Northeastern University, College of Computer and Information Science

CS 6220 Data Mining — Assignment 3

Due: October 2, 2017 (100 points)

Exploring Data with PCA

For this assignment you will be using a set of images scraped from google images. This dataset contains 40 color images capturing the four different seasons (summer, fall, sprint, winter – 10 images for each season). Each image consists of 62,500 pixels (250 x 250), and each pixel is defined by three color values varying from 0 to 255 that specify the degree of red, green and blue present on that pixel. You can download the dataset here.

Objectives:

1. Employ data reduction techniques such as principal component analysis
2. Visualize an interpret the results of (1)
3. Engineer numeric features from image datasets

Submission:
Submit your GitHub link through Slack as done in Assignments 0 and 1.

Grading Criteria:
Follow the instructions in the pdf, and complete each task. You will be graded on the application of the modules topics, the completeness of your answers to the questions in the assignment notebook, and the clarity of your writing and code.

The idea: PCA with Image Histograms

Your objective here will be to perform dimensionality reduction on the dataset to highlight the similarity between images within some lower-dimensional (feature) space. Specifically, we would like to perform a transformation on the images so that they can be viewed in a 2-dimensional space, which we can then plot and visualize. This can be accomplished by performing principal component analysis (PCA) on some property of the images and retaining only the top 2 principal components. Ideally, we want to apply PCA to a property of the images that can be used to capture some notion of the similarity between images.

Intuitively, since we typically view an image as a collection of pixels, we might consider performing PCA on the set of pixels, reducing an image from a collection of 62,500 pixel values to a new 2-dimensional space. In other words, we would consider each pixel as a feature of an image, and try to transform an image into a new set of 2 features. Ideally, images that have a similar set of pixels should also be similar in this 2-dimensional space (i.e., they should share similar feature values). However, due to the magnitude of this reduction, a significant amount of information would likely be lost in the transformation, meaning that pictures with similar pixel values many not appear similar in the the new 2-dimensional space.

A better approach might be to reduce the histogram of color values. Each color has 256 possible values, resulting in a histogram of 768 (256*3) color values distributed over the entire range of 65,000 pixels. Each value of the histogram is the number of pixels in the image with the corresponding color value. Here, we would consider each histogram value as a feature of an image. By performing PCA on an image histogram, we are only reducing an image from a set of 768 unique histogram values (rather than the 65,000 unique pixel values) to a new 2-dimensional space. As a result of this transformation, images that have a similar histogram should also be similar in this 2-dimensional space.

What to do

The Notebook provided with this assignment here includes functions to compute the histograms and plot the images within the transformed (2-dimensional) space (*load_images*() and *plot_image_space*(), respectively).
There are also functions to generate and plot the color palettes associated with each image (*cluster_image_colors*() and *plot_color_palette*(), respectively); the palettes are generated via (k-means) clustering of the pixel color values, and may be investigated at your own leisurethey are not needed to complete the assignment.
The images can be loaded and the histograms generated by using the provided functions. Please ensure that the directory provided to the *load_images*() function is correct. For example, if you have placed all the images in your base Jupyter Notebook directory in a folder called images, then the (relative) path to the images would be ./images.
Correctly completing the above should provide you with an output similar to what you see below (note that this was generated using a different set of images):

Projection of the Image Histograms into 2 Dimensions

What you need to provide

Your output should contain the following:

Part1 [25pts]:
The PCA projection of the image color histograms in 2 dimensions. Using the provided *plot_image_space*() function. This should be displayed as thumbnail images distributed within a 2-dimensional plot.
You will need to use PCA, which is implemented in scikit-learn. See this link for documentation here.

Part2 [25pts]:
Given this output, respond to the following questions:

1. What does it mean for two images to be close together in this plot? What does it mean for two images to be far apart?

2. Do images corresponding to one of the seasons tend to group together more closely than others? Why might this be the case?

Part3 [50pts]:
Once you completed the first two parts of the assignment, choose one of the following below:

1. Repeat this process while using a different set of images curated by yourself.

2. Repeat this process using a different data reduction method and describe any similarities/differences between that experiment when compared to applying PCA.

Share the visual output of this exercise on slack (#mod3activity) by Sunday Oct 1st (end of the day), and comment on the output of at least one of your peers by the end of the following day (Monday Oct 2nd).