Data Science assignment

Welcome to your Diamontech data science assignment. Below you can find some notes, a description of a dataset and a few tasks we'd like you to tackle.

Notes

- We have been trying to strip away any Diamontech-specific domain knowledge. Nevertheless, should you have doubts, do not hesitate to ask.
- We value process and reasoning over cutting-edge performance.
- We value clean and readable code.
- We are aware your time is a valuable and limited resource. Please set yourself trade-offs and be prepared to discuss them afterwards.
- Questions are meant to hint directions. They will be the likely starting point for discussing afterwards.

Dataset

The dataset we ask you to analyze is `dataset.zip`. It contains noisy synthetic absorbance spectra associated to glucose concentrations (in mg/dl). An absorbance spectrum is the result of a measurement done with our device, and is defined over a frequency continuum, that has been interpolated for your convenience. A way to span this frequency continuum is via wavenumbers. The predictive task is to infer the glucose concentrations from the absorbance spectra.

The dataset is clean, you don't need to go hunting for witches. The absorbance spectra in the dataset are not meant to simulate those of human tissues, rather of simple water solutions with glucose and another couple of constituents at varying concentrations. Every row in the dataset refers to a single (synthetic, i.e. simulated) measurement.

Tasks

- Predict glucose concentrations from absorbance spectra, using an algorithm that, according to your judgment, fits the problem.
- Prediction performance should be expressed in terms of MARD, defined as mean ARD over all measurements. ARD, which stands for absolute relative difference, is defined as in the formula in Figure 1 - for a measurement `t_k`, `y_ref` being the true glucose value while `y_CGM` the predicted. What's the advantage of using MARD instead of RMSE? It's difficult to interpret MARD values in absolute terms without domain knowledge.

Nevertheless, how would you claim your model is not predicting random values, but actually learning from the data?

$$ARD_k = 100\% \cdot \frac{\left| y_{CGM}(t_k) - y_{ref}(t_k) \right|}{y_{ref}(t_k)}$$

Figure 1: ARD

- A colleague with domain knowledge advises you to improve the performance of your model. Which strategy would you follow? According to the angle from which you tackle the problem, you could develop an intuition from the absorbance spectrum of pure glucose, displayed in `glucose-absorbance.png`, given that the noise affects every wavenumber "the same way".
- Absorbance values in the dataset are taken at specific wavenumbers. This is not only done for convenience of model computation, but needs to happen in a more aggressive fashion for development constrains: QCL lasers are manufactured to be able to span a handful of wavenumbers only. Based on the dataset at hand, which top 5 wavenumbers would you communicate to the manufacturer of the QCL lasers?
- Change of scenario. Let's assume that the measurements in the dataset are not obtained via simulation, but are the results of your colleagues' work in the optics lab, which produced the samples manually and measured them. They won't be able to produce such amount of data for every experiment and thus ask you how much data you need for your algorithms. What do you answer them, based on the dataset at hand? Would this change your algorithm selection?
- You are given the chance to attend a conference and learn with excitement about the powerful predictive power of boosting algorithms, especially boosted trees. Apply such algorithm to the provided dataset. Do you experience any advantages? What are the potential pros and cons, especially for more complex datasets?