Lecture 2: Datasets and variables

Introduction to Machine Learning

Sophie Robert

L3 MIASHS | Semestre 2

2023-2024

1 Datasets

- Definition
- Example

2 Variables

- Variable types
- Studying numeric variables
- Studying categorical variables

Reminder on previous session

In the previous session, we learned that **Machine Learning** algorithms are able to **learn**, **infer** and **predict** given **data**.

To build a Machine Learning algorithm, you need **data** !

Question

Can anyone tell me what a **dataset** is ?

Datasets

Datasets

Datasets

A **dataset*** can be thought of as a matrix $M = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$ with n the number of individuals in the population and m the number of variables.

Columns of a table represents a **particular variable** (also called **feature**), and each row corresponds to a given **record** of the data set in question for an **individual**.

Datasets

Example:

Individual	Variable 1	Variable 2	Variable 3
ID1	5	4	1
ID2	2	3	1

Question:

Give the value for:

$x_{1,3} =$

$x_{2,1} =$

Variable 1 for individual 1

All data regarding individual 2

Example of dataset

The Iris dataset was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems*.

ID	Sepal length	Sepal width	Petal length	Specie
1	2.1	3.1	4.1	Setosa
2	3.1	1.1	2.1	Setosa
3	4.1	5.1	3.1	Versicolor
4	1.1	2.1	2.1	Virginica

Example of dataset

The Iris dataset was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems*.

ID	Sepal length	Sepal width	Petal length	Specie
1	2.1	3.1	4.1	Setosa
2	3.1	1.1	2.1	Setosa
3	4.1	5.1	3.1	Versicolor
4	1.1	2.1	2.1	Virginica

Question

Does anyone from lecture 1 remember for what type of problem is the **Iris dataset used for** ?

Example of dataset

ID	Sepal length	Sepal width	Petal length	Specie
1	2.1	3.1	4.1	Setosa
2	3.1	1.1	2.1	Setosa
3	4.1	5.1	3.1	Versicolor
4	1.1	2.1	2.1	Virginica

The names of the variables are:
There are ______ individuals.
There are ______ variables.

Variables

Variable types

Question

Can anyone list the different types of variables that can be encountered in datasets ?

Variable types

Let's consider a dataset $M = (x_{i,j})_{1 \leq n, 1 \leq m}$, with n individuals and m variables.
A variable j can be:

Variable types

Let's consider a dataset $M = (x_{i,j})_{1 \leq n, 1 \leq m}$, with n individuals and m variables.
A variable j can be:

- **Numeric**: $(x_{i,j})_{1 \leq i \leq n} \in \mathbb{R}^n$.
  Example: **Petal width**.

Variable types

Let's consider a dataset $M = (x_{i,j})_{1 \leq n, 1 \leq m}$, with n individuals and m variables.
A variable j can be:

- **Numeric**: $(x_{i,j})_{1 \leq i \leq n} \in \mathbb{R}^n$.
  Example: **Petal width**.
- **Categorical**: $(x_{i,j})_{1 \leq i \leq n} \in \mathcal{X}^n$, with $\mathcal{X}$ a set of distinct values.
  A special case of categorical variables often encountered .
  Example: **Flower specie**.

Variable types

Let's consider a dataset $M = (x_{i,j})_{1 \leq n, 1 \leq m}$, with n individuals and m variables.
A variable j can be:

- **Numeric**: $(x_{i,j})_{1 \leq i \leq n} \in \mathbb{R}^n$.
  Example: **Petal width**.
- **Categorical**: $(x_{i,j})_{1 \leq i \leq n} \in \mathcal{X}^n$, with $\mathcal{X}$ a set of distinct values.
  A special case of categorical variables often encountered .
  Example: **Flower specie**.
- **Ordinal**: $(x_{i,j})_{1 \leq i \leq n} \in \mathcal{X}^n$, with $\mathcal{X}$ a set of **ordered** distinct values.
  Example: **Performance (low, medium, high)**.

Dataset analysis

To **analyze a dataset**, you can perform:

Dataset analysis

To **analyze a dataset**, you can perform:

- A **visual*** analysis: use graphs to better understand the dataset.

Dataset analysis

To **analyze a dataset**, you can perform:

- A **visual*** analysis: use graphs to better understand the dataset.
- A **statistical*** analysis: use statistical estimators to better understand the dataset.

Dataset analysis

To **analyze a dataset**, you can perform:

- A **visual*** analysis: use graphs to better understand the dataset.
- A **statistical*** analysis: use statistical estimators to better understand the dataset.

Analysis depends on the variable type !
A poor analysis of variables can cause misinterpretation of data.

Dataset analysis

Question

Can anyone give me:

- Possible **graphical representation** of **numeric** and **categorical** variables ?
- Possible **estimators** of **numeric** and **categorical variables** ?

ID	Sepal length	Sepal width	Petal length	Specie
1	2.1	3.1	4.1	Setosa
2	3.1	1.1	2.1	Setosa
3	4.1	5.1	3.1	Versicolor
4	1.1	2.1	2.1	Virginica

Analyzing numeric variables

Usual indicators include:

- **Arithmetical mean**: summarize to better understand the overall value.
  $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Analyzing numeric variables

Usual indicators include:

- **Arithmetical mean**: summarize to better understand the overall value.
  $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **Variance and standard error**: measures the **dispersion of the data**.
  $\text{var}(X) = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{X})^2$
  $\sigma(X) = \sqrt{\text{var}(X)}$

Analyzing numeric variables

Usual indicators include:

- **Arithmetical mean**: summarize to better understand the overall value.
  $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- **Variance and standard error**: measures the **dispersion of the data**.
  $\text{var}(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2$
  $\sigma(X) = \sqrt{\text{var}(X)}$
- **Quantiles**: divide the ordered vectors into equal parts of same

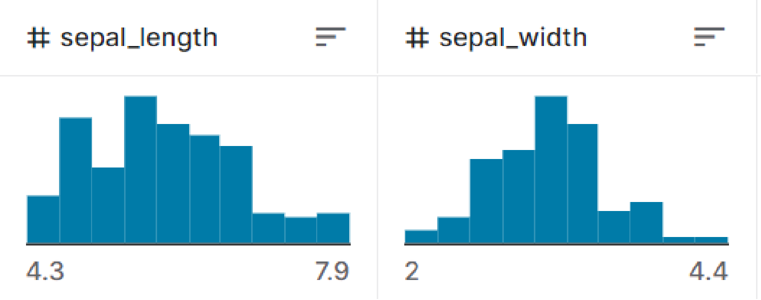
  $1/4$ quantiles, median

  Very useful for datasets with a lot of outliers*!

Representing numeric variables: histograms

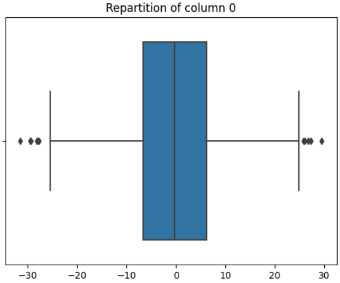
Histograms* consist in:

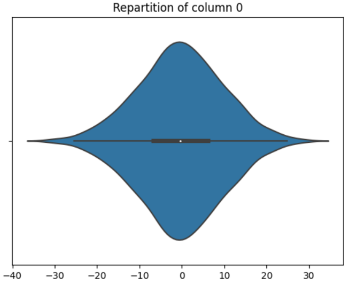
- Dividing the numerical space into intervals of regular length
- Computing the frequency of values per interval



Representing numeric variables: boxplots

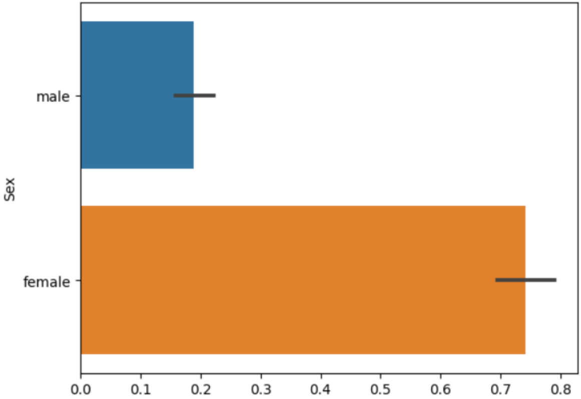
Boxplots* and **violin plots*** consist in representing all the values of the variables and their statistical indicators (usually, quantiles and medians).





Analyzing and representing categorical variables

Categorical variables are often **harder** to study.
Usual indicators are **counts** and **frequency**.
Usual graphical representation can be **bar graphs**.



Questions

Questions ?