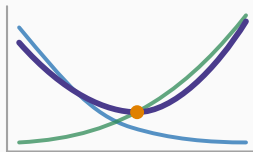


Apprentissage statistique

Chapitre 6 : La classification : logistique et analyse discriminante



Jaouad Madkour

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

Pourquoi la régression linéaire ne convient pas

La régression logistique

L'analyse discriminante

Le seuil de décision

Ce qu'il faut retenir

Pourquoi la régression linéaire ne
convient pas

Deux objections

Le problème

La cible est **qualitative** : défaut ou non. On code $Y = 1$ pour le défaut, $Y = 0$ sinon.

Régresser directement Y sur X par les MCO pose deux difficultés.

Les deux objections

1. Les prédictions **sortent de l'intervalle** $[0, 1]$: on obtient des « probabilités » négatives ou supérieures à un.
2. Avec plus de deux catégories **non ordonnées**, le codage numérique impose un ordre arbitraire qui n'existe pas.

Ce qu'il faut donc

Une fonction qui prenne n'importe quelle combinaison linéaire de X et la ramène **dans** $[0, 1]$.

C'est exactement ce que fait la fonction logistique.

La régression logistique

La fonction logistique

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Elle vaut toujours entre 0 et 1, quelle que soit la valeur de $\beta_0 + \beta_1 X$.

La forme linéaire cachée

En isolant, on obtient la **cote** puis son logarithme :

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \quad \Rightarrow \quad \ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

Le **logarithme de la cote** est linéaire en X . C'est ce qui rend le modèle maniable.

Interpréter un coefficient

β_1 n'est **pas** l'effet sur la probabilité, qui dépend du point où l'on se place.

e^{β_1} est le facteur par lequel la **cote** est multipliée quand X augmente d'une unité.

L'analyse discriminante

Une approche par les distributions

Le renversement du problème

La logistique modélise directement $P(Y | X)$.

L'**analyse discriminante** modélise la distribution de X dans chaque classe, puis retourne le raisonnement par le théorème de Bayes :

$$P(Y = k | X) \propto \pi_k f_k(X)$$

où π_k est la proportion de la classe k et f_k la densité de X dans cette classe.

C'est le théorème de Bayes du chapitre 8 des Probabilités

La structure est identique à celle du test médical : on connaît la probabilité du symptôme sachant la maladie, on veut celle de la maladie sachant le symptôme.

Ici, on connaît la distribution des revenus chez les défaillants, on veut la probabilité de défaut sachant le revenu.

Linéaire ou quadratique

Le seuil de décision

Une probabilité n'est pas une décision

Deux étapes distinctes

Le modèle produit une **probabilité**. La décision exige un **seuil** au-delà duquel on classe en défaut.

Le seuil de 0,5 est un choix par défaut, presque toujours mauvais.

Pourquoi 0,5 ne convient pas ici

Sur notre portefeuille, les défauts représentent 10 % des dossiers. Peu de clients atteindront une probabilité supérieure à 0,5 : le modèle classera presque tout le monde en « bon ».

Il faut donc **abaisser le seuil** si l'on veut détecter des défauts.

Application en gestion

Le bon seuil dépend du coût des deux erreurs :

- refuser un bon client coûte une marge perdue;
- accorder un crédit à un futur défaillant coûte le capital.

Le second coût dépasse largement le premier dans la plupart des portefeuilles. Le seuil optimal est donc **bien inférieur à 0,5**.

Aucune méthode statistique ne fixe ce seuil : il vient de l'économie du problème. Le chapitre 7 donne les outils pour le choisir en connaissance de cause.

Ce qu'il faut retenir

Cinq points

1. La régression linéaire donne des probabilités **hors de** $[0, 1]$: elle ne convient pas.
2. La **logistique** modélise $\ln(p/(1 - p))$ comme une fonction linéaire de X .
3. e^β est un facteur multiplicatif sur la **cote**, non sur la probabilité.
4. L'**analyse discriminante** passe par les distributions de X et le théorème de Bayes.
5. Le **seuil** de décision est un choix économique, pas statistique — et 0,5 est rarement le bon.

La suite

Il faut maintenant savoir mesurer la qualité d'un classifieur — et le chapitre 7 montrera que la mesure la plus naturelle est aussi la plus trompeuse.