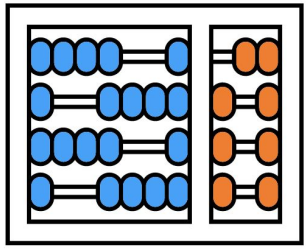


Instituto de Computação

UNIVERSIDADE ESTADUAL DE CAMPINAS

Capacitação profissional em tecnologias de Inteligência Artificial

Machine Learning Overview

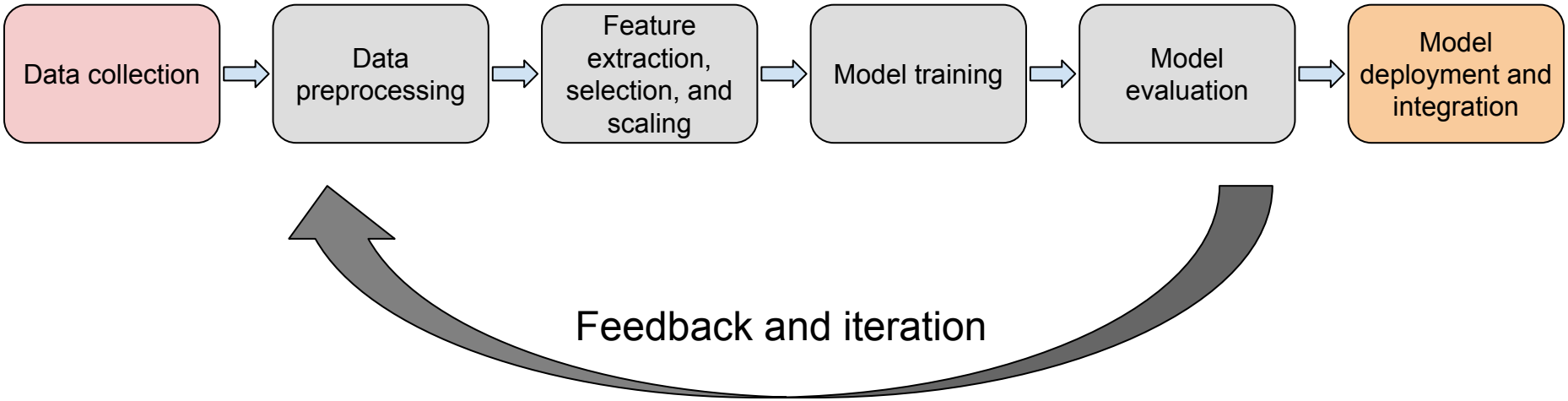
Prof. Edson Borin

https://www.ic.unicamp.br/~edson

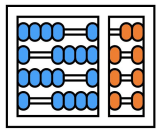
Institute of Computing - UNICAMP

ML Process Overview

ML Process


Data collection
Data preprocessing
Feature extraction, selection, and scaling
Model training
Model evaluation
Model deployment and integration
Feedback and iteration

Dataset Terminology



ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called a sample.
- Each sample i is represented by a set of n features $(x_1^i, \ldots, x_n^i)$
- Each sample i may contain a label (y^i)

ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called a sample.
- Each sample i is represented by a set of n features $(x_1^i, \ldots, x_n^i)$
- Each sample i may contain a label (y^i)

House price training dataset

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
House 3	1534 ft.^{2}	3	0	315,000 USD
House 4	852 ft.^{2}	2	12	178,000 USD
...	...	...	...	...

ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called a sample.
- Each sample i is represented by a set of n features (
- Each sample i may contain a label (y^i)

First sample (i=1)

House price training dataset

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
House 3	1534 ft.^{2}	3	0	315,000 USD
House 4	852 ft.^{2}	2	12	178,000 USD
...	...	...	...	...

ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called a sample.
- Each sample i is represented by a set c… x_n^i)
- Each sample i may contain a label (y^i)

House 1 **features**:
$x_1^1 = 2104$ (House size)
$x_2^1 = 4$ (# bedrooms)
$x_3^1 = 1$ (Age in Years)

House price training …aset

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
House 3	1534 ft.^{2}	3	0	315,000 USD
House 4	852 ft.^{2}	2	12	178,000 USD
…	…	…	…	…

ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called a sample.
- Each sample i is represented by a set o
- Each sample i may contain a label (y^i)

House 1 **label**:
y^1 = 460,000 (House price)

House price training dataset

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
House 3	1534 ft.^{2}	3	0	315,000 USD
House 4	852 ft.^{2}	2	12	178,000 USD
...	...	...	...	...

ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called a sample.
- Each sample i is represented by a set of n features $(x_1^i, \ldots, x_n^i)$
- Each sample i may contain a label (y^i)

House price training dataset

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
H				315,000 USD
H				178,000 USD
				...

Features and **labels** are usually represented as numerical (real or integer) values.

ML Process

Dataset: collection of data used in machine learning tasks.

- Each record is called
- Each sample i is repr… es $(x_1^i, \ldots, x_n^i)$
- Each sample i may contain … el (y^i)

Dataset may be split into subsets. Ex: **training** and **test** sets

House price training dataset

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)	
House 1	2104 ft.^{2}	4	1	460,000 USD	Training Set
House 2	1416 ft.^{2}	2	2	232,000 USD	
House 3	1534 ft.^{2}	3	0	315,000 USD	
House 4	852 ft.^{2}	2	12	178,000 USD	Test Set
…	…	…	…	…	

ML Process

Dataset organization:

- CSV file, database tables, python objects (Numpy array, pandas dataframe, Python lists, …)

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
…	…	…	…	…

train.csv

```
House ID,Size,# Bedrooms,Age,Price
1,2104,4,1,460000
2,1416,2,2,232000
...
```

Explicit object construction in python

```
dataset = [ [1,2104,4,1,460000],
            [2,1416,2,2,232000],
            ...
          ]
```

ML Process

Dataset organization:

- CSV file, database tables, python objects (Numpy array, pandas dataframe, Python lists, …)

train.csv

```
House ID,Size,# Bedrooms,Age,Price
1,2104,4,1,460000
2,1416,2,2,232000
...
```

List of lists in python

```
dataset = [ [1,2104,4,1,460000],
            [2,1416,2,2,232000],
            ...
          ]
```

	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
…	…	…	…	…

ML Process

Dataset organization:

- CSV file, database tables, python objects (Numpy array, pandas dataframe, Python lists, …)
- X (dataset attributes) are usually viewed as a table in which each line represents a dataset item and each column an attribute.
- Y (dataset labels) can be represented as a separated array of data or as a column on the same object (file) as the dataset attributes.

	X			Y
	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
…	…	…	…	…

ML Process

Dataset organization:

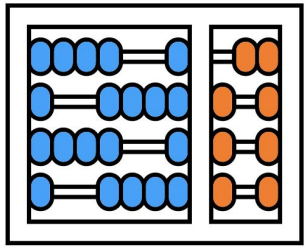
- CSV file, database tables, python objects (Numpy array, pandas dataframe, Python lists, …)
- X (dataset attributes) are usually viev ach line represents a dataset item and ea
- Y (dataset labels) can be represented data or as a column on the same object (f tes.

List of lists in python

```
X = [ [2104,4,1],
      [1416,2,2,],
      ...]

Y = [460000,
     232000,
     ... ]
```

	X			Y
	x_1 (Size)	x_2 (# bedrooms)	x_3 (Age in Years)	y (**Price**)
House 1	2104 ft.^{2}	4	1	460,000 USD
House 2	1416 ft.^{2}	2	2	232,000 USD
…	…	…	…	…



Instituto de Computação

UNIVERSIDADE ESTADUAL DE CAMPINAS

Capacitação profissional em tecnologias de Inteligência Artificial

Machine Learning Overview

Prof. Edson Borin

https://www.ic.unicamp.br/~edson

Institute of Computing - UNICAMP