ME 592: Data Analytics and Machine Learning for Cyber-Physical Systems

Homework 2

Homework Assigned on February 27, 2025 Homework Due on: March 13, 2025
One submission per group

Motivation

This homework is to provide an experience of Data Preparation involved in Data Analytics for Cyber-Physical Systems.

General Instructions

The dataset and problems for each group are slightly different, but the motivation remains the same. Following are some instructions for all the theme groups. Specific instructions for each group shall be provided in the relevant sections.

1. The final code must be pushed to git before the deadline.
2. Use the discussion board in Canvas in case of any issue.

Expected Outcome

1. A code pushed in git to preprocess each dataset provided.
2. A presentation video explaining your solution approach and results (video duration should be approximately 5 minutes). Submit the video (preferably link to video) and github repo link through Canvas.

.

Ag/Bio Applications and Image/Video Analytics

1. Image pre-processing and transformation The goal is to perform some preprocessing of images obtained for leaves that may or may not be diseased. The following are the tasks you have to perform with the leaf images provided in Agricultural_and_Image.zip :

1. Write code to randomly rotate, shift, scale and warp the image of the patches. Thus generate 100 images.
2. Obtain local patches (of a certain fixed size of your choice, much smaller than the original image size) from all the leaf images
3. Prewhiten (using ZCA whitening) the patches. (Refer to `http://ufldl.stanford.edu/tutorial/unsupervised/ExercisePCAWhitening/` for more information on how to whiten an image)
4. Determine the channel-by-channel distribution of the prewhitened images
5. Determine the channel-by-channel distribution of the original images

2. Image Segmentation The goal here is to extract the 36 subplots of soybean across the three time step data in the zip file. An example is shown in Fig. 1. Write a script to perform the same operation on the given images to obtain 36 separate subplots.

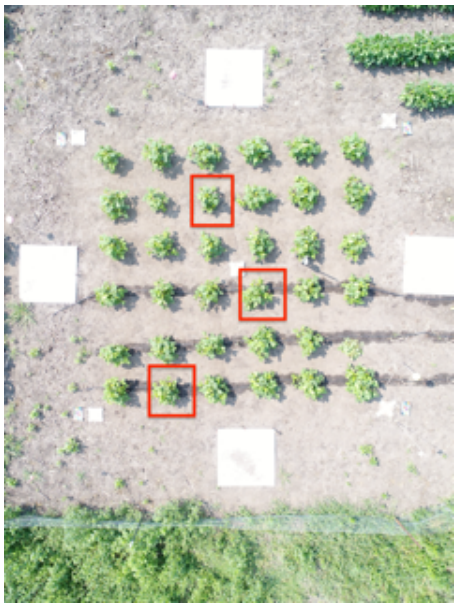

Figure 1: Example of an image highlighting examples of subplots - required for problem 2

Robotics

In robotics application, we often have to pre-process the input images being used by a robot to improve the learning capabilities of the robot. We will explore a few examples of such pre-processing techniques in this assignment. We will use the Grasping Dataset available at `https://www.kaggle.com/oneoneliu/cornell-grasp` to pre-process images so that a robot arm can learn to detect and grasp an object. The dataset contains raw images of the objects, coordinates and labels of grasping rectangles as well as point cloud data of the objects. The tasks to perform are as follows:

1. Download the dataset and overlay the positive and negative grasping rectangles for each image on the raw image using the coordinates of the rectangles provided.
2. Using the PNG images and point cloud data, register the point cloud data to the PNG images to obtain a 4 channel RGB-D image (Hint: Please take a look at the utils directory of the following repo as reference on how to create depth images from point cloud data. `https://github.com/skumra/robotic-grasping/`).
3. Using the RGB-D images above, extract sub-patches of images that are found within the coordinates of the positive grasping rectangles. Now, we will extract 4 channels worth of features from these sub-patches. Convert the first 3 channels of the RGB-D image to YUV color format to get features representing the color and intensity. Next, extract the last channel of the RGB-D image and convert it to a single image representing the depth of the features.
4. Next, apply PCA Whitening to the depth features extracted above to further reduce any bias in the depth features.
5. Visualize the point cloud data using your choice of software.

Scientific Simulations

The dataset in DM.zip contains the following:

1. 64 input geometries made of NURBS surface. Each input geometry contains 3 smesh files for three surfaces which are interacting in the analysis. these smesh files contain the position of the control points.
2. Deformed geometry (at 80th time step and 140th time step of the analysis) for the 64 geometries at 5 thickness and 3 pressure conditions.

Each of the geometry is placed in folder run1-64. Each of the smesh file contains the following:

- First 4 lines talk about the no. of components, the DOF, number of nodal points and something about the post-processing. You might find it easy to skip these 4 lines while parsing through the file.
- After the line 4, x,y and z coordinates of each node is specified. For simplicity, all the geometries are having 17 and 12 nodes in two directions of the surface.

Each of the final geometry is named as result tstep temp pressure geometry, where tstep refers to the time step of the deformed geometry (80 and 140). temp refers to the 5 temperatures (300K, 350K, 400K, 450K, 500K) and pressure refers to 3 pressure conditions (76mmHg, 80mmHg, 84mmHg) and geometry refers to the 64 geometries. Each file contains the final deformation of all the nodes of the three surfaces in the same order as the smesh files (skip line 1 which contains time step and analysis relevant details). Using the following data, perform the following tasks:

1. Create two ordered pairs of (input,output) corresponding to two time steps. Here, input refers to tuple of (geometry, temperature, pressure). Since, each of the geometry contains 17×12 nodes. Construct an array of shape 17×12×3 for each geometry. Shape of one element of the ordered pair may look something like ([[17,12,3],[1,],[1,]], [[17,12,3]]).
2. Consider the input geometry and output geometry to be images with three channels (x, y and z channels instead of RGB channels). Since, the range of x, y and z channel are completely different, we would like to apply PCA whitening to normalize the data. Refer to links below for more information on how to whiten an image

   - http://ufldl.stanford.edu/tutorial/unsupervised/PCAWhitening/
   - http://ufldl.stanford.edu/tutorial/unsupervised/ExercisePCAWhitening/

   You would need to whiten all the input geometries, output geometries of both the time steps.
3. Once, the output images are whitened, vectorize the output images of time-step 80 and use that for plotting t-SNE. While plotting the t-SNE, you would need to create three plots to mark the labels of the data based on geometry, temperatures and pressures. Comment on the data distribution based on the t-SNE results in terms of geometry, temperatures and pressures. You could use any of the implementations available at https://lvdmaaten.github.io/tsne/.

4. Repeat the same process on time-step 140. Comment on any changes in data distribution inference.

Time Series Analytics

The .mat files in the `electricity dataset.zip` contains energy consumption (in Watts) time series for different end uses as well as the main power for a house sampled at 1Hz.

The task is to explore the efficacy of different distance metrics and transformations for time series by computing differences among different end uses and the main power. Perform the following:

1. Perform pre-processing of all the variables (such as normalization and denoising).
2. Use direct Euclidean distance metric.
3. Convert the data to frequency domain using FFT and then do a comparison using euclidean distance metric.
4. Use KL Divergence metric and perform comparison.
5. Compare the data in wavelet transformed space.
6. Use windowed spectrogram to identify motifs in the main power data to detect changes in time-series characteristics.

Comment on your findings of the comparison and about the feature detection.

Attachments

1. Agricultural_and_Image.zip
2. DM.zip
3. electricity dataset.zip