

Analyse des données avancée

Chapitre 17 : Classer un nouveau client



Jaouad Madkour

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

La règle d'affectation

Mesurer les erreurs

Le scoring

Ce qu'il faut retenir

La règle d'affectation

Le principe

Affecter au centre le plus proche

On projette le nouvel individu sur les axes discriminants, on calcule sa distance aux K centres de groupes, **et on l'affecte au plus proche.**

C'est la règle géométrique, la plus simple, et celle que nous employons.

Nos trois centres sur l'axe 1

Groupe	Centre
Rare	-0,785
Mensuelle	-0,036
Hebdomadaire	+0,901

Un client dont le score discriminant vaut $+0,6$ sera classé « hebdomadaire » : c'est de ce centre qu'il est le plus proche.

Deux raffinements possibles

Ce que la règle géométrique ignore

Les effectifs. Si un groupe représente 5 % de la clientèle, l'y affecter devrait demander davantage de preuves. La règle bayésienne pondère par les probabilités a priori.

Les coûts d'erreur. Classer à tort un client fidèle en occasionnel n'a pas les mêmes conséquences que l'inverse.

En pratique, sur nos données

Les trois groupes font 112, 87 et 101 clients : les effectifs sont presque égaux, et la pondération a priori ne changerait presque rien.

Les coûts, eux, ne sont jamais égaux dans la vraie vie. Envoyer une offre de reconquête à un client déjà fidèle coûte un timbre ; ne pas l'envoyer à un client sur le départ coûte un client.

Mesurer les erreurs

La matrice de confusion

réel	prédit			total	% exact
	Rare	Mensuelle	Hebdo.		
Rare	78	23	11	112	69,6 %
Mensuelle	24	40	23	87	46,0 %
Hebdomadaire	1	17	83	101	82,2 %
taux global				300	67,0 %
validation croisée					65,3 %
hasard					33,3 %

Le tableau ne se lit pas par son total, mais par ses cases.

67 % de bons classements contre 33 % au hasard : la règle vaut quelque chose. **Mais la ligne du milieu s'effondre à 46 %** — les clients mensuels, coincés entre les deux extrêmes, sont les moins prévisibles. C'est logique : ils sont au centre de l'axe.

Une seule erreur grave : un seul client hebdomadaire classé « Rare ». Les confusions se font toujours entre voisins.

Et le taux calculé sur les données ayant servi à construire la règle est toujours optimiste. La validation croisée — on retire un client, on reconstruit la règle sans lui, on le classe — ramène à 65,3 %. C'est ce chiffre-là qu'on publie.

Trois lectures, du général au détail

Le taux global : 67,0 % contre 33,3 % au hasard. La règle apporte quelque chose.

Le groupe central : 46,0 % seulement. Les clients mensuels, situés au milieu de l'axe, sont les moins prévisibles — et c'est logique.

Les erreurs graves : un seul client hebdomadaire classé « rare ». Les confusions se font entre voisins.

Un taux global peut cacher un échec

Imaginons un groupe rare, 5 % des clients. Une règle qui ne le prédit **jamais** affiche 95 % de bons classements et ne sert à rien.

C'est pourquoi on lit toujours les taux ligne par ligne. Le taux global seul est un mauvais résumé, et il est pourtant celui que les rapports mettent en avant.

Le problème du resubstitution

Classer les mêmes individus qui ont servi à construire la règle donne un taux **systématiquement trop bon** : la règle a été ajustée sur eux.

Ce chiffre-là ne se publie pas.

La validation croisée, un contre tous

On retire un client, on reconstruit toute l'AFD sans lui, on le classe, on recommence trois cents fois.

Résultat : **65,3 %** au lieu de **67,0 %**. L'écart est ici modeste — signe que la règle n'est pas surajustée — mais il aurait pu atteindre dix points sur un petit échantillon.

Le scoring

Ce qu'on utilise vraiment en entreprise

Non pas la classe prédite, mais **le score lui-même** : la coordonnée sur l'axe discriminant, éventuellement transformée en probabilité.

On ordonne alors les clients du plus au moins probable, et l'on cible autant de clients que le budget le permet.

La différence est considérable

La règle dit : « ce client sera hebdomadaire ». **Le score dit : « ce client est à +0,62, au 78^e rang sur 300 ».**

Avec un budget pour deux mille courriers et cinquante mille clients, la classe prédite ne sert à rien — elle en désignerait quinze mille. Le score, lui, désigne les deux mille premiers.

Deux réserves à écrire dans le rapport

La règle prédit une association, pas un comportement. Un client dont le score est élevé ressemble aux clients fidèles; rien ne dit qu'il le deviendra.

Et le sens de la relation reste inconnu. Le client note bien le service parce qu'il vient souvent, ou vient souvent parce que le service est bon? L'AFD sépare les groupes; elle n'explique pas pourquoi ils diffèrent.

Où se trouve la suite

Établir un effet demanderait un dispositif que l'analyse des données ne fournit pas : une expérimentation, ou les méthodes d'inférence causale.

C'est très exactement le seuil où s'arrêtait le cours d'économétrie — et c'est le même pour toutes les méthodes descriptives.

Ce qu'il faut retenir

Six points

1. La règle géométrique affecte au **centre le plus proche**.
2. Deux raffinements : les **effectifs a priori** et les **coûts d'erreur**.
3. La **matrice de confusion** se lit ligne par ligne, jamais par son seul total.
4. Chez nous : **67,0 %** global, mais **46,0 %** sur le groupe central.
5. Le taux calculé sur les données d'apprentissage est optimiste : **validation croisée à 65,3 %**.
6. En entreprise, on utilise le **score** et non la classe : il permet de trier et de cibler.

La suite

Toutes nos méthodes partaient d'un tableau de mesures.

Le dernier chapitre de méthode part d'autre chose : un tableau où l'on n'a mesuré aucune variable, seulement demandé aux gens si deux choses se ressemblent.