Analyse et manipulation des données

*DigitalLab@LaPlataforme_*

Tools for data pre-processing

- Descriptive and inferential **statistics tools**

Tools for data pre-processing

- Descriptive and inferential **statistics tools**
- **Data transformations**: indexing, grouping and aggregation

Tools for data pre-processing

- Descriptive and inferential **statistics tools**
- **Data transformations**: indexing, grouping and aggregation
- **Feature Selection**

Tools for data pre-processing

- Descriptive and inferential **statistics tools**
- **Data transformations**: indexing, grouping and aggregation
- **Feature Selection**
- **Combination** of data sets

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

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**

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 look like standard normally distributed data.

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

Scaling

- **Standardization**: Common requirement for many ML estimators in scikit-learn; they might behave badly if the individual features do not 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,

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

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

Scaling

- **Standardization**: Common requirement for many ML estimators in scikit-learn; they might behave badly if the individual features do not 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

Given an ordinal categorical r.v N with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Enumeration

Primary	0
Secondary	1
University	2
Doctorate	3
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	Primary
1	Postdoc
2	University
3	Doctorate
4	Secondary
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Enumeration

Primary	0
Secondary	1
University	2
Doctorate	3
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	Primary
1	Postdoc
2	University
3	Doctorate
4	Secondary
5	Primary

Encoded dataframe

Index	Studies Level
0	Primary
1	Postdoc
2	University
3	Doctorate
4	Secondary
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < \ldots < C_n$ we enumerate them with integers $0 < \ldots < n-1$. This encoding preserves the order.

Enumeration

Primary	**0**
Secondary	1
University	2
Doctorate	3
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	**Primary**
1	Postdoc
2	University
3	Doctorate
4	Secondary
5	Primary

Encoded dataframe

Index	Studies Level
0	**0**
1	Postdoc
2	University
3	Doctorate
4	Secondary
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Enumeration

Primary	0
Secondary	1
University	2
Doctorate	3
Postdoc	**4**

DataFrame to Encode

Index	Studies Level
0	Primary
1	**Postdoc**
2	University
3	Doctorate
4	Secondary
5	Primary

Encoded dataframe

Index	Studies Level
0	0
1	**4**
2	University
3	Doctorate
4	Secondary
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Enumeration

Primary	0
Secondary	1
University	**2**
Doctorate	3
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	Primary
1	Postdoc
2	**University**
3	Doctorate
4	Secondary
5	Primary

Encoded dataframe

Index	Studies Level
0	0
1	4
2	**2**
3	Doctorate
4	Secondary
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Enumeration

Primary	0
Secondary	1
University	2
Doctorate	**3**
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	Primary
1	Postdoc
2	University
3	**Doctorate**
4	Secondary
5	Primary

Encoded dataframe

Index	Studies Level
0	0
1	4
2	2
3	**3**
4	Secondary
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < ... < C_n$ we enumerate them with integers $0 < ... < n-1$. This encoding preserves the order.

Enumeration

Primary	0
Secondary	1
University	2
Doctorate	3
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	Primary
1	Postdoc
2	University
3	Doctorate
4	**Secondary**
5	Primary

Encoded dataframe

Index	Studies Level
0	0
1	4
2	2
3	3
4	1
5	Primary

Ordinal Encoding

Given an ordinal categorical r.v X with categories $C_1 < C_2 < \ldots < C_n$ we enumerate them with integers $0 < \ldots < n-1$. This encoding preserves the order.

Enumeration

Primary	**0**
Secondary	1
University	2
Doctorate	3
Postdoc	4

DataFrame to Encode

Index	Studies Level
0	Primary
1	Postdoc
2	University
3	Doctorate
4	Secondary
5	**Primary**

Encoded dataframe

Index	Studies Level
0	0
1	4
2	2
3	3
4	1
5	**0**

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_encodings.ipynb