3945 – Advanced Machine Learning, Spring 2022
Home Assignment 3 – Explainable AI

Lecturers: Dr. Leon Anavy, Mr. Alon Oring

In this assignment, you will demonstrate the LIME method that was covered in class to explain image classification models. You will submit a single ipython notebook (.ipynb) file that includes all the code and outputs, as well as a brief report that explains your approach and findings.

Part 0: Choose pretrained image classification model and images to be explained

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) Select 3 images to be classified and explained

Part 1: Choose pretrained image classification model

For each image, x, perform the following:

3) Get the top 3 classes from the model $f_1(x), f_2(x), f_3(x)$
4) Interpretable (simplified) instances:
   a) Generate interpretable versions of the images you chose by either splitting them to super-pixels. You can use the CV2 package for that.
   b) Represent the interpretable instances as binary vectors. The entries of the vector correspond to inclusion/exclusion of the super pixels $x' \in \{0,1\}^{d'}$
5) For each class perform f_i:
   a) 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 generated interpretable instance, generate the corresponding image z and get its label $f_i(z)$
      iii. Calculate the similarity of the perturbed instance from the original image $\pi_x(z)$
   b) 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. Use K-Lasso for feature selection.
      ii. Find and present the set of important features (super-pixels/tiles) for the prediction $f(x)$

Submission Guidelines

- Submit the work in pairs. Only one submission for each pair.
- Your submission should include a single zip file containing:
  - A single ipython notebook (.ipynb) file that includes all the code and outputs.
  - A brief report (1-2 pages) in a pdf format that explains your approach and findings.
- The submitted file should follow the naming convention:
  - 3945_HW##_XXX_YYY.zip
    Where:
    - ## is the assignment number
    - XXX and YYY are your student numbers (IDs)

    For example: 3945_HW1_123456789_987654321.zip

- Make sure to run your notebook from start to finish before submitting to ensure that it runs without errors.
- You may use external libraries. Specify all required libraries in a proper manner.
- Grading will be based on correctness, elegance of solution, and style (comments, naming conventions, etc.)
- Your report should be clear, coherent, and concise.
- All figure and plots should include captions, labels and data units. Pay attention to data visualization guidelines.
- Make sure to use correct ML methodologies and justify your selections (split the data to train/test, tune hyperparameters, report relevant performance measures).