Student Name: Faris Hajdarpasic
Student ID: 50059467
Student Name: Ajay Ragh
Student ID: 50065136

Assignment 01

Explainable Machine Learning

1. **What is the meaning of explainable ML?**
   Explainable ML refers to the methods used to interpret and explain different ML models. These methods enable us to get insights about different information related to data itself, input-output relationship, role and impact of specific parts of model architecture, etc. It's broad field, and it does not have 100% general/templated approach, and it can also be defined subjectively.

2. **Why do we need to understand the decision-making mechanism of a ML algorithm?**
   We need to understand decision-making mechanism of ML model to be able to:
   - Understand role of specific parts of model architecture
   - Modify correctly model architecture with respect to goal
   - Interpret intermediate results in model architecture (e.g. results between layers)

3. **What are the differences between explainable models and interpretable models?**

   Interpretable models are models in which we can provide information about relationship between input features and output prediction. Simply models for which we can interpret results. Two interpretable models are linear regression and decision tree.

   Explainable models are models in which we can provide information about relationship between input features and output prediction based on some background domain knowledge. Simply, models for which we can intepret results and explain results using domain knowledge. In short, explainability is equal to the interpretability plus domain background knowledge.

   Explainable model example: The relation between inputs and outputs are analysed via background knowledge for example lets say we have a network that identifies tree age by looking at images of tree cross-sections, by background knowledge we know that this can be achieved by counting the tree rings in the cross section. So based on this knowledge we can analyse the network as say how it is identifying tree rings from the image and counting them.

   This same idea can be used for the case of lets say a clustering algorithm that is using decision trees, in this case lets say there is a parameter that cluster points should have minimum distance of 1 meter between them and there should bet at least 3 points to create a cluster. Then lets say the model is a decision tree architecture with our previous knowledge about parameters for a cluster we can easily explain that at each decision node the model is checking whether a given point satisfies the given conditions.

4. **What are the differences between global explainability methods and local explainability methods?**

   Local explanation explain one specific decision. Given one input, explain output. This gives characteristic of one specific sample. Two short examples:

   - Given image of whale, why that whale was classified as whale with specific ID
   - Given properties of the house, why house price was estimated that much

   Global explanation explains entire model behaviour. By calculation prototype, we can explain whole dataset. Prototype is sample that gives maximal output. Based on that prototype, we can later explain why model gave specific output for some input.

   Universal example: Binary classification (cat-1 and dog-0):
   Given the prototype image of a cat (image for which model would gave maximal score as output - 1.00) we can explain for which images model would give which score, and therefore we could make conclusion about whole dataset i.e. which images would be classified as cat, and which as dog.

5. **What are the differences between model-agnostic methods and model-specific methods?**

   Model-agnostic explanation explains how model maps input to the output, without knowing anything about the model architecture (i.e. model itself is black-box).

   Model-specific explanation explains how certain parts of model (or whole model) contribute to specific output given specific input. In this case, model cannot be viewed as black-box.

6. **What are the differences between ad-hoc and post-hoc methods?**

   Ad-hoc and post-hoc methods are two different approaches that is aimed at achieving better interpretability and transparency about the inner-working of deep-learning/machine learning models.

   Ad-hoc method as the name suugests are methods that are aimed at making the model itself more interpretable during the training phase. As a result of this we are able to have better understanding on how exactly the model learned while training on the data. To achieve this we prefer more understandable or interpretable learning architectures like decision trees, regression models etc.

   As a result of this the models themselves are simpler and computationally efficient, though as a consequence these kinds of architectures can't perform complex tasks like for example the image generation models that we see nowadays.

   Post-hoc methods does not take into account the model itself and its training. These are also known as model agnostic or black box methods. Since these methods are used externally on the entire model itself to understand its working without modifying it.

   So because of this, post-hoc methods unlike ad-hoc ones can be used to understand different models irrespective of their architectures. These are useful for understanding how complex models whose inner workings are harder to understand works.

7. **What are differences between data modality agnostic methods and data modality specific methods?**

   Data modality based techniques can be agnostic or specific techniques. The idea behind data modality based techniques is to understand and explain the working of a model based on the type of input data being processed.

   Data modality agnostic methods are designed to be broadly applicable irrespective of the type of data that is being processed. As a result these techniques are much more flexible and can be used to explain a wide variety of techniques.

   Data modality specific techniques are designed to explain the working of a network based on the modality of the data being used for example a segmentation network that uses depth and RGB images inputs. These techniques are tailored specifically to particular types of data or domains. As a result of this tailoring we can expect much more precise explanations about the workings of the network.