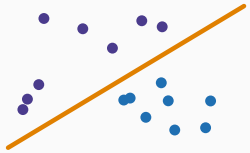
Analyse quantitative

Chapitre 14 : Les méthodes d'apprentissage automatique



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 29 août 2026

Plan du chapitre

Avant de commencer

Deux philosophies

Préparer les données

Deux techniques fondamentales

Synthèse

Avant de commencer

Pourquoi ce chapitre

Pourquoi ce chapitre ?

L'**apprentissage automatique** procède d'une philosophie différente de l'économétrie classique : on ne teste plus une théorie posée a priori, on laisse les données désigner la structure. Ce chapitre expose cette différence de méthode, la préparation des données qu'elle impose, la typologie des approches, et deux techniques fondamentales — l'**ACP** et les **K-moyennes**.

Deux philosophies

La démarche classique

L'analyste postule, à partir d'une théorie économique, un processus générateur et un ensemble de variables. L'algorithme se borne à estimer les paramètres et à tester leur significativité. La conclusion porte sur la **validité de la théorie** initiale.

La démarche par apprentissage

Les données décident des variables à retenir et de la forme de la relation. L'analyste ne teste pas une hypothèse précise : il cherche la structure la plus **prédictive**. Le gain est la capacité à capter des **non-linéarités** et des **interactions** que les modèles linéaires manqueraient.

Une raison technique de ce basculement

Quand le nombre d'observations devient très grand — dizaines de milliers et au-delà —, les erreurs types tendent vers zéro et **presque toute** hypothèse nulle finit par être rejetée. Des variables économiquement négligeables deviennent hautement significatives. Le test d'hypothèses perd ainsi son pouvoir discriminant, ce qui justifie de changer de critère et de juger sur la **performance prédictive**.

Les trois familles

Supervisé, non supervisé, par renforcement

L'apprentissage **supervisé** dispose de données **étiquetées** et vise la prédiction — d'une valeur, comme le prix d'un logement, ou d'une classe, comme « remboursera » ou « fera défaut ». L'apprentissage **non supervisé** travaille sans étiquettes et cherche la structure interne des données. L'apprentissage **par renforcement** apprend par essais et erreurs une suite de décisions dans un environnement changeant.

Applications en finance

Non supervisé : **détection d'anomalies**, quand la banque ne sait pas a priori quelles caractéristiques rendent une transaction suspecte. Supervisé : **score de crédit**, prévision de défaut. Renforcement : exécution optimale d'un ordre de grande taille, gestion dynamique d'un portefeuille, couverture de positions dérivées.

Préparer les données

Un travail majoritaire

Le nettoyage peut absorber jusqu'à **80 %** du temps d'un projet. Il traite les **enregistrements incohérents**, les **observations non pertinentes**, les **doublons** qui biaisent les résultats, les **valeurs aberrantes** à examiner soigneusement, et les **données manquantes** — le problème le plus fréquent.

Que faire des données manquantes

Si elles sont rares, on supprime les observations concernées. Sinon on **impute** : par la moyenne ou la médiane de la variable, ou par estimation à partir des autres variables. Aucune solution n'est neutre — chacune injecte une hypothèse sur la raison de l'absence.

Remise à l'échelle

Standardisation

$$x_{ij}^{\text{std}} = \frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}_i}$$

Chaque variable obtient une moyenne nulle et une variance unitaire.

Normalisation (min-max)

$$x_{ij}^{\text{norm}} = \frac{x_{ij} - x_{i,\min}}{x_{i,\max} - x_{i,\min}}$$

Chaque variable est ramenée dans l'intervalle $[0, 1]$, sans moyenne ni variance imposées.

Laquelle choisir

La **standardisation** est préférable en présence de valeurs extrêmes : la normalisation comprimerait alors l'essentiel des observations dans une plage étroite, peu représentative. Dans les deux cas, la transformation doit s'appliquer à **toutes** les variables explicatives — mais il n'est pas nécessaire de remettre à l'échelle la variable à prédire.

Découper l'échantillon

Trois sous-échantillons

L'échantillon d'**entraînement** sert à estimer les paramètres. L'échantillon de **validation** sert à choisir entre modèles et à régler les hyperparamètres. L'échantillon de **test**, tenu à l'écart jusqu'au bout, fournit la seule mesure honnête de la performance finale.

Sous-ajustement et sur-ajustement

Un modèle **sous-ajusté** est trop rigide pour capter la structure réelle : il échoue partout. Un modèle **sur-ajusté** apprend le bruit de l'échantillon d'entraînement : excellent en apprentissage, médiocre en test. Les remèdes au sur-ajustement sont la **régularisation** (chapitre 15), la simplification du modèle, l'accroissement des données et la **validation croisée**. C'est l'arbitrage biais-variance du chapitre 9, sous un autre nom.

Deux techniques fondamentales

L’analyse en composantes principales

Réduire la dimension

L’**ACP** construit un petit nombre de variables — les **composantes** — **non corrélées** entre elles, restituant l’essentiel de l’information contenue dans un grand nombre de variables corrélées. Les observations s’expriment alors, en bonne approximation, comme combinaisons linéaires de ces composantes.

La courbe des taux

Appliquée à dix ans de variations quotidiennes de taux sur des maturités allant d’un mois à trente ans, l’ACP dégage systématiquement trois composantes interprétables : un **déplacement parallèle** de toute la courbe — de loin la plus importante —, une variation de **pente**, puis une variation de **courbure**. Trois variables se substituent ainsi à une dizaine, sans perte notable.

Un outil déjà rencontré

L’ACP répond aussi à la **multicolinéarité** du chapitre 9, puisqu’elle produit des composantes orthogonales. Elle est enseignée en licence dans le cours d’analyse des données et utilisée en gestion des risques pour réduire la dimension des facteurs de VaR : c’est le même outil, appliqué à trois fins différentes.

L'algorithme des K-moyennes

Le principe

Les **K-moyennes** répartissent les observations en K groupes. On initialise K centres, on affecte chaque observation au centre le plus proche, on recalcule chaque centre comme la moyenne de son groupe, et l'on répète jusqu'à stabilisation.

Les points délicats

Le nombre K doit être fixé **a priori** ; on l'arbitre par la méthode du coude, en observant à partir de quelle valeur l'inertie intra-groupe cesse de baisser franchement. L'algorithme dépend de son **initialisation** et converge vers des optima locaux, d'où l'usage de plusieurs départs. Enfin, il repose sur une **distance**, ce qui rend la remise à l'échelle préalable indispensable.

Usages

Segmenter une clientèle, regrouper des actifs au comportement voisin, identifier des régimes de marché, ou isoler des transactions atypiques dans une démarche de détection de fraude.

Exploiter le texte

Le **traitement du langage naturel** convertit du texte en données exploitables : découpage en unités, normalisation, puis représentation numérique permettant de mesurer la proximité entre documents. On accède ainsi à des sources jusque-là inutilisables quantitativement — rapports annuels, comptes rendus de banques centrales, dépêches, réseaux sociaux — notamment pour construire des indicateurs de **sentiment**.

Synthèse

Ce qu'il faut retenir

Synthèse du chapitre

L'apprentissage automatique renverse la démarche : les données désignent la structure au lieu de valider une théorie posée d'avance, changement d'autant plus justifié que sur de très grands échantillons le test d'hypothèses perd son pouvoir discriminant. Trois familles : **supervisée** (prédiction et classification, données étiquetées), **non supervisée** (structure, détection d'anomalies), **par renforcement** (décisions séquentielles). La préparation des données — **nettoyage**, puis **standardisation** ou **normalisation** — occupe l'essentiel du travail. Le découpage **entraînement / validation / test** est la seule protection contre le **sur-ajustement**, forme renouvelée de l'arbitrage biais-variance. Enfin, l'**ACP** réduit la dimension par composantes orthogonales — trois suffisent pour la courbe des taux — et les **K-moyennes** regroupent les observations, à condition de fixer K et de remettre les variables à l'échelle.