimensional plot with the clouds of points. Also show the labels of the points. Did PCA make it easier to distinguish the two labels in two dimensions ? **Again, do not simply use sklearn PCA. You are only allowed to use matrix operations and np.linalg.eig to find eigenvalues and eigenvectors of a matrix.**

Problem 2: Low rank approximation of Mona Lisa.

1. Load the Mona Lisa image (in grayscale) and treat it as a matrix M. Perform a singular value decomposition on this matrix using **linalg.svd**. You can perform a low-rank approximation by zeroing out singular values and keeping only the top k. Show the best rank $k = 2$, $k = 5$ and $k = 10$ approximation to Mona Lisa.

2. If each pixel is represented by two bytes, how many bits is your compressed Mona Lisa for each of those k rank approximations?

Problem 3: Using Low Rank Structure for Corrupted Entries.

Download files `CorrMat1.csv` and `CorrMat3.csv` from Canvas. These are each 100 by 100 matrices. Look at the data and find which entries are corrupted. Then try to correct these corrupted entries. Explain your approach. (Hint: The corrupted entries have values that are completely out of the range of the others. This should help you identify which are the corrupted ones. For completing them, the hint is that we have been talking about PCA, low rank matrices and low-rank approximations.)

Problem 4: OBGYN diagnostic and Bayes.

Probability theory is nothing but common sense reduced to calculation.
P-S. Laplace

Here is a problem on Bayesian inference:
Based on medical statistics, 10 out of every 1,000 women have breast cancer. Of these 10 women with breast cancer, 9 test positive. Of the 990 women without cancer, about 89 nevertheless test positive. A woman tests positive and wants to know whether she has breast cancer for sure, or at least what the chances are.

What is the best answer?
(1) The probability that she has breast cancer is about 90 percent;
(2) Out of 10 women with a positive mammogram, about 1 has breast cancer;
(3) The probability that she has breast cancer is about 81 percent;
(4) Out of 10 women with a positive mammogram, about 9 have breast cancer;
(5) Out of 10 women with a positive mammogram, about 7 have breast cancer;
(6) The probability that she has breast cancer is about 1.

Solve this problem by carefully describing all events and applying the rules of probability and Bayes rule. First, write down all the conditional probabilities.

For example, define the event C (i.e. a patient has cancer). The first piece of information 'Ten out of every 1,000 women have breast cancer' can be written as $P(C) = 10/1000 = 0.01$. Similarly, the second piece of information is describing $P(TP/C)$ (Probability to test positive conditioned that the patient has cancer).

Use Bayes rule to solve this problem (which is equivalent to computing the PPV (Positive Predictive Value) of the mammogram screening test). Write down each step as a probability calculation in detail.