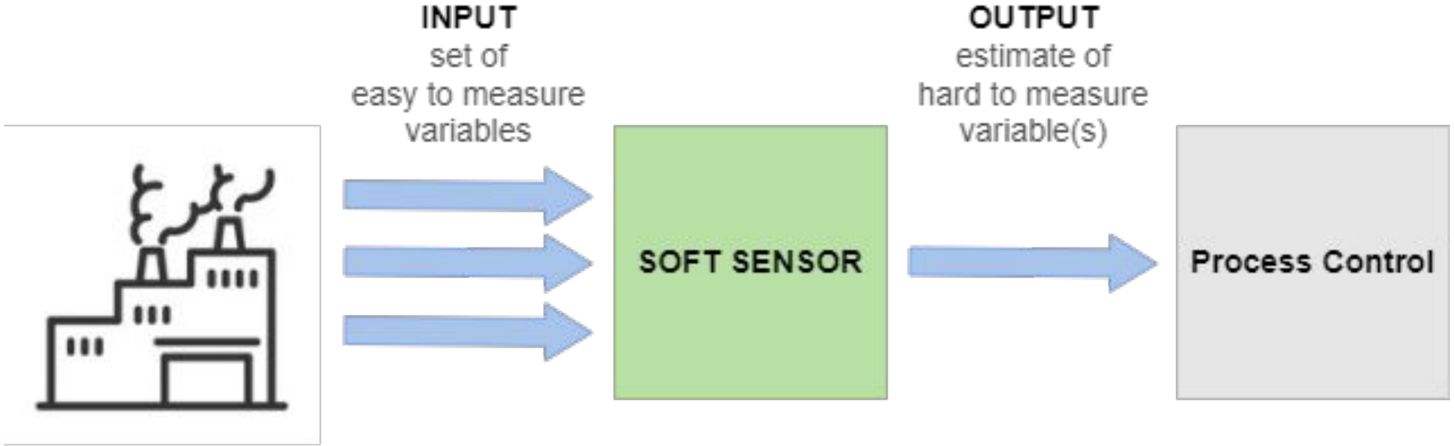
Soft Sensors

Técnicas Avançadas de Análise de Dados
2022/2023

Joana Leiria, 2018274420
Rafaela Carvalho, 2018277699

1290
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

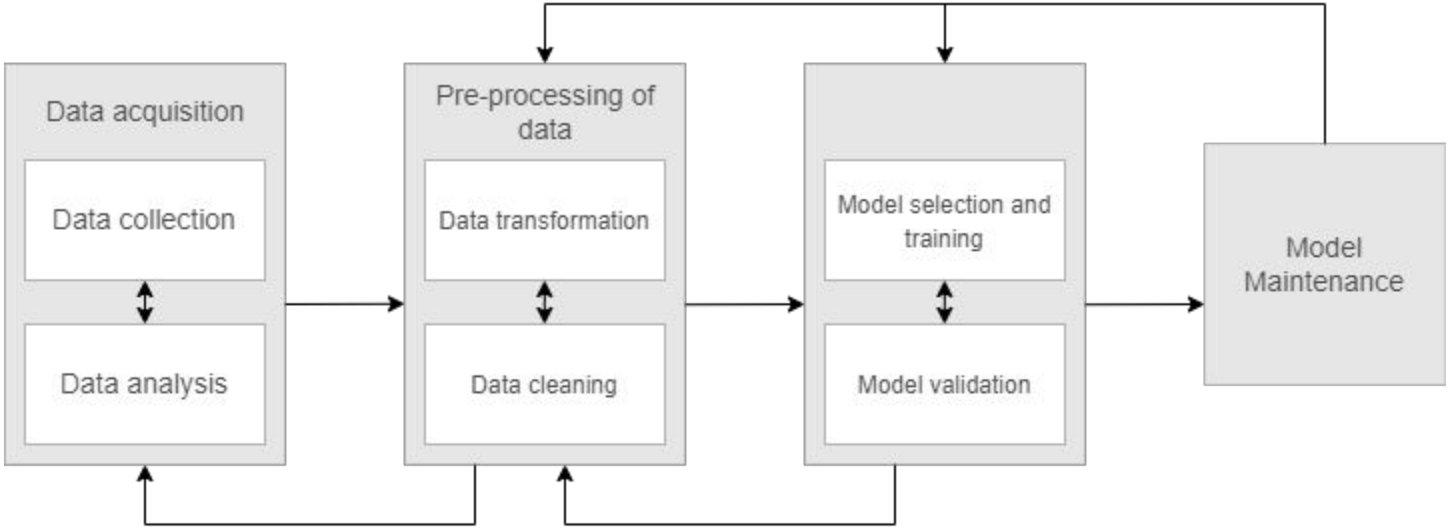
Introduction



Soft Sensor Applications

- Prediction of variables in real time that cannot be measured by physical sensors.
- Detection and diagnosis of failures during the process.
- Reduce expenses related to the process control.

Project Goal



Variable Selection

Pearson Correlation

- Decreases over-fitting
- Improves Accuracy
- Reduces Training Time

Pearson Correlation: Quantify linear dependence between two or more variables. Values ranging from -1 to 1:

$$R = \frac{\Sigma_{i=1}^{n}[(x_i - \mu_x)(y_i - \mu_y)]}{\sqrt{\Sigma_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \mu_y)^2}}$$

Variable Selection

Feature Expansion

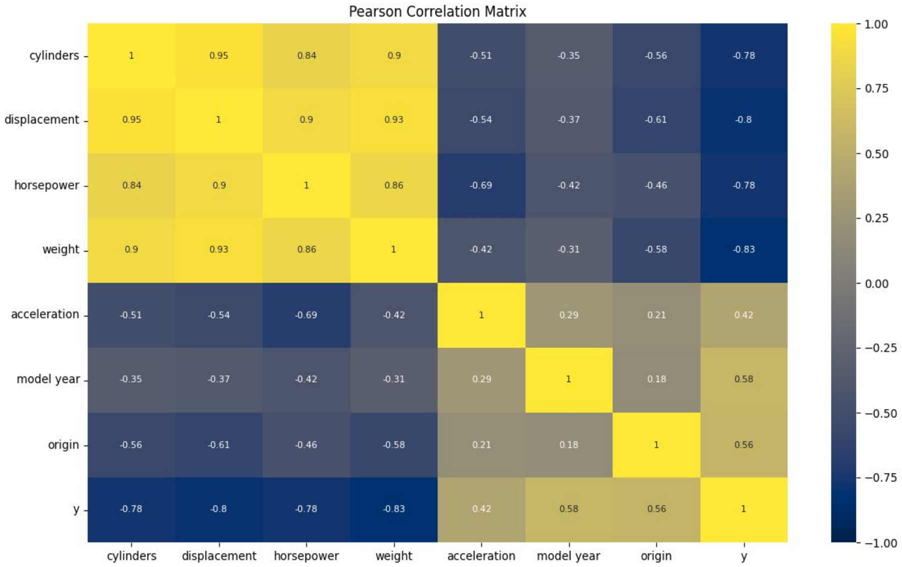
Create new variables to introduce non linearity to the model:

- x^2
- $\sqrt{x}$
- $1/x$
- xi*xj

$$|X|_{FE} = 4|X| + |X|!$$

Variable Selection

- Without VS
- With FE
- With FE and 3 different Pearson Correlation thresholds



Multiple Linear Regression

Uses several independent variables to predict a response variable (dependent).

Assumes a linear relation between independent variables.

$$\hat{y}_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i$$

$$\hat{\beta}_i = \underset{\beta}{argmin} \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Ridge Regression

Prevents overfitting.

Uses the L2-norm.

Reducing the model parameter through a penalty term, λ.

- λ = 0 : same result as MLR
- high λ : smaller parameter β (under-fitting)

$$\hat{\beta}_{RR} = \underset{\beta}{argmin} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 - \lambda \sum_{j=1}^{m} \beta_j^2$$

Least Absolute Shrinkage and Selection Operation

Uses shrinkage: data values are shrunk towards a central point.

Identical to RR but uses the L1-norm.

Less important variables can have null parameter.

$$\hat{\beta}_{Lasso} = \underset{\beta}{argmin} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{m} |\beta_j|$$

Elastic Net

Combines capability of variable selection (LASSO) and a better prediction performance (RR).

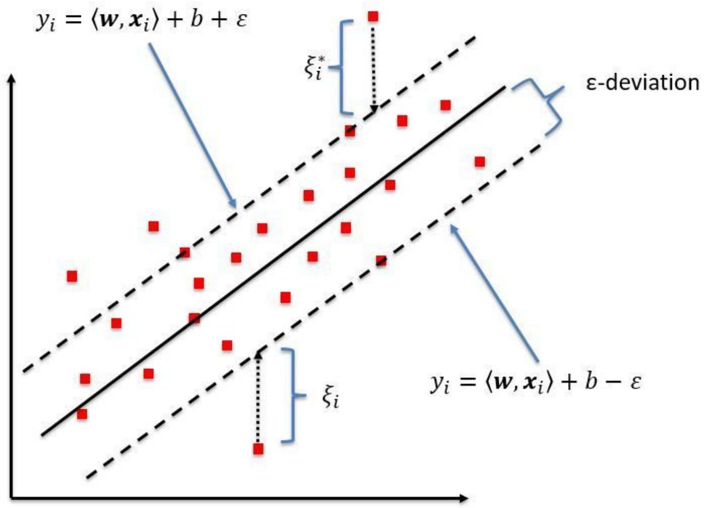
- α: strength of each penalty term
- λ: controls the trade-off between variance and bias.

$$\hat{\beta}_{EN} = \underset{\beta}{argmin} \frac{1}{2n} \cdot \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda(\alpha \sum_{j=1}^{m} |\beta_j| + \frac{1-a}{2} \sum_{j=1}^{m} \beta_j^2)$$

Support Vector Regression

Minimize hyperplane coefficients for fixed error

- ε: fixed maximum error
- ζ: slack (deviation from the margin)
- C: strength of ζ parcel



Support Vector Regression

$$L = \underset{\beta,\zeta}{argmin} \frac{1}{2}||\beta||_2^2 - C\sum_{i=1}^{n}\zeta_i + \zeta_i^*$$

- α: Lagrange multiplier
- K: kernel function
  - linear
  - rbf

$$f(x_n) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(x_i^T, x_i) + b$$

Datasets

Dataset 1:

- Name: Concrete Compressive Strength
- y = concrete_compressive_strength
- |X| = 8

Dataset 2:

- Name: Auto-Mpg Data
- y = mpg
- |X| = 7

Dataset 3:

- Name: Box–Jenkins Gas Furnace
- y = y(t) (Carbon Dioxide)
- |X| = 8

Dataset 4:

- Name: Water Treatment Plant
- y = y(k) (fluoride in the effluent)
- |X| = 55

Métricas para avaliar os resultados

R² - Coefficient of determination

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

close to 1: good fit

RMSE - Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

low RMSE: good fit

Pseudo Code

```
# Code to test a variable selection method and a training model
Import dataset
Preprocess data (delete outliers, non numeric variables and empty variables)
Choose a variable selection method: without VS or with FE or with FE and 3 Pearson
correlation thresholds
Input model #MLR or RR or LASSO or EN or SVR
Input hyperparameters values to try in a dictionary # except for MLR

for i = 1, ..., iterations:
    Split dataset in train (70%) and test (30%) randomly (but reproducible)
    Tune hyperparameters using a gridsearch and a 10 fold Cross Validation
    Fit the model with the best hyperparameters to the training dataset
    Predict y for the test dataset
    Calculate scores (r2 and rmse) between predicted y and real y for the test dataset
    Save model path, parameters and scores in a dataframe

Calculate final results (the mean of r2 and rmse)
Save results to a csv
Load and plot the best model from the dataframe
```

R^2 results

Dataset	Variable Selection	$\|X\|$	MLR	RR	LASSO	EN	SVR (linear)	SVR (rbf)	GMM
Dataset 1	without VS	8	0,604	0,604	0,604	0,604	0,581	0,865	
	with FE	57	0,869	0,869	0,859	0,858	0,852	0,908	
	with FE, PC = 0.2	32	0,842	0,842	0,837	0,837	0,824	0,886	
	with FE, PC = 0.3	26	0,835	0,834	0,834	0,834	0,826	0,887	
	with FE, PC = 0.4	11	0,726	0,726	0,726	0,726	0,713	0,829	
Dataset 2	without VS	7	0,808	0,807	0,808	0,807	0,802	0,866	
	with FE	49	-	0,867	0,865	0,866	0,864	0,879	
	with FE, PC = 0.7	27	-	0,845	0,834	0,834	0,848	0,846	
	with FE, PC = 0.75	24	0,836	0,844	0,836	0,836	0,844	0,840	
	with FE, PC = 0.8	9	0,733	0,731	0,732	0,732	0,707	0,711	
Dataset 3	without VS	8	0,993	0,993	0,993	0,993	0,992	0,992	
	with FE	52	0,991	0,992	0,992	0,992	0,992	0,991	
	with FE, PC = 0.7	28	0,989	0,990	0,990	0,990	0,990	0,990	
	with FE, PC = 0.75	27	0,988	0,990	0,990	0,990	0,990	0,990	
	with FE, PC = 0.8	23	0,989	0,990	0,990	0,990	0,990	0,990	
	with FE, PC = 0.9	18	0,989	0,990	0,990	0,990	0,990	0,990	
Dataset 4	without VS	55	0,616	0,690	0,695	0,693	0,643	0,758	
	with FE, PC = 0.7	52	0,522	0,621	0,598	0,592	0,553	0,689	
	with FE, PC = 0.72	26	0,546	0,567	0,553	0,569	0,532	0,674	
	with FE, PC = 0.73	11	0,554	0,565	0,557	0,567	0,543	0,679	

TABLE I
R^2 RESULTS

RMSE results

Dataset	Variable Selection	$\|X\|$	MLR	RR	LASSO	EN	SVR (linear)	SVR (rbf)	GMM
Dataset 1	without SV	8	10,439	10,443	10,441	10,444	10,731	6,082	
	with FE	57	6,002	5,994	6,230	6,234	6,374	5,024	
	with FE, PC = 0.2	32	6,593	6,594	6,696	6,698	6,958	5,588	
	with FE, PC = 0.3	26	6,739	6,759	6,749	6,760	6,923	5,570	
	with FE, PC = 0.4	11	8,690	8,686	8,679	8,684	8,881	6,849	
Dataset 2	without VS	7	3,446	3,452	3,448	3,452	3,482	2,861	
	with FE	49	-	2,866	2,884	2,874	2,879	2,708	
	with FE, PC = 0.7	27	-	3,085	3,199	3,197	3,040	3,060	
	with FE, PC = 0.75	24	3,169	3,103	3,183	3,184	3,087	3,122	
	with FE, PC = 0.8	9	4,062	4,075	4,066	4,066	4,232	4,197	
Dataset 3	without VS	8	0,267	0,267	0,267	0,267	0,278	0,280	
	with FE	52	0,304	0,278	0,273	0,273	0,278	0,296	
	with FE, PC = 0.7	28	0,338	0,324	0,319	0,319	0,321	0,320	
	with FE, PC = 0.75	27	0,341	0,324	0,319	0,319	0,322	0,320	
	with FE, PC = 0.8	23	0,339	0,321	0,318	0,319	0,322	0,320	
	with FE, PC = 0.9	18	0,331	0,318	0,317	0,318	0,321	0,323	
Dataset 4	without VS	55	0,024	0,021	0,021	0,021	0,023	0,019	
	with FE, PC = 0.7	52	0,026	0,023	0,024	0,024	0,025	0,021	
	with FE, PC = 0.73	26	0,025	0,025	0,025	0,025	0,026	0,022	
	with FE, PC = 0.72	11	0,026	0,025	0,025	0,025	0,026	0,022	

TABLE II

RMSE RESULTS

Conclusions

- SVR with rbf kernel has, generally, the best results
- Feature Expansion improves the metrics
- A high Pearson Correlation (with feature expansion) can lead to worse results
- Datasets with few variables and delays (D3 and D4) applied work best with simpler model but show low variance in results

Future Work

- Test other variable selection methods (as Mutual Information and Fast Tracker)
- Test other training model (GMM - gaussian mixture model)
- Use another dataset with similar properties as the first 2