# Machine Learning — Programming Assignment

June 2022

## 1 Problem definition and data

Healthcare is one ofthe major sources of applications for machine learning. It includes interesting problems ranging from image recognition to time series analysis. Many healthcare applications involve some form of natural language understanding.

Here we want to build a system able to automatically classify medical reports. To this purpose a data set of reports written by doctors has been collected. Each report describes a single case and is stored as a text file with the first line summarizing the case, and the others providing more detailed information.

The following is, for instance, one of the reports in the data set:

```
CT of chest with contrast.  Abnormal chest x-ray.

EXAM: , CT chest with contrast.,HISTORY: , Abnormal chest x-ray, which
demonstrated a region of consolidation versus mass in the right upper
lobe.,TECHNIQUE: ,Post contrast-enhanced spiral images were obtained
through the chest.,FINDINGS: ,There are several, discrete, patchy
air-space opacities in the right upper lobe, which have the
appearance most compatible with infiltrates.  The remainder of the
lung parenchyma is clear.  There is no pneumothorax or effusion.  The
heart size and pulmonary vessels appear unremarkable.  There was no
axillary, hilar or mediastinal lymphadenopathy.,Images of the upper
abdomen are unremarkable.,Osseous windows are without acute
pathology.,IMPRESSION: , Several discrete patchy air-space opacities
in the right upper lobe, compatible with pneumonia.
```

The classes to be identified correspond to the medical specialty to which the case belongs. The classes considered are:

- cardiovascular / pulmonary;
- consult;

- gastroenterology;
- general medicine;
- neurology;
- orthopedic;
- radiology;
- surgery.

The class can be obtained from the filename, which follows the format:

`<class>-<id>.txt`

where `<id>` is a numerical code.

The data set has been divided in a training set 2731 reports, a validation set of 300 reports, and a test set of 300 reports.

## 2 Assignment

We want to build a classifier that is able to predict the medical specialty of a given report. For the programming assignment you are expected to:

1. analyze and comment the data;
2. design and implement a suitable data pre-processing procedure;
3. implement, train and evaluate one or more classification models;
4. use suitable data processing and visualization techniques to analyze the behavior of the trained model(s); in particular try to determine if the short description is enough for a good prediction and how much the detailed prediction increases classification accuracy.

All the above should be implemented as scripts in the Python programming language. Any machine learning library (included `pvml`) can be used. Data and code developed during the course can be used if needed.

## 3 Report

Prepare a report of three to five pages documenting all your work. Provide detailed instructions on how to reproduce the results. The report must be in the PDF format. Include your name in the report and conclude the document with the following statement: "I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher."

Make a ZIP archive with the report and the Python scripts, and submit it from the course web page. To keep the size of the submission manageable, do not include files containing the original data, the features etc.