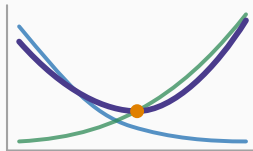


Apprentissage statistique

Chapitre 3 : Rééchantillonnage et validation croisée



Jaouad Madkour

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

Estimer une erreur qu'on n'observe pas

La validation croisée

Les découpages qu'il faut respecter

Le bootstrap

Ce qu'il faut retenir

Estimer une erreur qu'on
n'observe pas

Ce que le chapitre 2 a laissé en suspens

Le compromis biais-variance se règle sur l'**erreur de test**. Or celle-ci n'est pas observable : si on la mesurait pour choisir, elle cesserait d'être une mesure honnête.

Il faut donc l'**estimer** à partir des seules données d'apprentissage.

L'approche par ensemble de validation

Découper une fois : **70 %** pour apprendre, **30 %** pour valider.

Simple, mais deux défauts sérieux :

- l'estimation dépend fortement du **découpage tiré** ;
- on n'apprend que sur une partie des données, ce qui **surestime** l'erreur.

La validation croisée

Le principe des K blocs

Validation croisée à K blocs

1. Partager les données en K blocs de taille égale.
2. Pour chaque bloc : apprendre sur les $K - 1$ autres, évaluer sur celui-ci.
3. Moyenner les K erreurs obtenues.

Sur le portefeuille de crédit

1 000 dossiers, $K = 5$:

	Effectif
Par bloc de validation	200
Pour l'apprentissage	800
Modèles estimés	5

Chaque dossier sert **quatre fois** à l'apprentissage et **une fois** à la validation. Toutes les données servent aux deux usages, sans jamais servir aux deux **en même temps**.

Définition

K	Biais de l'estimation	Variance	Coût
petit (2-3)	élevé — on apprend sur peu	faible	faible
5 à 10	modéré	modérée	modéré
$K = n$ (leave-one-out)	quasi nul	élevée	très élevé

Cinq ou dix constituent le compromis usuel.

Intuition

On pourrait croire que $K = n$ est le meilleur choix, puisqu'on apprend sur presque tout.

Mais les n modèles sont estimés sur des échantillons qui diffèrent d'une seule observation : ils sont **presque identiques**, donc leurs erreurs sont fortement corrélées. Moyenner des quantités corrélées ne réduit guère la variance.

Moins de blocs donne des modèles plus différents, donc une moyenne plus stable.

Les découpages qu'il faut
respecter

Trois situations où le tirage aléatoire est faux

Classes déséquilibrées : la stratification

Avec 10 % de défauts, un bloc tiré au hasard peut n'en contenir presque aucun.

La **validation croisée stratifiée** conserve la proportion de chaque classe dans chaque bloc. Elle est la règle dès que les classes sont déséquilibrées.

Données groupées

Si un même client apparaît dans plusieurs lignes, il ne doit pas se retrouver à la fois en apprentissage et en validation — sinon le modèle le reconnaît.

On découpe alors **par groupe**, non par ligne.

Séries temporelles

Un découpage aléatoire ferait apprendre sur le futur pour prévoir le passé.

On procède par **fenêtre glissante** : apprendre jusqu'à t , valider sur $t + 1$, avancer. L'ordre chronologique est toujours respecté.

Application en gestion

Le cours *Modélisation ARMA* évalue ses prévisions exactement ainsi : estimation sur une fenêtre, prévision hors échantillon, puis glissement.

C'est la même idée que la validation croisée, adaptée à une contrainte que les données transversales n'ont pas — le temps ne se mélange pas.

Le bootstrap

Un second outil de rééchantillonnage

Bootstrap

Tirer, **avec remise**, un échantillon de même taille que l'original, et répéter un grand nombre de fois.

La dispersion des résultats obtenus estime la variabilité de la statistique étudiée.

Deux usages, deux questions

La **validation croisée** répond à : quelle sera l'erreur sur des données nouvelles ?

Le **bootstrap** répond à : quelle est l'incertitude de mon estimation ?

Ce sont deux questions distinctes, et deux outils distincts — souvent confondus.

Le bootstrap fonde les forêts aléatoires

Le chapitre 9 construira des centaines d'arbres, chacun sur un échantillon bootstrap différent, puis les moyennera.

C'est une application directe : rééchantillonner produit des modèles différents, et moyenner des modèles différents **réduit la variance**. Tout le chapitre 2 est là.

Ce qu'il faut retenir

Cinq points

1. L'erreur de test s'**estime** par rééchantillonnage, sans jamais y toucher.
2. La **validation croisée à K blocs** : ici 5 blocs, 200 en validation, 800 en apprentissage.
3. $K = 5$ ou 10 est le compromis usuel; le **leave-one-out** a une variance élevée.
4. **Stratifier** si les classes sont déséquilibrées, découper **par groupe** si les données le sont, **respecter le temps** en séries temporelles.
5. Le **bootstrap** mesure l'incertitude d'une estimation; il fonde les forêts aléatoires.

La suite

Nous savons mesurer. Il reste à disposer de modèles dont on puisse régler la flexibilité de façon continue.

Le chapitre 4 en donne les deux premiers, en partant de la régression linéaire.