

RECSM Summer School: Machine Learning for Social Sciences

Session 3.1: Introduction to Unsupervised Learning

Reto Wüest

Department of Political Science and International Relations
University of Geneva



- ① Unsupervised Learning
- ② The Challenge of Unsupervised Learning

Unsupervised Learning

- Recall that in an **unsupervised learning** problem we only have a set of features X_1, X_2, \dots, X_p measured on n observations.
- We cannot make predictions because we do not have an associated response variable Y .
- The goal in unsupervised learning is to **discover patterns** in our measurements on X_1, X_2, \dots, X_p .

- Our focus is on **two types** of unsupervised learning techniques:
 - **Principal components analysis:** Used for data visualization or data pre-processing before supervised learning techniques are applied;
 - **Clustering methods:** Used for discovering unknown subgroups in the data.

The Challenge of Unsupervised Learning

The Challenge of Unsupervised Learning

- In supervised learning, we usually have
 - a clear goal (prediction of Y on the basis of X_1, X_2, \dots, X_p),
 - and we know how to assess the quality of our results (CV, validation on an independent test set).
- Hence, in supervised learning, we can **check our work** by evaluating how well our model $\hat{f}(X)$ predicts Y on observations not used in fitting $\hat{f}(X)$.

The Challenge of Unsupervised Learning

- Unsupervised learning is often **more challenging** than supervised learning.
 - It is more **subjective** (there is no clear goal such as the prediction of Y),
 - and it is more difficult to **assess the results**.
- This means that in unsupervised learning, we **cannot check our work** because we do not know the true answer (the problem is unsupervised!).