Mapping IR absorbance in wheat seeds to crucial properties using machine learning

UPPSALA UNIVERSITET

Johannes Bohlin
johannes.bohlin.3852@student.uu.se

Ellen Lindgren
ellen.lindgren.8023@student.uu.se

Oskar Lundberg
oskar.lundberg.1178@student.uu.se

Ruiyun Wang
ruiyun.wang.1761@student.uu.se

Supervisor
Carl Nettelblad
carl.nettelblad@it.uu.se
Department of Information Technology and Department of Cell and Molecular Biology, Uppsala University

Course Coordinator
Maya Neytcheva
maya.neytcheva@it.uu.se
Department of Information Technology, Uppsala University

Project in Scientific Computing, 2020

Introduction

Certain properties of the wheat seeds and the bread are crucial to the outcome when baking bread. If a machine learning approach could be used to learn these properties then both time and money would be saved in comparison to the methods used today.

Data

The data consists of NIRS absorbance data for 100 wavelengths from 288 samples and data for the 23 properties we want to be able to predict. Due to noise in the measurements, data preprocessing is needed which is illustrated below.

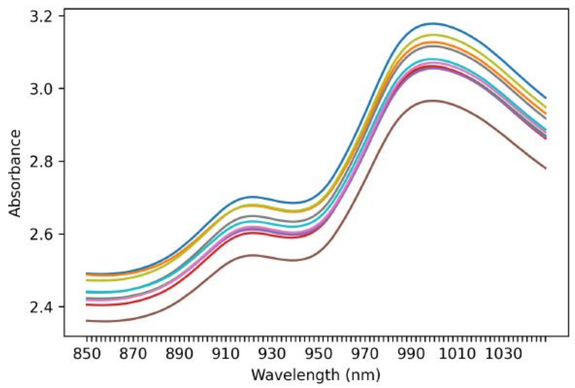

Figure 1: The raw NIRS absorbance data for 10 random samples.

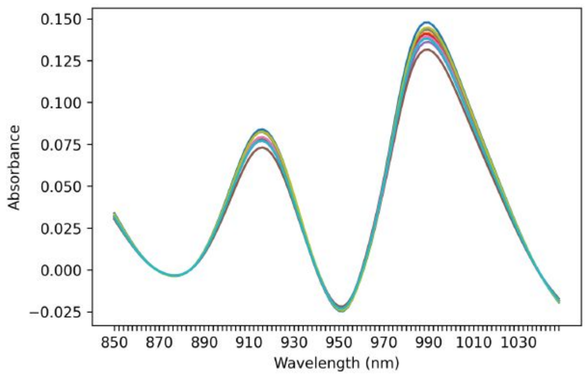

Figure 2: The data after baseline removal for the same 10 samples.

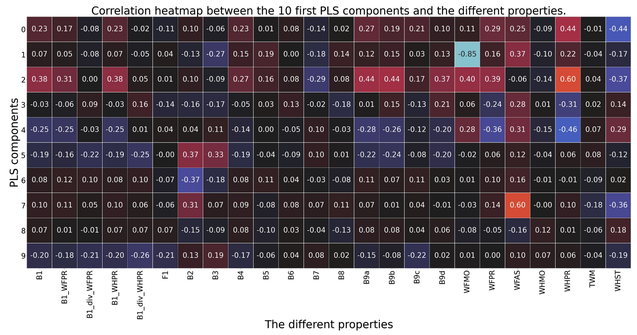

Figure 3: A Pearson correlation heatmap to illustrate the properties linear correlation to the data, after PLS dimensionality reduction.

Method

The heatmap shows that only a few properties have a linear correlation to the data. Therefore, three different linear models were chosen to investigate which performs best. To investigate if we can predict non-linear relationships, neural networks were implemented. Also, an artificial property was created from the data to test if our Neural Networks structures could predict a property with complex non-linear correlations.

Linear models and dimensionality reduction

The linear methods implemented are Linear Regression and two dimensionality reduction methods, PCA and PLSR which respectively maximizes the variance and covariance of the data and reduces the problem into fewer dimensions.

Neural networks

Neural networks consists of layers of neurons which are connected by weights that are learned by calculating how to minimize the cost function. Neural Networks were chosen since they are very flexible and can learn complicated non-linear relationships.

property	Dense 8,16,16 CV	Dense 8,16,16 Validation	Dense 16,32 CV	Dense 16,32 Validation	PLSR CV	PLSR Validation
B1	0.175	0.150	0.098	0.001	0.247	0.140
B1_WFPR	0.082	0.432	0.028	0.391	0.161	0.300
B1_div_WFPR	-0.109	0.028	-0.528	-0.049	0.019	0.030
B1_WHPR	0.120	0.139	0.033	0.389	0.2401	0.131
B1_div_WHPR	-0.168	0.058	-0.409	-0.331	0.026	-0.018
F1	-0.025	-0.047	-0.229	-0.142	-0.070	-0.175
B2	-0.402	0.14	-0.724	-0.267	0.393	0.242
B3	-0.251	-0.107	-0.613	-0.341	0.178	0.033
B4	0.121	0.308	0.025	0.234	0.182	0.284
B5	-0.665	-0.219	-0.523	-0.912	0.019	-0.109
B6	-0.445	-0.281	-0.636	-0.748	-0.052	-0.006
B7	-0.222	-0.381	-0.403	-0.520	0.084	0.276
B8	-0.421	-0.123	-0.513	-0.084	-0.042	-0.055
B9a	0.332	0.326	0.320	0.249	0.358	0.332
B9b	0.234	0.062	0.224	-0.170	0.323	0.065
B9c	-0.295	-0.269	-0.843	-0.852	0.031	0.033
B9d	-0.110	-1.622	0.394	0.669	0.218	-0.062
WFMO	**0.955**	**0.977**	**0.912**	**0.964**	**0.978**	**0.979**
WFPR	**0.961**	**0.975**	**0.826**	**0.922**	**0.996**	**0.933**
WFAS	**0.514**	**0.617**	0.338	0.357	**0.800**	**0.797**
WHMO	-0.524	-0.400	-0.306	-0.413	-0.010	0.037
WHPR	**0.898**	**0.937**	**0.868**	**0.918**	**0.925**	**0.951**
TWM	-0.251	-0.341	-0.152	0.269	-0.036	-0.040
WHST	0.329	0.669	0.313	0.351	**0.618**	**0.729**
Artificial property	**0.882**	**0.891**	**0.815**	**0.848**	-0.010	-0.020

Table 1: R2-values, where a score close to 1 indicates a good model, for both cross-validation and on the validation set for two NN structures and PLSR (the best performing linear model).

Figure 4: Predicted vs observed values on the validation set for WFMO (left) and B1 (right).

Result

In summary, the regression models managed to predict 5 out of the total of 23 properties, using the NIRS absorbance data. These 5 properties have a linear correlation to the absorbance data. The implemented neural network could predict the non-linear artificial property. However, the network did not manage to predict any property that did not have the linear correlation.

Conclusion

Only 5 properties could be predicted. The lack of results of the non-linear properties is because either that the model is not refined enough or that there simply does not exist a connection between the data and the properties.