3684 – Advanced Topics in Machine Learning, Spring 2022
Home Assignment #2b – LIME demonstration on image data

Lecturer: Dr. Leon Anavy
Teaching Assistant: Mr. Alon Oring

General instructions:

1. Submission is **individual.**
2. Submission must include python code **and** a written report.
3. You may use external libraries. Specify all required libraries in a proper manner.
4. Your code must be reproducible. Code that will not run will result in a grade reduction.
5. Your report should be clear, coherent, and concise. The report should not exceed 10 pages.
6. Invest thoughts and considerations to the way you choose to present data and experimental results.
7. All figure and plots should include captions, labels and data units. Pay attention to data visualization guidelines.

Assignment tasks:

The goal of this assignment is to demonstrate the LIME method that was covered in class used to explain image classification models.

1. Choose a pretrained image classification model f to be explained. The model will be used as a black box. You only need to be able to classify new images using the model. You can use the following resource: https://pytorch.org/vision/stable/models.html
2. Choose 2-3 images to be classified and explained (x).
3. For each image perform the following:
   a. Get the top 3 classes from the model $f(x)$
   b. Interpretable (simplified) instances:
      i. Generate interpretable versions of the images you chose by either splitting them to tiles or to super-pixels. You can use the CV2 package for that.
      ii. Represent the interpretable instances as binary vectors. The entries of the vector correspond to inclusion/exclusion of the tiles/super pixels $x' \in \{0,1\}^{d'}$
   c. Local dataset generation
      i. Generate a set of random perturbations of the interpretable instances by uniformly choosing which parts to include $z' \in \{0,1\}^{d'}$
      ii. For each generate interpretable instance, generate the corresponding image z and get its label (as a binary classification result for each of the three classes) $f(z)$
      iii. Calculate the similarity of the perturbed instance from the original image $\pi_x(z)$
   d. Fit a local surrogate model g and generate explanations
      i. Fit a linear model with locally weighted loss (using π_x) and L_1 regularization on the generated dataset
      ii. Find and present the set of important features (super-pixels/tiles) for the prediction $f(x)$

Summarize all your work in a scientific/professional report.

Class presentation:

If you have chosen this assignment as your class presentation assignment you are required to prepare a 20 minutes presentation in which you will need to showcase your work. You should cover all aspects of your work in the presentation.