

Inferential statistics

Applied Data Analysis (ADA) - October 2025

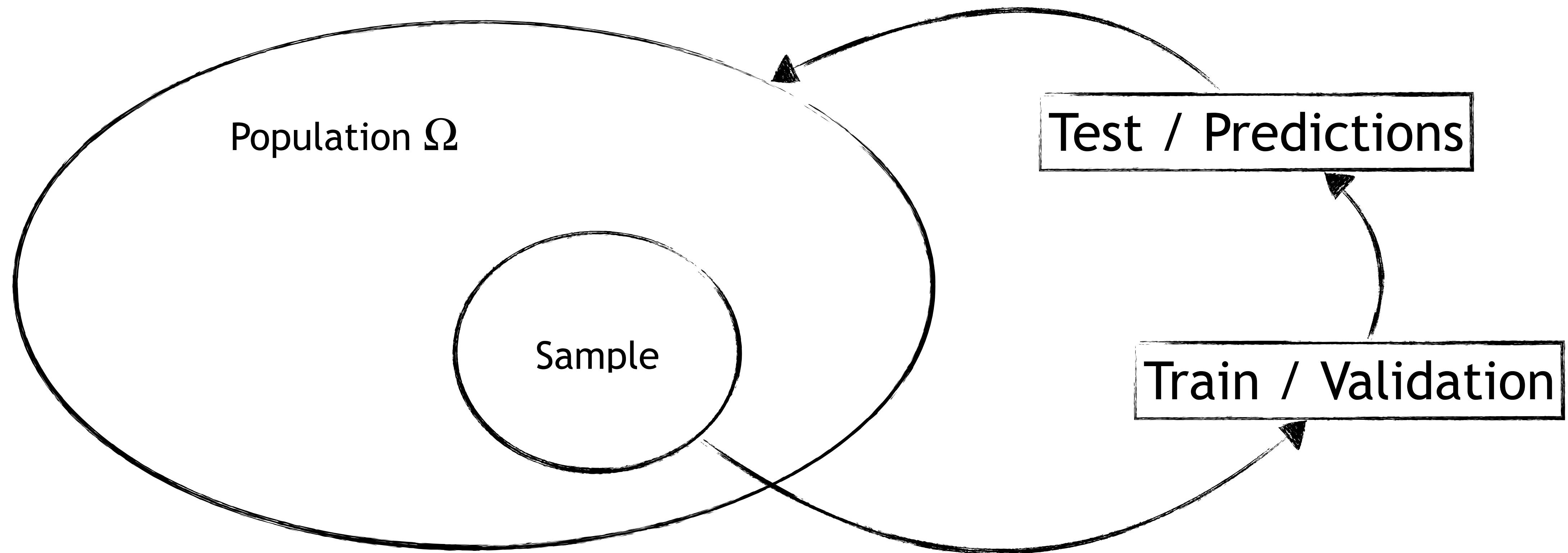
Nomades Advanced Technologies

Gaspard Villa

- ❖ Monday : Metrics and training process
 - Supervised & Unsupervised learning
 - Evaluation metrics
 - Training process
- ❖ Tuesday : Linear models
- ❖ Wednesday : Support Vector Machine models
- ❖ Thursday : Decision trees
- ❖ Friday : Time series analysis

Inferential statistics

Definition : Inferential statistics is the idea of drawing conclusions about a population base on data from a sample.



Supervised learning

Definition : Supervised learning is a type of machine learning where the model is trained on a labeled dataset, meaning each input data point is paired with the correct output (target).

=> The goal is for the model to learn the mapping from inputs to outputs and make predictions on new unseen data.

Supervised learning

- Example -

- ❖ Spam detection in mails => Input : email text / Label : spam or not spam
- ❖ Prediction of house prices => Input : number of rooms, area, location / Label : price.
- ❖ Medical diagnosis => Input : patient symptoms and test results / Label : disease present or not.
- ❖ Image classification => Input : image pixels / Label : “cat”, “dog”, etc...

Unsupervised learning

Definition : Unsupervised learning is a type of machine learning where the model is given unlabeled data.

=> The goal is to find patterns, groupings, or structures within the data without knowing the “correct” answers.

Unsupervised learning

- Example -

- ❖ Customer segmentation => Task : grouping customers by purchasing behaviour.
- ❖ Anomaly detection => Task : Finding fraudulent transactions in banking data.
- ❖ Topic modelling => Task : discovering themes in a collection of new articles.
- ❖ Dimensionality reduction => Task : visualizing high-dimensional data (e.g., genetics, images) using PCA.

Supervised learning

- Regression methods -

Definition : It models the relationship between one or more independent variables (features) and a continuous dependent variable (target).

=> It's a supervised learning method.

Regression methods

- Illustration of some models -

- Simple linear regression : single feature vs target.
- Multiple linear regression : multiples features to predict the target.
- Polynomial regression : Useful for modelling non-linear trends.
- Regularized regression : Penalizes weights to prevent overfitting and manage multicollinearity => Improves generalisation.

Supervised learning

- Classification methods -

Definition : It models the relationship between one or more independent variables (features) and a categorical dependent variable (target).

=> It's also a supervised learning method.

Classification methods

- Example -

- Logistic regression : Probability model for binary tasks.
- Decicision tree : Tree based model, interpretable model.
- Random forest : multiple random trees.
- Support Vector Machines (SVM) : Find optimal separating hyperplane.
- Neural Networks : Flexible and powerful (especially for high-dimensional inputs).

Evaluation of models

- Metrics -

How do you know your model is working better than the others ?

- ➡ You define metrics to evaluate the model's performance.
- ➡ But can you use the same metrics for regression and classification tasks ?
- ➡ How do you choose your metric ?

Evaluation of models

- Metrics for regression -

Goal : We need to have a metric that evaluate how close predictions of continuous variables are to the actual target variables.

1 - Error terms : Let y_i be the true value and \hat{y}_i the predicted value:

$$\epsilon_i = y_i - \hat{y}_i$$
$$\Rightarrow \text{Error} = \sum_{i=1}^n \epsilon_i$$

Evaluation of models

- Metrics for regression -

2 - Mean Absolute Error (MAE) : It measures average absolute difference.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Remark : Easy to interpret but doesn't penalise large errors.

Evaluation of models

- Metrics for regression -

3 - Mean Squared Error (MSE) : It measures the square of errors.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Remark : Commonly used in optimisation!

Evaluation of models

- Metrics for regression -

4 - Root Mean Squared Error (RMSE) : Its measure has the same units as the target variable.

$$\text{RMSE} = \sqrt{\text{MSE}}$$

Remark : More sensitive to outliers than MAE.

Evaluation of models

- Metrics for regression -

5 - R-Squared (R^2) : It measures a proportion of variance explained.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Remark : 0 means that the prediction is not better than the mean as a naïve prediction.

Evaluation of models

- Metrics for regression -

6 - Mean Average Percentage Error (MAPE) : It measures a prediction accuracy of the models.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Remark : Very intuitive error interpretation in terms of relative error.

Evaluation of models

- Metrics for classification -

Goal : We need to have a metric that evaluate the quality of the prediction given by a specific model.

0 - Basic concepts:

- True Positive (TP) : true value = 1, predicted value = 1
- True Negative (TN) : true value = 0, predicted value = 0
- False Positive (FP) : true value = 0, predicted value = 1
- False Negative (FN) : true value = 1, predicted value = 0

Evaluation of models

- Metrics for classification -

1 - Accuracy: It measures the proportion of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Remark : Issues when facing imbalanced classes.

Evaluation of models

- Metrics for classification -

2 - Precision : It measures how many predicted positives were correct.

$$\text{Precision} = \frac{???}{???}$$

Remark : Useful for imbalanced classes.

Evaluation of models

- Metrics for classification -

2 - Precision : It measures how many predicted positives were correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Remark : Useful for imbalanced classes.

Evaluation of models

- Metrics for classification -

3 - Recall : It measures how many actual positives were captured.

$$\text{Recall} = \frac{???}{???}$$

Remark : Useful for imbalanced classes.

Evaluation of models

- Metrics for classification -

3 - Recall : It measures how many actual positives were captured.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Remark : Useful for imbalanced classes.

Evaluation of models

- Metrics for classification -

4 - F1-score : It measures harmonic mean of precision and recall.

$$\text{F1-score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Remark : Useful for imbalanced classes.

Evaluation of models

- Metrics for classification -

5 - Confusion matrix : It shows all 4 outcomes (TP, TN, FP, FN) in a 2x2 table.

		Prediction	
		0	1
Truth	0	TN	FP
	1	FN	TP

Evaluation of models

- Metrics for classification -

6 - Log-loss (cross-entropy loss) : It measures probabilistic confidence in predictions.

$$\text{Log-loss} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Remark : Commonly used in optimisation!

Underfitting

Definition : The model is too simple to capture underlying patterns. It can be visualised by a poor performance both on training and validation sets.

=> Make the model more complex by modifying its parameters or completely change the model.

Overfitting

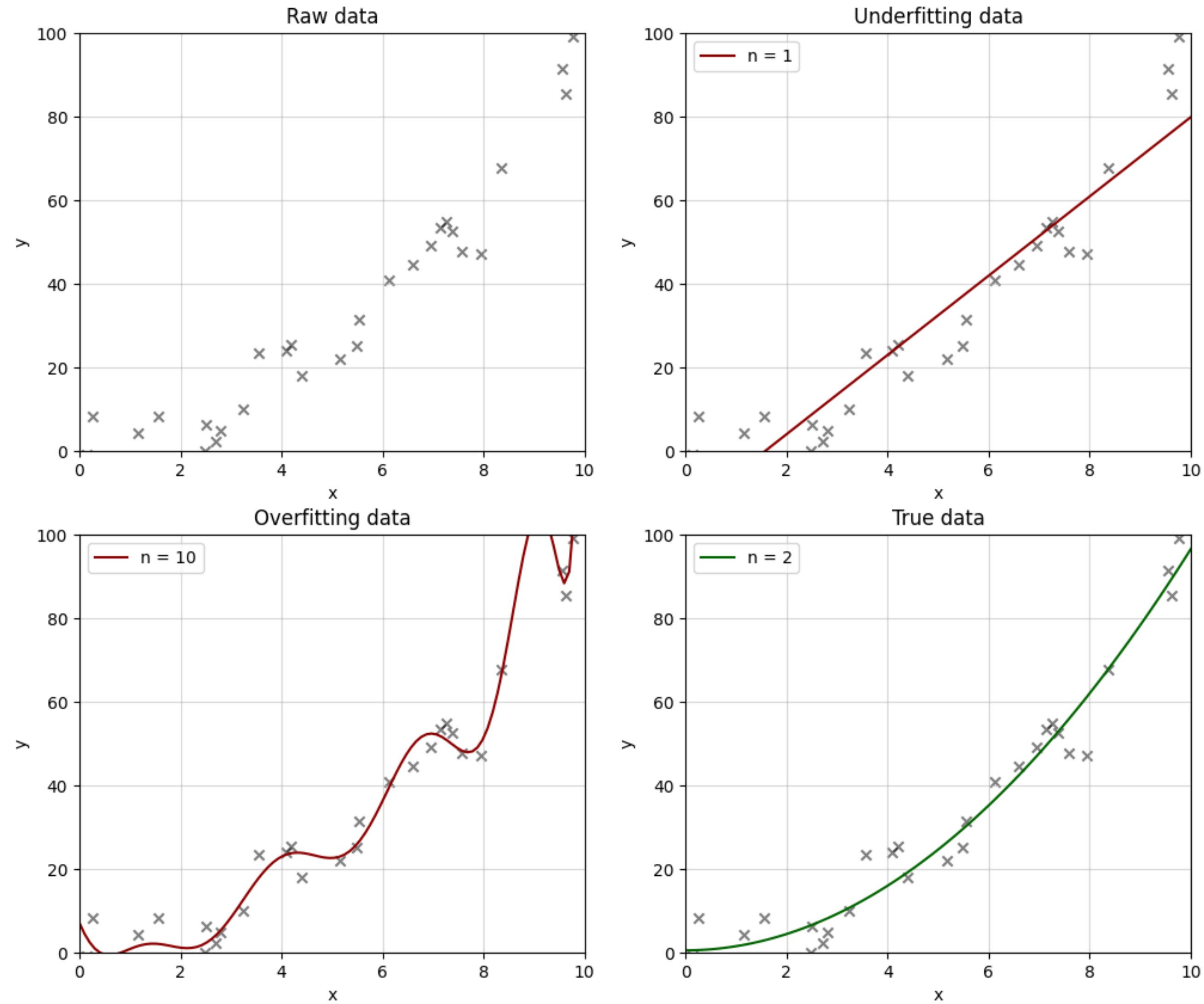
Definition : The model learns noise and details from the training data. It can be visualised by a good performance on training set but poorly on validation set.

=> Simplify the model, regularisation or training data set not well defined regarding the validation set (split to redo).

Underfitting / Overfitting

Practice : Use the data set `overfitting_data.csv`, train a Polynomial Regression model on it and plot the results. Do it for $n = 1, 2$ and 10 .

Underfitting / Overfitting



Practice

- Cross-validation using sklearn -

Training process

- General overview -

- **Data preprocessing** : Cleaning, normalisation, encoding, split, etc... (EDA)
- **Model selection** : Chose one or multiples model types.
- **Loss function definition** : Quantify how wrong the model is (not performance!)
- **Optimization** : Algorithms to minimize loss such as gradient descent.
- **Evaluation** : Measure performance on training and validation sets.

Practice

- KNN -