CS189: Introduction to Machine Learning

Homework 7

Due: 11:59 p.m. Monday, May 5th, 2014

1 Warmup

Run k-means on MNIST with 5, 10, and 20 cluster centers and visualize the centers. You must implement the k-means algorithm yourself.

2 Eigenfaces

You will now implement the Eigenfaces algorithm to find a common bases between celebrity faces and student faces.

You are provided three files as part of the homework.

- `CelebrityDatabase.zip` contains images of (aligned) faces of celebrities from which you will construct eigenfaces.
- `StudentDatabase.zip` contains images of aligned students' faces for which you will reconstruct from the eigenfaces.
- `mask.mat` contains a boolean mask that zeroes out the background in the face images. In order to keep only the relevant pixels in a face image and convert it into a vector execute:

  ```
  unmasked_pixels=find(mask); im_vector=im(unmasked_pixels);
  ```

  In order to obtain a masked image from `im_vector`, use:

  ```
  full_im = zeros(size(binary_mask));
  full_im(unmasked_pixels) = im_vector;
  ```

You need to do the following steps to find eigenfaces and the best matches. Note that you are NOT allowed to use any inbuilt functions for PCA in MATLAB.

1. Compute eigenfaces from the celebrity faces (after applying the binary mask). Visualize the top 10 eigenfaces. What kind of variations do the top eigenfaces seem to correspond to?

2. Reconstruct 5 celebrity faces from only the top 10 eigenfaces and visualize some of them with the original image and the reconstructed image side-by-side. Plot the average L_2 error as a function of the number of eigenfaces used to reconstruct the original image.

3. Pick 5 student faces and reconstruct them with top 20 eigenfaces. Visualize these reconstructions as you did above in your write up. Plot the average L_2 error as a function of the number of eigenfaces to reconstruct the original image.

You need to include all the figures, numbers and answers to the questions in your report.

3 SVD Practice

Consider the following 2×2 matrices:

$A = \begin{bmatrix} 2 & 2 \\ 1 & -1 \end{bmatrix}$ and $A = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$

For each of the above matrices, work out the following steps in order to compute the SVD of the above matrix.

1. Compute U by calculating eigenvectors AA^T.

2. Compute the entries of Σ by calculating the positive square roots of the eigenvalues of AA^T.

3. Compute V by calculating eigenvectors of A^TA.

4. Verify that $A = U\Sigma V^T$.

4 Extra Credit

Using the Million Song Dataset located at `http://labrosa.ee.columbia.edu/millionsong/`, come up with your own machine learning mini-project. Make sure you provide a writeup with the appropriate pictures/graphs.

Here are some suggestions:

- Cluster the lyrics of songs from the musixMatch dataset and see if there are any meaningful groupings.

- See if there are any patterns in the top principal components of musixMatch

- Try implementing a genre classifier based on song metadata