Different setups in supervised Machine Learning: class workshop

Julián D. Arias-Londoño
Dpto. of Systems Engineering and Computer Science
Universidad de Antioquia

Abstract
This document aims to introduce students to various machine learning paradigms and setups, providing definitions and case examples to enhance understanding. After some definitions, students are given some exercises where hypothetical application contexts are described and they must identify the corresponding learning paradigm.

Contents

1 Introduction

Fundamental definitions of Machine Learning (ML) typically begin by introducing supervised and unsupervised learning, as they form the foundation for understanding how data availability influences different scenarios. These concepts do not define specific tasks, as various problems can fall under each category. Instead, their core distinction lies in whether target values or labels are available for each sample in a dataset $\mathcal{D}$. In certain cases, a problem can even be approached using both supervised and unsupervised methods.

It is equally essential to understand the fundamental characteristics of these learning paradigms:

- Supervised learning problems: classification, regression.
- Unsupervised learning problems: clustering, dimensionality reduction, matrix completion*

In simple terms, classification problems involve a discrete target variable y, while regression problems involve a continuous one. Additionally, although dimensionality reduction is typically an unsupervised task, it can be performed in a supervised manner when y is available and the chosen technique incorpo-

*Matrix completion, commonly used in recommender systems, is closely related to data imputation.

rates it.

While these definitions cover most ML problems, real-world applications often involve variations that do not neatly fit into standard categories. Accurately identifying the specific setup of a problem is crucial, as it determines not only the appropriate models and learning techniques but also whether additional post-processing is required after model predictions[4]. Moreover, it influences the choice of evaluation metrics and validation methodologies.

This workshop aims to introduce students to different extensions of traditional supervised and unsupervised learning paradigms. Through practical case studies, participants will develop the ability to recognize these variations and apply suitable ML approaches accordingly.

2 Different setups of supervised or unsupervised learning

2.1 Imbalanced learning

Imbalanced learning occurs when the distribution of classes in the dataset is skewed, meaning some classes have significantly more samples than others. This often leads to biased models that favour majority classes[1]. *Example*: Fraud detection in banking, where fraudulent transactions are much rarer than legitimate ones.

2.2 Multi-instance learning

In multi-instance learning, the dataset consists of labelled bags of instances rather than individually labelled samples. The label of the entire bag may depend on the problem; in some cases, all instances that form a sample are given the same label, which means that during inference, a new sample is assigned the label according to a consensus mechanism among the instances composing it. However, in some cases, a sample, i.e., a bag, is positive if it contains at least one positive instance. *Example*: Drug discovery, where each molecule (bag) contains multiple molecular conformations (instances), and the molecule is labelled as effective if at least one conformation exhibits the desired property.

2.3 Multi-output learning

Multi-output learning involves predicting multiple target values simultaneously for a given input. This problem is also called **multi-tasks** learning. The multiple variables to be predicted must hold some relationship among them, so approaching their prediction simultaneously makes sense. *Example*: Predicting both temperature and humidity from satellite images in climate modelling.

2.4 Multi-label learning

Multi-label learning allows instances to belong to multiple categories simultaneously[8]. *Example*: A news article classified under multiple topics such as politics, business, and technology.

2.5 Multi-labeller learning

Multi-labeller learning arises when multiple annotators provide labels for the same data, requiring methods to aggregate diverse annotations. *Example*: Sentiment analysis based on user-generated ratings, where different users may rate the same review differently.

2.6 Multiview Learning

Multiview learning is a paradigm where data for each instance is described using multiple, complementary feature sets (or "views"). The goal is to exploit the redundancy and complementarity across views to improve learning performance[7]. *Example*: In image classification, an object can be captured from multiple camera angles (front, side, top). Each angle represents a different view of the same modality (images). The model learns from all views to improve classification accuracy.

2.7 Multimodal Learning

Multimodal learning refers to the integration of information from different modalities (e.g., text, images, audio, sensor data) to improve learning. Each modality represents a fundamentally different type of data, unlike multiview learning, where the views are different feature sets describing the same modality[2]. *Example*: An AI assistant that answers questions by jointly analyzing spoken language (audio), the transcript (text), and an accompanying diagram (image).

Difference between Multiview and Multimodal learning paradigms

- Multiview learning involves multiple perspectives or feature sets of the same modality.
- Multimodal learning integrates fundamentally different modalities of data (e.g., combining text, audio, and images).

2.8 One-class classification

One-class classification focuses on distinguishing a single target class from outliers, often in anomaly detection tasks[5]. Although it has a relationship with the outlier identification and removal task, there is a fundamental difference: the one-class classification task faces a prediction problem, not just the identification of some samples, which is performed only once. *Example*: Detecting manufacturing defects in products, where most samples are non-defective and defects are rare.

2.9 Ordinal regression

Ordinal regression is a type of regression where the target variable has a natural order, but the differences between consecutive levels are not necessarily equal. *Example*: Predicting customer satisfaction on a scale from 1 to 5, where the order matters but the gap between 1 and 2 may not be the same as between 4 and 5. It could be approached as a multiclass classification problem. Still, the error metrics must take into consideration that the amount of error committed by the systems is proportional to the absolute differences between labels, so regular multiclass error classification metrics are not applicable.

3 Other learning paradigms

3.1 Semi-supervised learning

A learning paradigm that leverages a small amount of labelled data along with a large amount of unlabelled data to improve model performance[3]. *Example*: Classifying emails as spam or not with only a small subset of labelled emails.

3.2 Self-supervised learning

A form of unsupervised learning where the model generates pseudo-labels from the data itself to learn useful representations. Technically speaking, the models are trained in a supervised fashion, but the fact that no external labels are required implies that the task is unsupervised. *Example*: Training a model to predict missing words in sentences to improve natural language processing capabilities.

3.3 Active learning

Active learning involves an iterative process where the model selectively queries the most informative instances for labelling[6]. *Example*: A medical diagnosis system that actively asks doctors to label uncertain cases to refine its predictions.

3.4 Federated learning

Federated learning is a decentralized approach where multiple devices or servers collaboratively train a model without sharing their local data, preserving privacy and security. Basically, this paradigm corresponds to a supervised learning problem that must also satisfy restrictions related to data accessibility. *Example*: Training a predictive text model across users' smartphones without transferring their personal typing data to a central server.

---

4 Class activity 1. Identify the setup/paradigm

Every subsection describes a problem that could be solved using ML techniques. Identify the ML learning paradigm or setup that better matches the description. Support your answer.

Case study 1.

A company is developing a handwriting recognition model that can categorize each handwritten sample under multiple styles (e.g., cursive, print, bold).

Answer: ______________________________

Case study 2.

A machine learning system detects unusual network activity that could indicate cybersecurity threats, but it only has access to normal traffic patterns.

Answer: ______________________________

Case study 3.

A multinational company wants to train an AI assistant across different regions, but local governments do not allow the data collected in their countries to be shared with other countries.

Answer: ______________________________

Case study 4.

A company builds a recommendation system that suggests both movies and books based on user preferences.

Answer: ______________________________

Case study 5.

A customer support chatbot improves itself by asking users to confirm whether its responses were helpful or not.

Answer: ______________________________

Case study 6.

Medical imaging systems detect cancerous tumors by analyzing multiple scans of the same patient, and a diagnosis is made based on at least one positive scan.

Answer: ____________________

Case study 7.

A speech recognition system that must predict both text and emotion from a speaker's audio input.

Answer: ____________________

Case study 8.

A ride-sharing company aims to rank customer feedback based on service quality, with ratings ranging from "poor" to "excellent."

Answer: ____________________

Case study 9.

An online product review system collects multiple ratings from different users for the same product, where the final rating is determined by aggregating the opinions of multiple users.

Answer: ____________________

Case study 10.

A financial institution develops an algorithm to identify customers at risk of defaulting on loans, but the number of defaulters is much smaller than that of non-defaulters.

Answer: ____________________

Case study 11.

A medical diagnostic tool combines radiology images, patient medical history (text), and lab test results (numeric data) to detect diseases.

Answer: ____________________

Case study 12.

A manufacturing quality control system uses multiple camera angles (top, side, and zoomed views) of the same product to detect defects.

Answer: ____________________

Case study 13.

A global AI system aims to train a model across multiple hospitals without transferring patient data.

Answer: ____________________

Case study 14.

A hospital aims to detect rare genetic disorders in newborns, but the dataset contains significantly more healthy samples than affected ones.

Answer: ____________________

Case study 15.

An online tutoring system uses both students' facial expressions during video lessons and their clickstream behavior as input features to predict engagement levels.

Answer: ____________________

5 Class activity 2. Mathematical descriptions

One key element to better understanding the learning paradigm that matches a problem is being able to describe the problem in mathematical terms. The purpose of this exercise is to try to put some of the learning setups and paradigms described in sections 2 and 3 in mathematical terms.

Let's see a couple of examples:

5.1 Multi-output problems

In a multi-output problem, the dataset is going to be described as a set $\mathcal{D}\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{x}_i$ corresponds to the input vector and $\mathbf{y}_i$ to the output target vector. In this case, the output is a vector instead of a scalar (y), which is typical of a standard supervised learning problem. The problem description could continue by clarifying that $\mathbf{y}_i = [y_{i1}, y_{i2}]$, where y_{i1} corresponds to the first target variable and y_{i2} to the second one.

5.2 Multi-labeller problems

In this case, the mathematical description could be very similar to that of the multi-output problems ($\mathcal{D}\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$), but the difference is going to be in the meanings of the components forming the output vector $\mathbf{y}_i$. In this case, y_{i1} corresponds to the label provided by expert 1 and y_{i2} the corresponding label given by expert 2.

5.3 Assignment

Try to provide mathematical descriptions for:

- Multi-label learning
- Multi-instance learning
- Ordinal regression problems
- Multimodal learning

References

[1] Charu C Aggarwal et al. *Neural networks and deep learning*, volume 10. Springer, 2018.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[4] Salvador García, Julián Luengo, Francisco Herrera, et al. *Data preprocessing in data mining*, volume 72 of *Intelligent systems reference library*. Springer, 2015.

[5] Shehroz S Khan and Michael G Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014.

[6] Alaa Tharwat and Wolfram Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820, 2023.

[7] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[8] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.