Middle East Technical University | Department of Computer Engineering

CENG 499

Introduction to Machine Learning

Fall 2016-2017

Homework 3

Due date: **08.01.2017 - 23:59**

1 Objectives

To familiarize yourselves with K-means clustering and Principle Component Analysis through implementing and applying them on a real data set, for image compression problem.

2 Specifications

This homework consists of two parts. In the first part you will implement the K-means clustering algorithm. In the second, you will implement the Principle Component Analysis. In both parts, you will use the "automobile" subset of "CIFAR-10 dataset" which you can find in

`https://www.cs.toronto.edu/~kriz/cifar.html`

This dataset actually consists of 10 categories of 32x32 color images. However the providers have chosen to hold the images as 1x3072 vectors.

In the first part of this assignment, your aim is to find K "mean" colors so that we can decrease the number of colors used in the images from 256 to K, which will require less bits to represent the same image leading to a lossy compression.

In the second part, your aim is to find K principle components, so that we can express the images in terms of their linear combination which reduces 32-by-32 images into K coefficients as a lossy compression.

Homework file is continued with a review of K-means and PCA algorithms.

2.1 K-means Clustering

K-means clustering is a very intuitive and easy to implement algorithm. It consists of an initialization step and an alternating procedure that stops when there is no change as a result of alternations. At the initialization step K random points are chosen to be cluster means. The alternation process is as follows

- Assign data instances to the closest cluster center using the Euclidean Distance given in (1).
- Update each cluster center as the mean of its assigned data instances.

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

2.2 Principal Component Analysis

Assume that you are given a dataset $D = \{(\mathbf{x}^{(i)} \mid i \in [1..m]\}$, in which each $\mathbf{x}^{(i)}$ is an n-dimensional multivariate vector. Before starting the Principal Component Analysis, one should zero center and scale the dataset according to (2).

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \tag{2}$$

in which the subscript corresponds to j^{th} feature, superscript corresponds to i^{th} data example, μ and σ are the mean value and the standard deviation.

Principal components are the K set of uncorrelated variables learned from correlated ones representing the data, which have the largest variance. These correspond to the eigenvectors of the covariance matrix of the data related with the K largest eigenvalues. The covariance matrix is given in (3)

$$\Sigma = \frac{1}{m} \sum_{1}^{m} (x^{(i)})(x^{(i)})^T \tag{3}$$

where Σ is the covariance matrix.

Having found the principal components, any specific data example $x^{(ex)}$ can be represented as a linear combination of them. The coefficient vector $c^{(ex)}$ is given by (4)

$$c^{(ex)} = P^T x^{(ex)} \tag{4}$$

where P is an n-by-K matrix whose columns are the vectors corresponding to principal components.

2.3 Programming and Interpretation Tasks

Change the default format of the MATLAB workspace into long fixed-decimal format to avoid possible numerical errors.

First, you need to input the complete dataset and take the "automobile" subset. The dataset is divided into parts originally. Converting these parts into a single (2D or 3D) data design matrix according to your way of solving problems is highly recommended. As in the previous homeworks, you must use vectorized implementations. In this homework we will use **only the training part** of the dataset.

For the first part of the homework, you should write two MATLAB **functions**. The first function called **findClusterCenters** will take all the images in the training set (data design matrix is recommended) and the number of clusters (K) as input and run the K-means algorithm on all of the 3x1 pixels as data instances. To be more specific, you will use all the 3x1 pixels of all the training images to compute K 3x1 cluster centers. The function will output these cluster centers. The second one called **kmeansCompress** will take the cluster centers and an image to compress as input and assign each pixel of the image to index of the closest center using Euclidean Distance. Having done that, the image is compressed from 32x32x3

to 32x32x1. This function should also reconstruct the image such that each pixel value is replaced with the assigned cluster center. The function returns the compressed and reconstructed images.

For the second part of the homework, you should also write two MATLAB **functions**. The first one called **findPrincipalComponents** takes all the training images and the number of principal components (K) as input, computes and returns the first K principal components of the dataset as described in the section 2.2. The second one called **pcaCompress** takes the K principal components and an image to compress, and computes the coefficients to represent the image in terms of the principal components. It then reconstructs the image using these and return the coefficients and reconstructed image.

Having implemented the two methods, run the algorithms for K=2,4,16 and test using the ex1 and ex2 images provided in **examples.mat**. You should prepare a report discussing the methods. The report should contain at least the followings for the three different K values;

- The cluster means as a figure in which each color is represented as a 32x8 rectangles, and a discussion on why and how these colors are the results.
- The principle components as 32x32x3 images and the discussion of their appearance.
- Comparison of the original and reconstructed images
- Comparison of the algorithms and detailed discussion explaining the reasons behind the results
- Discussion of the effects of the value of K
- A recommendation for choosing best K
- Possible problems that can be encountered using these algorithms.

3 Restrictions and Tips

- You may need reshape and imwrite functions to save the vectors as images. Since the images are too small you may prefer to use imresize to scale them.
- Note that the training dataset consists of rotated images.
- Do not use any available MATLAB repository files without referring to them in your report.
- Toolbox function use other than Image Processing Toolbox is not allowed in this homework. Do not use any ML-related toolbox. However, you may cross-check your results utilizing toolboxes. Your homework submission must not include any high-level toolbox function other than the ones provided by the Image Processing Toolbox.
- If you encounter trouble regarding vectorization, first try to implement the tasks by using loops. Vectorization is required, since typical ML projects deal with massive amounts of data. If at the end you fail to vectorize your code, submit the current version. Although vectorization appears to be essential in MATLAB programming, since this is not a MATLAB course, unvectorized versions will only result in a minor decline (at most 10%) in the grade you are going to receive.
- Implementation should be of your own. Readily-used codes should not exceed a reasonable threshold within your total work. In fact you shouldn't need any code on repositories.
- Don't forget that the code you are going to submit will also be subject to manual inspection.

4 Submission

- **Late Submission:** You have 3 days for all the homeworks (except the first one) and the project.
- Your scripts and function files together with a 3-to-4 pages long report focusing on theoretical and practical aspects you observed regarding this task should be uploaded on COW before the specified deadline as a compressed archive file whose name should be <<student_id>_hw3> preceding the file extension.
- The archive must contain **no directories** on top of MATLAB/Octave scripts and function files.

5 Regulations

1. **Cheating: We have zero tolerance policy for cheating**. People involved in cheating will be punished according to the university regulations.
2. **Newsgroup:** You must follow the newsgroup (news.ceng.metu.edu.tr) for discussions and possible updates on a daily basis.