Data Challenge

Data Scientist - Ferring Pharmaceuticals

Salvatore Prioli
Ph.D and MSc in Computational Chemistry

University of Southern Denmark
Universita degli Studi di Siena

(+45) 50312038
salvatore.prioli@gmail.com
prioli@sdu.dk

June 24, 2022

Data set description

In the data set provided can be found:

- **The training data**: a collection of 500 samples for which *yref* has been measured along a spectrum in the interval 1100-2498 *nm* (2 *nm* stride)
- **The validation data**: a collection of 80 samples for which *yref* must be determined based on 400 repeated spectra in the interval 1118.19-2424.92 nm (6.28nm stride)

The training set and the validation set are sampled at different wavelengths.

Defining a Strategy

Arbitrary assumptions

Based on the information on the provided data, I make three assumptions to implement my strategy:

- *yref* is obtained through a spectrophotometer operating in the near infrared region (NIR)
- *yref* is probably the response of ingredient/s concentration or mixing/blending level of two or more ingredients. It is likely that these properties depens on a small interval of the spectrum (assuming mimimum signal overlap)
- Typically, linear laws (such as Lambert-Beer) govern the relation between absorption and response (excluding overtones occurring in IR absorption)

Defining a Strategy

Step by Step Operations

- Prefer supervised simple liner regression models (*yref* is continuous):
- Proceed by enhancing the sophistication of the model until good performances are obtained (starting from Linear Regression up to Ensemble models)
- Simple linear models can better manage collinearities and provide a good level of variance (avoid overfitting)
- Redundant wavelenghts can be excluded if have low correlation with *yref*.
  - Various models with different number of feautures will be trained.
  - These models will be more solid to be used with different instruments (with different sampling strides) and faster to train in the future.

Data Preprocessing

Check the integrity of the training data set:

- Null values (NaN): not present in the training data set (excluding 'Unnamed x' columns).
- Calculation of the data set statistic (mean median difference, skweness, kurtosis):
  - mean median difference, skweness, kurtosis: values in the range of a Normal (Gaussian) distribution – no transformation needed.
  - outliers are searched using the Zscore:
    - removing the outliers do not make significant change in the statistic
    - removing outliers do not improve the performance of the models (not reported)

Study of the Distribution: outliers

- The data follows a normal distribution
- The outliers have no effect on the models performance
- No further reason to remove outliers

The outliers are not removed from the CAL_500 data set

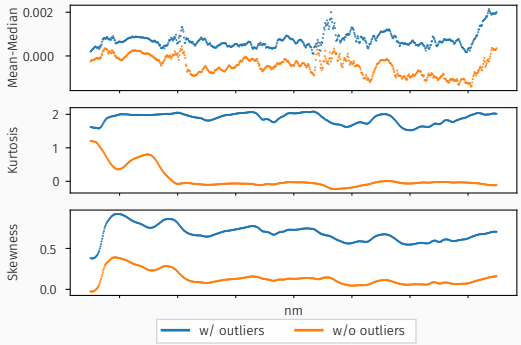

Figure 1: Statistics for each wavelenght (nm) with and without outliers (blue and orange dots respectively).

Model definitions

All the calculation have been performed with python (refer to nb for libraries version)

- Simple Linear regression (LR)
- Ridge Linear regression (LR_R)
- Support Vector regression (SVR) (with grid search for best parameters)

Other methods tried (reported in the nootebook, not discussed further in this presentation because of the poor performance and/or the longer training time)

- Random Forest regressor
- Bagging regressor
- Gradient Boosting regressor
- StackingRegressor (with Ridge and Support vector regression)

Model result: using all the feautures available

	LR	LR_R	SVR
MAE	6.711606	1.005797	0.300109
MSE	89.500996	1.793347	0.167567
RMSE	9.460497	1.339159	0.409350
R2_score	-32.043367	0.337904	0.938135
Wall Time (s)	0.262228	0.049515	0.456951

Model with low error (MAE, MSE, RMSE) and high R^2 are preferred.

- low error - high confidence predicting testing data; high R^2 - high total variance explained by the model
- SVR (with grid search) is the best model (probably due to the underlying linear law govening the relation reinforced by the supporting vectors)
- Other important metrics are AIC, AICc, BIC (Singh et al., 2021)

Selecting most important feautures

Recalling the (arbitrary) assumption that the property depends on a small frequency (wavelength) interval:

- The most important feautures are searched ranking them by correlation with *yref*
- To exclude redundant feautures, the models are trained with increasing number of most important feautures (starting from 100 up to 700 with 50 stride)
- Good compromise to avoid overfitting

Because of computer power limit (calculations performed on a ”slow” macbook air):

- Not possible to go below 50 as step size
- Gridsearch of C and gamma parameters for SVR for every stride not performed
- K-fold cross validation (suggested when performing these calculations) not considered

Model result: with KBest Feautures

	LR	LR_R	SVR
MAE	0.295506	1.005797	0.300560
MSE	0.155953	1.793347	0.168167
RMSE	0.394908	1.339159	0.410082
R2_score	0.942423	0.337904	0.937913

Based on the former considerations:

- SVR is still the best models (low errors and high R^2)
- The accureacy can be improved by tuning the SVR parameters
- A total of 650 features are used (Kbest)

Therefore, the SVR model will be used to obtain the final values

Prediction

Is not possible to make prediction if the number of feautures available in the prediction set is different from the number of feautures used to train the model. For this reason:

- The common wavelengths between the prediction data set and training data set have been selected (considering only integer part – i.e. 1124 in CAL500 is considered the same as 1124.47 in VAL_HSI)
- 101 common feautures have been found
- The model is trained again in the same way as found with the complete training data set
- The predictions are obtained for each of the 80 samples averaging all the avalilabe 400 measures

The final predictions can be found in the `Predictions_SalvatorePrioli.csv` file.