**Machine Learning Homework #1**

**Question 1 explanation:**

The central idea behind principal component analysis (PCA) is to reduce the dimensionality of a data set composed of many interrelated variables while retaining as much variation as possible. The entire process consists of six steps:

1. Ignore the labels and turn our d+1 dimensional dataset to a d dimensional dataset
2. Compute the ***mean*** for every dimension of the whole dataset.
3. Compute the ***covariance*** *matrix* of the whole dataset.
4. Compute ***eigenvectors*** and the corresponding ***eigenvalues***.
5. Sort the **eigenvectors** by decreasing **eigenvalues** and choose k eigenvectors with the largest eigenvalues to form a *d × k dimensional* matrix **W.**
6. Use this *d × k eigenvector matrix* to transform the samples onto the new subspace.

Let’s start:

* Firstly, we ignore the labels and calculate the mean which is sum of all values divided by total number of values for each column in the matrix.
* Next, we calculate the covariance matrix using the formula,



* We will end up with a square matrix.
* We can then calculate our eigen values by the below formula,



The lambda values shall give us our eigen values

* Lastly using our eigen values we can calculate their eigen vectors

In the calculations performed for question 1 we only perform the steps up to step 5, as we only need the first and second PC. Here the eigen vector with the higher eigen value becomes our first principal component. Roughly speaking, the eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data, and those are the ones we want to drop.

Question 2:

To run the code:

* Mentioned libraries need to be downloaded if not already
* Path needs to be changed in ipnby file as per user’s path for the image