University of Fribourg | *Lecturer: Dr. Andreas Fischer*
Department of Informatics | *Teaching Assistant: Michael Jungo*

Spring Semester | Pattern Recognition

Exercise 5 - 29th of April 2024

Fourth Group Task – Molecule Classification

Deadline: 20th of May 2024 (end of day)

In this exercise you will be developing a machine learning approach for molecule classification. Given a molecule as a graph, where an atom is a node and the covalent bonds the edges, the goal is to identify whether the molecule is *active* or *inactive* against HIV. This is a graph classification task, where graphs are compared with each other using the graph edit distance (GED) to determine the most likely label.

Dataset

You can find the dataset on ILIAS in `exercises/5-Molecules`.

In the dataset, you will find the molecules as graphs defined in an XML file (`gxl`). There is quite a bit of information about the molecules in the graph, but you can reduce the scope by only using the nodes with their labels (chemical symbols) and the (unlabelled and undirected) edges between the nodes. Therefore, a graph will consist of atoms and the bonds between them. If you want to incorporate additional information, feel free to include them, but make sure the graph edit distance is calculated accordingly.

The file structure is as follows:

- `gxl/`: XML files with the graph definitions for each molecule.
- `train.tsv`: Ground truth for all training samples, where each molecule is classified as either *active* or *inactive*.
- `validation.tsv`: Ground truth for all validation samples.

Data Format

The molecules are given as graphs defined in `gxl` files (Graph eXchange Language). Each graph contains `<node>` tags with an `id` and `<attr>` tags with additional information, such as the chemical *symbol*. The edges are given as `<edge>` tags between two node ids. Even though

the edge definitions are given in a directed manner, with `from` and `to`, they are considered to be undirected, and only given once (not separately in both directions).

Example

```
<?xml version="1.0"?>
<!DOCTYPE gxl SYSTEM "http://www.gupro.de/GXL/gxl-1.0.dtd">
<gxl>
  <graph id="molid624331" edgeids="false" edgemode="undirected">
    <node id="_1">
      <attr name="symbol">
        <string>C </string>
      </attr>
      [...]
    </node>
    [...]
    <edge from="_1" to="_2">
      [...]
    </edge>
    [...]
  </graph>
</gxl>
```

Expected Submission

- Access to your Git(Hub) repository so that we can inspect your code.
- Short report in your Git repository (Markdown or PDF) with your results.