Mini Project 3: Biomass characterization through NIR spectra

The objective of the present project is to evaluate machine-learning approaches for biomass moisture determination by NIR spectroscopy. The study includes a dataset of NIR spectral data and reference biomass moisture determination according to standardized laboratory methods. Most importantly, the complex chemometric approaches consisting of individual methods for data pre-processing, wavelength selection, and machine-learning regression are employed and compared according to the proposed evaluation methodology. The ensuing results are presented in the form of the coefficient of determination (R2) and the root mean squared error (RMSE) computed from the selected cross-validation round (RMSECV).

The dataset was obtained by experimental measurements on solid biomass samples (random blends of pine and spruce wood chips, bark, forest residues, and sawdust) collected from biomass processing facilities. Spectral data were acquired using a FT-NIR spectrometer. Near-infrared spectra was collected on samples moving on a turntable at 1 m·s-1 and recorded as relative absorbance. The acquisition parameters are summarized in Table 1.

Table 1: Parameters of the NIR dataset

Parameters	**Value**
Recorded spectral range	834-2500 nm (12000 – 4000 cm^{-1})
Number of measurements/sample	5-7
Spectral resolution	16 cm-1
Number of data points in each spectra	1037
Number of tested samples	125

Task 1: Load the dataset and explore the data. Perform data analysis and pre-processing, find outliers, and discuss their usage. Explore different pre-processing techniques to clean the spectral data (Advanced Preprocessing, ei pvm; Savitzky-Golay, ei pvm).

Task 2. Use PLS (Partial Least Squares), SVR (Support Vector Regression), and ANN to build an ML solution(with the three types of approaches) for predicting the moisture content of the biomass selection.

Task 3. Present the results of the two ML algorithms as scores of R2 and RMSE with cross validation, applying on each ML technique different pre-processing approaches.

Task 4. Write a report including the results of your work, focusing on the following sections:

- Introduction
  - What is the problem you are trying to solve? What is the relevance?
- Data analysis
  - What variables exist in the dataset, what values are in the feature vectors, and what issues arise with this type of dataset?
- Pre-processing techniques
  - What data preprocessing techniques are used in the chemometric approaches, and what are their characteristics? Why are they used?
- Modelling approaches
  - PLS(characteristics, usage scenarios, benefits)
  - SVR(characteristics, usage scenarios, benefits)
  - ANN(characteristics, usage scenarios, benefits)
- Results
  - What are the results in terms of R2 and RMSECV in different combinations of processing techniques and ML methods
- Conclusions
  - What are the results achieved with this work? Which ML approach has better accuracy, and what type of pre-processing technique gives better results? What could be done better and what were the bottlenecks?

You need to submit your Python code (preferably in a Jupyter notebook or Google Colab notebook) along with a written report.

Bibliography

Advanced Preprocessing. (n.d.). Retrieved from wiki.eigenvector: https://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Sample_Normalization

Savitzky-golay. (n.d.). Retrieved from https://nirpyresearch.com/savitzky-golay-smoothing-method/