PRML Programming Assignment 1

This assignment is due on: 30/09/2020
This assignment should be submitted either as an ipython notebook or a Latex compiled document with codes, results and conclusion.
Use python 3 for implementing.
Use numpy, matplotlib libraries.
You can download the required data sets by clicking on the link text.

Question 1:

Perform the following linear algebra operations on any arbitrary matrices of your choice, using objects and functions from numpy library.

- Transpose of a matrix
- Inverse of matrix
- Trace of a matrix
- Determinant of a matrix
- Rank of a matrix
- Multiply two matrices
- Perform element wise multiplication
- Find Eigen values and vectors of a matrix
- Find the null space or kernel of a matrix.
- Find the L1, L2 and L-infinity norm of an arbitrary 5 dimensional vector.
- Compute L1 norm, Forbenous norm of a matrix

Question 2:

Given the following 3D input data.
1 1 9
2 4 6
3 7 4
4 11 4
5 9 2

- Plot the 3D data points.
- Compute the sample covariance matrix.
- Compute the eigen vectors and eigen values of the sample covariance matrix.
- Project the data points along the 2 major Eigen vectors (i.e Eigen vectors corresponding to 2 dominant Eigen values) and plot the 2D projected data points

Note: *This is called PCA: principal component analysis , used for dimensionality reduction. It projects the data along maximum variance directions that preserves the information.*

Question 3:

1. Using the Iris flower dataset , plot the histogram for each feature(any three) for any one of the species of your choice.
2. Assuming a 1D Gaussian distribution plot the distributions for each species using any one feature.(use the same feature for all 3 species)
3. Assuming a 2D Gaussian, select any 2 features and plot the distribution for each species.

Question 4:

Generate random numbers in a range, with mean and variance of your choice from:

1. Gaussian distribution
2. Binomial distribution
3. Poisson distribution
4. Uniform distribution

of sizes 50, 500 and 5000 data points, and plot the points.

Question 5:

This is a simple linear regression problem, which is to be solved by least squares technique. Using the data provided , the best straight line fitting the data points is to be evaluated. i.e find the coefficients b_0, b_1 of the line $y = b_0+b_1x$.

1. Read the following Sweden auto insurance data set, split it into training and testing sets.
2. Using the training set Compute the mean and variance of each variable (i.e. x and y)
3. Compute the covariance between the two variables (i.e Cov(x,y)).
4. Find the coefficients of the best fitting straight line.The coefficients are given by following formulae
   b_1 = Cov(x,y) / var(x)
   b_0 = mean(y) - b_1 mean(x)
   And plot the straight line along with all the data points.
5. Perform the prediction on test data set using the regression solution and compute the root mean squared error between original data values and predicted values.

Question 6:

Solve the above problem(question 3) with linear algebra.
The idea is to form the overdetermined system of linear equations of the form $y = b_0+b_1x$
The least squares solution for:

$$X\ b = Y$$

is given by

$$b = (X^T X)^{-1} Y$$

Plot the regression line and data points.