Assignment for "Model explanations using Influence Analysis"

Due: June 20, 2025

In this assignment you will explore the fragility / robustness of some of the ML model explanation techniques discussed in class, specifically those based on Influence Analysis. These methods are designed to provide an explanation for a specific model inference (i.e., predicting for one instance in the test set), and this takes the form of a list of training data points, ranked in order of their relative "influence" on the model when making the inference. In class we have seen the the example of a Titanic passenger (in the test set) who is predicted to survive, where the explanation consists of top-k other passengers from the training set.

In class we have pointed out two potential problems with these explanations. Firstly, that the top-k lists are sensitive to the choice of training and test datasets, as well of model, its hyper-parameters, and the training process. Secondly, that even when we fix all these choices, different influence methods will produce different top-k lists. We suggested the terms "stability" (within one method) and "consistency" (across two methods) to refer to these problems.

Task description.

Setup.
Your task is to design and implement experiments to analyse and quantify stability and consistency, using specific datasets, models, and influence methods. Specifically:

- you will use the well-known Titanic benchmark dataset and corresponding survival prediction tasks: https://github.com/pandas-dev/pandas/blob/main/doc/data/titanic.csv (there are many versions on kaggle. this has nearly 900 records which seems fairly complete). Task:.
- Head to the Influenciae git repo https://github.com/deel-ai/influenciae. You will use the following two methods from the available suite:
    - first-order influence calculator https://colab.research.google.com/drive/1WlYcQNu5obhVjhonN2QYi8ybKyZJl4iY
    - Tracin https://colab.research.google.com/drive/1E94cGF46SUQXcCTNwQ4VGSjXEKm7g21c
- Use the tutorials to guide you through:
    - learning a model using TensorFlow. the Tutorials provide fairly clear code if you are not familiar with TensorFlow
    - generating explanations for specific test data points in the model

Challenge: the code is designed to work with an image dataset, not with tabular data. For convenience, I have included an implementation of the modelling task in TensorFlow, which (1) uses Titanic, and (2) works with both influence methods. Feel free to use this (it's a pdf so you will need to import bits and pieces into your own code).

Experimenting.

- Explore the distribution of each explanation list. Does it have a long tail? how quickly does the influence score drop off? plot a few charts for each of the two methods
- quantify sensitivity and consistency. For this, you will need to identify a generic method way to quantitatively compare two ranked lists, providing a score (hint: **Kendall Tau** is commonly used, but you may suggest others).
    - How do influence lists compare across the two methods? use the list comparison scores
    - are some of the same records in each of those lists? what do they look like and what is special about these records??
    - what happens if you remove the top-k points from the training set and retrain the model? is the model performance (accuracy, AUC) affected?

Deliverables:

- Code in a git repo demonstrating the tasks above,
- a report showing your experiment design and presenting your findings. Your choice of approach and design should be justified.

Guidelines and support You may work independently, or in pairs.

You may contact Paolo: p.missier@bham.ac.uk. but please note: my reaction times may be quite long.