CS 369 Assignment 2 2017
Due 6:00 pm Friday April 7

Marked out of 24 points. 6% of the final grade.

Write your code in Python 3 and present your code and report as a Jupyter Notebook.

In your report, include explanations of what you are doing and why you are doing it. Write in whole sentences and present your results clearly.

Submit your Jupyter notebook in the raw .ipynb form and as a .html file to `https://adb.auckland.ac.nz/` by 6:00 pm on the due date.

Problem 1: Pseudoinverse [*5 points*]. For this problem, you will a need 120×100 pixel grayscale image of your own face as you used in Assignment 1. You can use any digital camera and image processing software to capture a colour image of own face, crop and scale it down to 120 pixels (height) $\times$ 100 pixels (width), and convert it to be grayscale in the pgm format. This will be used as a 120×100 rectangular matrix $\mathbf{A}$

a. For the grayscale image $\mathbf{A}$, form a pseudoinverse matrix $\mathbf{P}_{\text{inv}}$ that behaves like an inverse of $\mathbf{A}$ in that the product $\mathbf{P}_{\text{inv}}\mathbf{A} = \widehat{\mathbf{I}}$ yields an approximation to the 100×100 identity matrix $\mathbf{I}_{\mathbf{100}}$. Illustrate the obtained results in your report by displaying $\mathbf{P}_{\text{inv}}$, $\mathbf{A}$ and $\widehat{\mathbf{I}}$ as images (linearly scale $\mathbf{P}_{\text{inv}}$ and $\widehat{\mathbf{I}}$ to the range [0,255] before displaying them).

b. Calculate the elementwise error matrix $\mathbf{E} = \mathbf{I} - \widehat{\mathbf{I}}$ between the computed identity, $\widehat{\mathbf{I}}$, and the actual identity matrix $\mathbf{I}_{100}$. Display the matrix of absolute values of entries of $\mathbf{E}$, $\mathbf{E}^{abs}$ where $\mathbf{E}^{abs}_{ij} = |\mathbf{E}_{ij}|$ as an image (linearly scale the matrix to the range [0,255] before displaying it). Calculate the range, mean and standard deviation of entries of $\mathbf{E}$.

Problem 2: Principal components to find geographic structure [*11 points*].

For this problem, you will need the `snps.txt` data file from the Resources page of the course website.

This question is based on the geographic analysis of genetic SNP data from Novembre et al 2008, Nature, doi:10.1038/nature0733 discussed in Lecture 8.

The data matrix stored in `snps.txt` is a binary matrix with 100 rows and 2734 columns. Each row corresponds to an individual person and each column corresponds to a position in the individual's genome which is known to vary between individuals, known as a *single nucleotide polymorphism* or SNP. So each row is a trial and each column is a measurement (note that this is the transpose of the PCA setup in the notes, so simply take the transpose of this matrix after you read it in).

The data comes from a geographically structured population with an unknown number of sub-populations. Your task here is to perform a principal components analysis of the data to determine the number of populations and allocate individuals to each population.

a. With the help of `numpy.linalg.svd`, write code to read in the data matrix and perform a principal components analysis of it (remember to centre and scale the matrix).

b. Print out the first 5 entries of each of the first 5 principal component vectors, and print out the singular values corresponding to each of the 100 principal components in order.

c. Project each data vector along the first two principal components to get a 100×2 matrix $\mathbf{L}$ in which entry L_{ij} corresponds to the amount of the ith individual in the direction of the jth principal component. Print out $\mathbf{L}$ in your report.

d. Plot $\mathbf{L}$. From your plot, determine the number of distinct subpopulations in the data set and describe a rule for partitioning the space into the different subpopulations. Include a list of the individuals (corresponding to row numbers) in the largest subpopulation.

Problem 3: Sampling from some common distributions [*8 points*].

Use the `numpy.random` library to simulate random samples from the distributions below. When plotting histograms, make sure the axes are labeled appropriately and you don't choose far too many or far too few bins (something like $\sqrt{n}$ bins for n points is a good rule of thumb).

For each problem i-iv below:

a. state the value of the parameters of the distribution that are appropriate for the problem;

b. make a histogram of the sampled values; and

c. compare the mean and variance of the sampled values to the theoretical mean and variance for the distribution with the given parameters.

i. On average, one customer comes into a shop every 15 mins. Use the Poisson distribution to simulate the number of customers arriving every hour for 300 hours.

ii. The death rate for a type of tree is 0.01 per year. Use the exponential distribution to simulate the number of years each of 150 trees lives.

iii. A certain species of bird lays exactly 24 eggs per year but on average 75% of the eggs are predated before hatching. Use the binomial distribution to simulate the number of chicks each of 500 of these birds successfully hatch in a year.

iv. Red blood cells are about 7 μm in diameter with a standard deviation of 1μm. Use the normal distribution to simulate the size of 1000 red blood cells.