Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
Winter Call, A.Y. 2023/2024

Last update: January 5, 2024

1 Project dates

Start date: January 5, 2023 at 23:59 (CET)
Due date: January 26, 2023 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Within the field of particle physics, a long-standing task of interest is detecting the positions where particles (e.g., electrons) pass in their trajectories. Some sensors are capable of detecting the passage of these particles and, following that, the position of the particles themselves. One such sensor is the RSD (Resistive Silicon Detector). This sensor has a 2-dimensional surface within which it can detect the passage of particles. Figure 1 shows an example of such a detector.

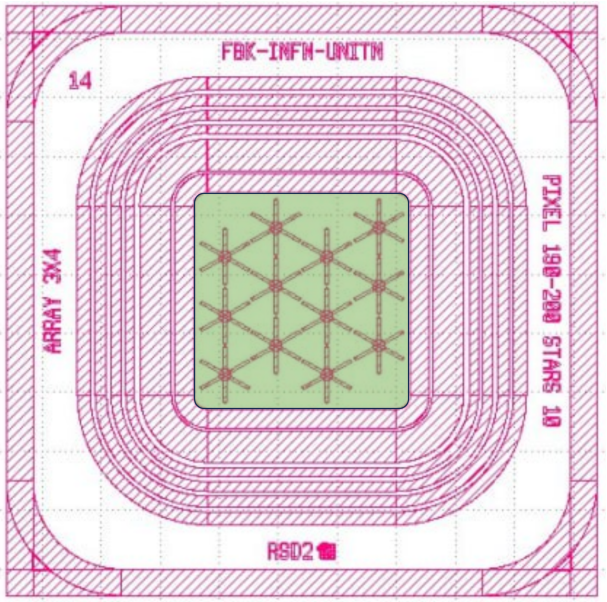

Figure 1: Example of a detector. The detector can observe a particle passing through the green area by measuring signals on the asterisk-shaped pads.

When a particle passes through the sensor (green area in Figure 1), various metallic *pads* are used to measure a signal. In the specific sensor of interest, the pads have a "snowflake" (or star, or asterisk) shape. In particular, in Figure 1 there are 12 pads within the sensor.

The passage of a particle through the sensor is referred to as an *event*. For each event, each pads measures a signal: the properties of this signal (the inputs to the problem) are correlated with the *position* of the passing particle. Since the surface is 2-dimensional, the position of the particle (your target) is defined as a pair of (x,y) coordinates.

The goal of this project is to build a data science pipeline that predicts for each event, given as inputs the characteristics of the signals measured by each pad, the target (x,y) coordinates where the particle of interest passed.

2.1 Dataset

The dataset is comprised of 514,000 events: 385,500 events for the development set and 128,500 for the evaluation. Each event has been conducted in a controlled setting, where the passage of a particle has been enforced (i.e. the (x,y) coordinates are known).

For each event, the signals measured by all pads have been recorded, and some relevant features have been extracted. An example of a signal measured by one of the pads, for a given event, is the one shown in Figure 2

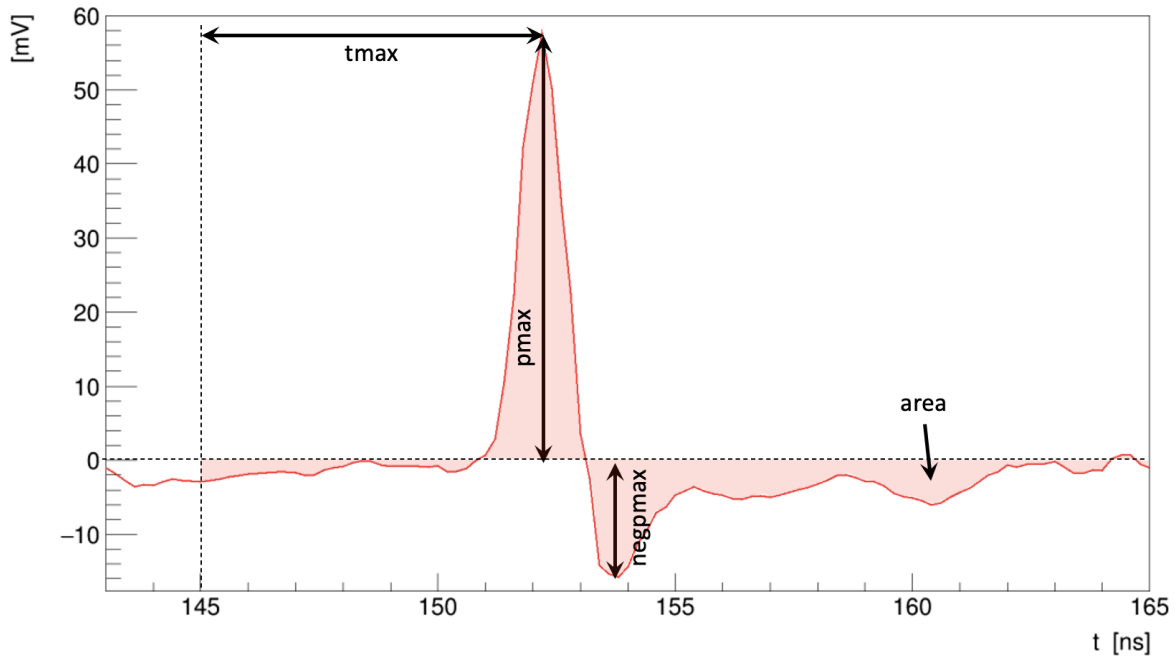

Figure 2: An example of a signal measured by one of the pads of a sensor, for a given event. The features extracted and considered for this study (*pmax*, *negpmax*, *area* and *tmax*) are shown.

For each signal measured by each of the 12 pads, some features are extracted and comprise the dataset. In particular:

- `pmax[0]`, `pmax[1]`, ... `pmax[17]`: the magnitude of the positive peak of the signal, in mV
- `negpmax[0]`, `negpmax[1]`, ... `negpmax[17]`: the magnitude of the negative peak of the signal, in mV
- `tmax[0]`, `tmax[1]`, ... `tmax[17]`: the delay (in ns) from a reference time when the positive peak of the signal occurs

- `area[0], area[1], ... area[17]`: the area under the signal
- `rms[0], rms[1], ... rms[17]`: the root mean square (RMS) value of the signal

Notice that 18 readings of each features are provided for each event, whereas only 12 pads are present in the sensor. This occurs because of hardware constraints in the data acquisition phase: a subset of the 18 features, as such, does not contain actual readings but rather noise.

For each event, the (x, y) coordinates (in μm) are the two targets to be predicted (`x` and `y` columns in the dataset).

Info: If you visualize the data, you may notice that some areas of the sensor are not covered by any event. That occurs because, at those coordinates, either pads or wires used to read the signals from the pads (due to their reflective properties) are present.

The dataset is located at this URL.

Within the archive, you will find the following elements:

- **development.csv** (development set): a comma-separated values file containing the 385,500 events for the development set. This portion has the x, y coordinates to be predicted for each event. These two are the target values to be used to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the 128,500 events corresponding to the evaluation set. This portion does not contain the (x,y) coordinates.
- **sample_submission.csv**: a sample submission file.

2.2 Task

You are required to build a regression pipeline to predict the (x,y) coordinates, in μm. Do note that this is a multi-output regression problem (i.e., you are required to predict two separate outputs for each input).

Do note that, in scikit-learn, many regression algorithms implemented already support multiple outputs by default. Other algorithms instead can instead be extended by using a MultiOutputRegressor.

2.3 Evaluation metric

Your submissions will be evaluated in terms of average (Euclidean) distance of your predictions from the targets. In other words, for a single event with coordinates (x_1, y_1) and prediction $(\hat{x}_1, \hat{y}_1)$, the distance is computed as $\sqrt{(x_1 - \hat{x}_1)^2 + (y_1 - \hat{y}_1)^2}$.

For all n events $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ with predictions $(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2), ..., (\hat{x}_n, \hat{y}_n)$, the metric used will be:

$$d = \frac{1}{n} \sum_i \sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{1}$$

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a `CSV` file formatted as follows:

```
Id,Predicted
1,521.0|654.2
2,433.2|340.0
3,320|440.2
4,767.0|412.4
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the `Id` of the corresponding record in the Evaluation set. It corresponds to the column `Id` in the evaluation CSV file.
- the `Predicted` the (x,y) coordinates predicted. Note that both coordinates should be specified as a part of this field and must be separated by a "pipe" character (`|`), as `x|y` (e.g. `521.0|654.2`).

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to `lorenzo.vaiani@polito.it`. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_winter_2024**.

Info: A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing `.zip` extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.

5 Fill in the LLM usage form

As discussed in the exam rules, adoption of Large Language Models (e.g. ChatGPT) is allowed for the production of the report (**not** for the implementation of the solution). Each team **must** provide information about whether they used, and to which extent they did, LLM-based tools.

To do so, please fill in this form by the due date of this project. Failure to do so will result in a void project.

Warning: This is an additional requirement that was not required in past years. Make sure you remember to fill in the form by the due date, or your project will not be considered valid!