Data Science

Msc Supply Chain & Retail Management (2025-2026)

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

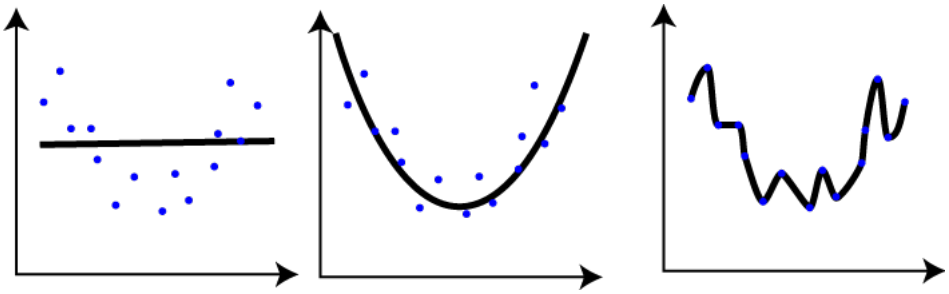
Context: You are working as a data scientist in a retail supply chain company. Your goal is to use machine learning to improve decision-making in purchasing, supplier selection, and inventory management.

Exercise 1: Generalities in Machine Learning

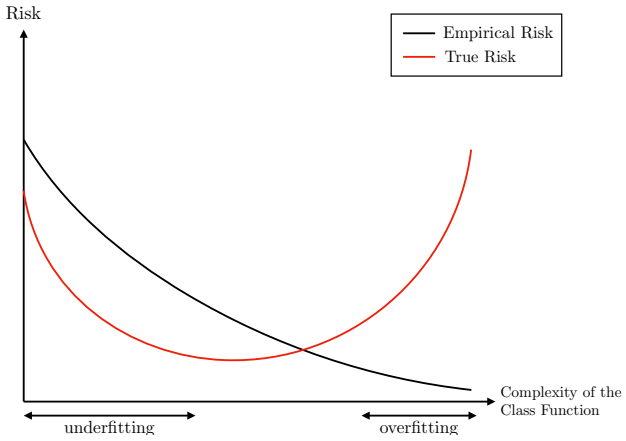
In this first part, we are dealing with some generalities in Machine Learning.

Generalities

1. What are the three main categories of learning algorithms we can meet in Machine Learning?
2. Among the set supervised algorithms, what are the two types of problem that we can solve (or types of hypothesis we can learn)?
3. Recall the definition of the $0-1$ loss, given a binary classifier h.
4. Explain the cross validation procedure.
5. Associate each of the following graph with the following situations:
    - Overfitting
    - Underfitting
    - Ideal



6. Explain the following graph



Learning a model

Let $S = \{\mathbf{x}_i, y_i\}_{i=1}^m$ a sample of size , ℓ a loss and $\mathbf{w}$ the model parameters. Consider any classification algorithm that seeks to solve the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_{\mathbf{w}}(\mathbf{x}_i), y_i) + \lambda \|\mathbf{w}\|, \tag{1}$$

where $h_{\mathbf{w}}$ represents a hypothesis (a classifier or regressor) that depends on the parameter $\mathbf{w}$.

1. In the optimization problem (1), identify what is call the *empirical risk* and the *regularization term*.
2. What is the name of λ and its role?
3. Recall the general objective of Machine Learning (particularly with regard to empirical risk and generalization risk).

4. Describe precisely the procedure you should put in place to solve the problem (1) in order to answer to the previous question.
   The aim is to explain how you will conduct the experiment.

Exercise 4: Ensemble Methods

1. Explain, in your own words, the way the decision tree algorithm works and what are its weaknesses.
2. How can we build a random forest using decision trees? How it can solve a weakness of the decision tree algorithm?
3. Cite another method seen in class which is closed to the previous one, *i.e.*, another ensemble method.

Exercise 3: Procurement Cost Prediction

You are predicting procurement cost from:

- `Region (A, B, C)`
- `Quantity`
- `Delivery Time`

Region	Quantity	Delivery Time	Cost
A	50	5	1100
B	200	4	400
A	130	6	2100
C	300	7	3100
B	200	5	900

1. Propose a split on one feature (e.g., `Quantity` < 150). Compute the **variance of cost** in each resulting subset.
2. Calculate the **reduction in variance (information gain)** after the split.
3. Which split (by region, quantity, or delivery time) gives the best regression performance?
4. Discuss the risk of overfitting in regression trees and how pruning can help.

Exercise 4: Delivery Delay Prediction

You aim to predict a potential delay in supply chain using the following data:
Exam

Supplier	Items	Delayed?
A	57	Yes
A	40	No
B	12	Yes
A	30	No
B	20	Yes

1. Calculate the **Gini impurity** of the root node.
2. Compute the Gini impurity of the children nodes after splitting on `Supplier`.
3. Calculate the **information gain** from this split.
4. Suggest an alternative split on `Items < 45` and perform the same calculations.
5. Which split is better? Explain.