CS 215 : Data Analysis and Interpretation : Assignment 4

Instructor : Suyash P. Awate

Due Date : 6 Nov 2016, Sun, 11:55 pm; Maximum Points 75

Submission Instructions:
• Submit your solution, i.e., code, resulting graphs, and the report (in Adobe PDF format), for each question on moodle.
• Submit a single zip file that contains the solution to each problem below in a separate folder.
• To get partial credit for the code, ensure that the code is very well documented.
• To get partial credit for the derivations, include all derivation steps in their full details.
• 5 points for submission in the proper format.

1. (20 points) Estimating π Using Simulation.

   Initialize the random-number generator, i.e., rng(), in Matlab with seed 0.

   Consider a bivariate random variable $X := (X_1, X_2)$, where X_1, X_2 are both independent and have a uniform distribution over $(-1, 1)$.

   (a) What is the probability that the random variable X take values within a circle of radius 1 ? Derive the expression.

   (b) Use the previous result to estimate the value of π purely relying on simulation of X. Justify your approach in words.

   (c) Report the estimates of π using sample sizes $N = 10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8$. Use the "single" datatype for the simulation. How will you write code to handle the situation when you desire to increase the accuracy of the estimate with a sample size as large as $N = 10^9$ or larger ? Does your code handle this case ?

   (d) How will you estimate for the sample size M required to have the estimate of π within $[\pi - 0.01, \pi + 0.01]$ with 0.95 probability ? Describe an algorithm and justify it. Compute this estimate for the sample size M and report it.

2. (25 points) Multivariate Gaussian.

   Initialize the random-number generator, i.e., rng(), in Matlab with seed 0.

   Generate N points (with N taking the values $10, 10^2, 10^3, 10^4, 10^5$) from a multivariate $2D$ Gaussian probability density function with mean $\mu = [1, 2]'$ and a covariance matrix C with the first row as $[1.6250, -1.9486]$ and the second row as $[-1.9486, 3.8750]$. For this simulation, you are only allowed to use the randn() and eig() functions in Matlab.

   For each data sample of size N, compute the maximum likelihood (ML) estimates of the mean and the covariance matrix.

(a) Describe and justify your method for generating sample points from the 2D Gaussian.

(b) For each value of N, repeat the experiment 100 times, and plot a box plot of the error between the true mean μ and the ML estimate $\widehat{\mu}_N$ (which depends on N), where the error measure is $\| \mu - \widehat{\mu}_N \| / \| \mu \|$. Use a logarithmic scale on the horizontal axis, i.e., $\log_{10} N$.

(c) For each value of N, repeat the experiment 100 times, and plot a box plot of the error between the true covariance C and the ML estimate $\widehat{C}_N$ (which depends on N), where the error measure is $\| C - \widehat{C}_N \|_{\text{Fro}} / \| C \|_{\text{Fro}}$. Use a logarithmic scale on the horizontal axis, i.e., $\log_{10} N$.

(d) For each value of N, for a single data sample, within a single figure, plot the $2D$ scatter plot of the generated data and show the principal modes of variation of the data by plotting a line starting at the empirical mean and going a distance equal to the empirical eigen-value along a direction given by the empirical eigen-vector.

3. (25 points) Multivariate Gaussian fitting for Principal Component Analysis.

Download the dataset comprising images of handwritten digits in `http://yann.lecun.com/exdb/mnist`.

Each image is stored as a matrix (28×28) of numbers. You can visualize these images (or matrices) in Matlab using the functions imagesc() or imshow(). Use the Matlab command "axis equal" to use the same units on each axis of the image.

For the following computations, make sure to convert (cast) the integer data type to a floating point type.

For every digit, from 0 to 9, compute (i) the mean μ, (ii) the covariance C, and (ii) the first mode of variation determined by the eigenvector v_1 and the corresponding eigenvalue λ_1 (where λ_1 is the largest of all eigenvalues) of the covariance matrix C.

Note: Before computing the mean and covariance matrix, convert each 28×28 image matrix to a $28^2 \times 1$ vector by concatenating its columns. To visualize the $28^2 \times 1$ mean vector, convert it back to a matrix and then visualize it using imagesc(). Use the reshape() function to change matrices to vectors and vice versa. The covariance matrix will be of size $28^2 \times 28^2$.

• For each digit, sort the 28^2 eigenvalues of the covariance matrix and plot them as a graph. Comment and justify what you observe. How many "principal" / significant modes of variation (i.e., number of "large" eigenvalues) do you find, for each digit ? Are the significant modes of variation equal to 28^2 or far less ? Why ?

• For each digit, show the 3 images side by side: (i) $\mu - \sqrt{\lambda_1} v_1$, (ii) μ, and (iii) $\mu + \sqrt{\lambda_1} v_1$, to show the principal mode of variation of the digits around their mean. Comment and justify what you observe. For a certain digit, say 1, what does the principal mode of variation tell you about how people write that digit ?