Analyse et manipulation des données

DigitalLab@LaPlataforme_

Tools for data pre-processing

- Descriptive and inferential statistics tools
- Data transformations: indexing, grouping and aggregation
- Feature Selection
- Combination of data sets
- Encoding of categorical variables
- Dimensionality reduction with PCA, LDA
- Explainability with MDS, Isomap, LLE, T-sne

Other Encodings

Scaling

- **Standardization**: Common requirement for many ML estimators in scikit-learn; they might behave badly if the individual features do not more or less look like standard normally distributed data.

  $$z = (x - u) / s$$

- **MinMaxScaler**: Scales features between a given minimum and maximum value, often between zero and one,

  $$x\_s = (x - min) / (max- min)$$

  $$x\_s (R - L) + L$$

- **MaxAbsScaler:** Special case of MinMaxScaler but for [-1, 1].

Ordinal Encoding

To convert categorical features to nteger codes, we can use ordinal encodersr. This estimator transforms each categorical feature to one new feature of integers (0 to n_categories

Country
Argentina
Brazil
France
Brazil

Encoded Country
0
1
2
1

Discretizers

We can take a numerical variable and segment it equally in categories.

For example, if we are dealing with the salary of developers, we can discretize it in three groups, in such a way these groups have more or less the same number of instances.

Polynomial Features

Often it’s useful to add complexity to a model by considering nonlinear features of the input data. One possibility is to use polynomial features.

For example, if we have the features of X1 and X2, we can create six features from them by combining through multiplications obtaining:

(1, X1, X2, X1.X1, X1.X2, X2.X2)

Demo notebook 10_pipelines_and_other_enc.ipynb