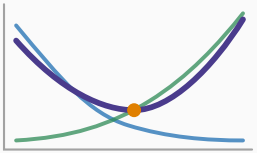
Apprentissage statistique

Chapitre 1 : Expliquer ou prévoir : deux objectifs distincts



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

Prérequis

Deux usages d'un même modèle

Le cadre

Supervisé et non supervisé

La règle absolue

Ce qu'il faut retenir

Prérequis

Ce que ce cours suppose acquis

Les cours qui précèdent

Le niveau Licence dans son ensemble, en particulier l'*Introduction à l'économétrie* et l'*Algèbre linéaire*.

Les notions mobilisées

Les moindres carrés et leur interprétation. L'écriture matricielle, la norme d'un vecteur, l'inversion. La loi normale et la notion d'estimateur.

La pratique de R ou de Python est supposée acquise : le cours ne l'enseigne pas.

Ce que le cours établit lui-même

Ce que ce cours ajoute

L'économétrie estime pour **expliquer**, et juge un modèle sur son ajustement dans l'échantillon.

Ce cours estime pour **prévoir**, et juge un modèle sur ce qu'il fait **hors** échantillon. Le déplacement paraît mince ; il change tous les critères, à commencer par le sens du mot « bon modèle ».

Deux usages d’un même modèle

La distinction qui commande tout le cours

Expliquer et prévoir

Expliquer : obtenir un coefficient interprétable, assorti d'un écart-type, pour répondre à « quel est l'effet de X sur Y ? »

Prévoir : obtenir une erreur faible sur des observations **jamais vues**, pour répondre à « quelle sera la valeur de Y ? »

Ce ne sont pas deux degrés d'une même exigence

Un modèle peut prévoir remarquablement et n'expliquer rien. Un modèle correctement identifié peut prévoir médiocrement.

L'économétrie du niveau Licence poursuit le premier objectif. Ce cours poursuit le second, et il faut accepter dès maintenant ce qu'il faudra abandonner en chemin : l'interprétabilité des coefficients, la significativité, souvent le sens même du modèle.

Un exemple qui sépare nettement les deux

Exemple

Pour prévoir le défaut d'un emprunteur, le **nombre de relances déjà envoyées** est un excellent prédicteur.

Pour comprendre **pourquoi** il fait défaut, cette variable ne sert à rien : elle est une conséquence de la difficulté, non sa cause. La supprimer dégrade la prévision et améliore l'explication.

Le cadre

Ce qu'on cherche à estimer

Le modèle général

$$Y = f(X) + \varepsilon$$

f est la partie systématique, inconnue. ε est un bruit d'espérance nulle, **indépendant de X**, qu'aucun modèle ne peut capter.

Deux sources d'erreur

L'erreur de prévision se décompose en :

- une part **réductible** — l'écart entre notre $\hat{f}$ et le vrai f ;
- une part **irréductible** — la variance de ε.

Tout l'effort porte sur la première. **La seconde fixe une limite qu'aucune méthode ne franchira.**

Pourquoi c'est libérateur

Intuition

Savoir qu'une part de l'erreur est irréductible évite de poursuivre indéfiniment une performance impossible.

Un modèle qui atteint la borne du bruit est parfait, même si son erreur reste élevée. Un modèle qui semble faire mieux que le bruit est **suspect** : il a probablement vu les données de test.

Le fil rouge du cours

Un portefeuille de crédit, du chapitre 1 au chapitre 12

Le cours travaille sur un même jeu : **mille dossiers de crédit**, dont **cent défauts**, décrits par une vingtaine de variables — revenu, ancienneté, encours, historique.

L'objectif est de prévoir le défaut. Ce cadre permet de comparer toutes les méthodes sur les mêmes chiffres, et il porte une difficulté que le chapitre 7 exploitera longuement : les classes sont **déséquilibrées**.

Supervisé et non supervisé

Deux familles

La distinction

Apprentissage supervisé — on dispose de Y, et l'on cherche à le prévoir. Régression si Y est quantitatif, classification s'il est qualitatif.

Apprentissage non supervisé — il n'y a pas de Y. On cherche une structure : groupes, axes, réductions.

Le non supervisé, vous le connaissez déjà

L'analyse en composantes principales, les correspondances, la classification ascendante hiérarchique : c'est tout le cours d'**Analyse des données** du niveau Licence.

Ce cours-ci est donc consacré au **supervisé**, et suppose l'autre acquis.

Le vocabulaire, à traduire

Apprentissage statistique	Économétrie
variables explicatives, attributs	variables exogènes, régresseurs
étiquette, cible	variable dépendante
apprentissage	estimation
observation, exemple	individu

La règle absolue

Ne jamais évaluer sur les données d'apprentissage

Le découpage minimal

Ensemble	Rôle
Apprentissage	estimer les paramètres
Validation	choisir entre modèles et régler les hyperparamètres
Test	mesurer la performance finale, **une seule fois**

Pourquoi l'erreur d'apprentissage ne vaut rien

Un modèle assez flexible peut passer exactement par tous les points d'apprentissage. Son erreur y est nulle, et sa performance sur des données nouvelles peut être catastrophique.

L'erreur d'apprentissage mesure la capacité à **mémoriser**. Seule l'erreur sur données non vues mesure la capacité à **généraliser**.

La fuite de données, erreur la plus coûteuse du métier

Il y a **fuite** quand une information du test s'introduit dans l'apprentissage. Les cas fréquents :

- normaliser ou imputer **avant** de découper — les moyennes utilisées contiennent le test ;

Ce qu’il faut retenir

Le résumé du chapitre

Cinq points

1. **Expliquer** et **prévoir** sont deux objectifs distincts, parfois contradictoires.
2. $Y = f(X) + \varepsilon$: l'erreur a une part **réductible** et une part **irréductible**.
3. Ce cours traite l'apprentissage **supervisé** ; le non supervisé est le cours d'analyse des données.
4. On évalue **toujours** sur des données non vues : apprentissage, validation, test.
5. La **fuite de données** est l'erreur la plus commune et la plus coûteuse.

La suite

Reste à comprendre pourquoi un modèle plus flexible n'est pas toujours meilleur. C'est le compromis biais-variance, et c'est le chapitre 2.