Machine Learning (Pattern Recognition) in Python

In this task you are asked to implement machine learning algorithms in Python and interpret their results in real data (images) and synthetics data.

You can use any python library you want e.g. scikit learn, pandas, OpenCV etc.

The work consists of two questions. You can consult and use any hardware and / or code that is available on the Internet, provided that the source and / or link to website where you obtained information.

You will submit a single IPython Notebook file (Jupiter notebook).

Both the code as well as your answers to comprehension / interpretation questions should be built into the same IPython notebook. You can use cells header to further organize your document.

[Question 1 - Pre-processing, dimensionalization, visualization and image sorting]

Data:

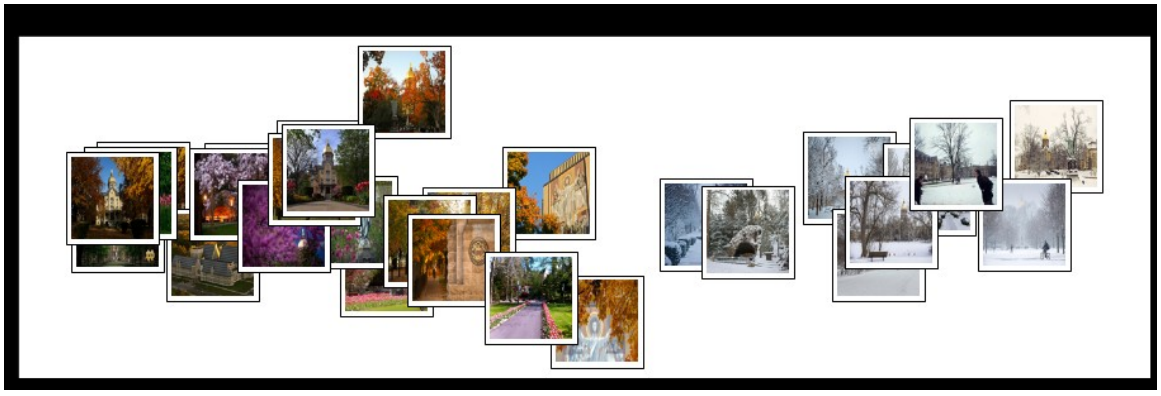
The data set consists of 30 color RGB images set record landscapes in spring (spring), autumn (fall) and winter (winter) (10 images for each season).

The first letter in the file name of each image identifies the time at which the image was recorded, e.g. the image F1.jpg recorded in the fall (fall) while the image W10.jpg was recorded in the winter (winter). Therefore the naming of the files completely determines the category to which it belongs each image. Images are made up of many different pixels. Each pixel consists of three color values ranging between 0 and 255 and determine the brightness of red, green and blue respectively at any point in the image. The data is available in the images.zip file.

Required:

1) Write a loadImages (path) function that takes the path as input to where is the folder of images e.g. loadImages (“C: / images”), reads the images, converts them to 100 x 100 pixels and returns a table 30 column data, where each image is represented as a vector column. The function also returns the categories (labels) to which the different ones belong images encoded in integers (eg 0 for photos captured in winter, 1 for the photos taken in the fall and 2 for those that recorded in the spring).

2) Write a function PCA_ImageSpaceVisualization (X) which takes the input data table, calculates the first two main components (principal components) of the data and displays the data in the first two main components. The function returns a plot in which images are visualized in the two-dimensional space resulting from the projection of the data in the first two main components. The plot is expected to be in the form of:



2.1) What does it mean when images are located near this two - dimensional space that is shown in the above plot? What does it mean when images are far apart? Can we generalize these conclusions for the original image space which is very large in size?

2.2) Images corresponding to one of the seasons tend to be grouped closer together than the rest? Why is this happening?

3) Compare the accuracy of the nearest neighbor classifier (1-NN) and of linear support vector machine (SVM) to the problem of recognizing the season in which an image was recorded. In other words compare the

performance (in terms of classification accuracy) of the above classifiers in the classification of image data into categories winter, spring and autumn.

You are asked to address the classification problem using:

(1) the initials large dimension images in vector format.

(2) low characteristics dimension that you will export through PCA.

3.1) Mathematically define the measure of classification accuracy.

3.2) Use 5-fold cross validation and indicate the average classification accuracy for both classifiers for both large data and low dimensional characteristics.

3.3) How to determine the dimension of the features to be exported through PCA?

3.4) Which classifier has the best performance and why?