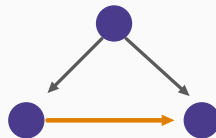


Inférence causale et machine learning

Chapitre 5 : Appariement et pondération par le score de propension



Comparer ce qui est comparable

Le score de propension

La vérification qui compte

La limite de fond

Ce qu'il faut retenir

Comparer ce qui est comparable

Le principe

Sous indépendance conditionnelle, il suffit de comparer traités et non traités à **caractéristiques X égales**, puis d'agréger.

$$ATT = E_X[E[Y | T=1, X] - E[Y | T=0, X] | T = 1]$$

Appariement exact

Pour chaque traité, trouver un non traité ayant **exactement** les mêmes caractéristiques, et comparer.

Pourquoi cela échoue vite

Avec cinq variables binaires, il y a déjà **32** profils. Avec dix, plus de mille. Avec une variable continue, aucun appariement exact n'existe.

C'est la **malédiction de la dimension** du chapitre 2 d'*Apprentissage statistique*, rencontrée ici sous une autre forme.

Le score de propension

Résumer X en un seul nombre

Score de propension

$$e(X) = P(T = 1 \mid X)$$

C'est la probabilité d'être traité, compte tenu des caractéristiques observées.

Le théorème de Rosenbaum et Rubin

Si l'indépendance conditionnelle vaut sachant X , alors elle vaut sachant $e(X)$ seul :

$$(Y(0), Y(1)) \perp T \mid X \implies (Y(0), Y(1)) \perp T \mid e(X)$$

Pourquoi ce résultat est précieux

Il ramène un problème de **dimension quelconque** à un problème de **dimension un**.

Apparier sur vingt variables est impossible ; apparier sur un score compris entre 0 et 1 est immédiat.

Le score de propension ne crée aucune information nouvelle : il réorganise celle de X . Si X est insuffisant, le score l'est aussi.

Trois usages

Ce qu'on peut en faire

Méthode	Principe
Appariement	associer chaque traité au non traité de score le plus proche
Stratification	découper en tranches de score, comparer dans chacune
Pondération	repondérer les observations pour équilibrer les groupes

La pondération par l'inverse de la probabilité

On donne à chaque observation un poids $1/e(X)$ si elle est traitée, $1/(1 - e(X))$ sinon.

Les traités **rare**s pour leur profil reçoivent un poids élevé, et réciproquement : la population repondérée ressemble à ce qu'aurait produit un tirage au sort.

Application en gestion

Un score très proche de 0 ou de 1 produit un poids **immense**.

Quelques observations dominent alors toute l'estimation, dont la variance explose. Deux remèdes : **écrêter** les poids extrêmes, ou **restreindre** l'analyse à la zone de support commun.

Un tableau de poids doit toujours être examiné avant de publier un résultat.

La vérification qui compte

L'équilibre, et non l'ajustement du score

Ce qu'il faut vérifier

Après appariement ou pondération, comparer la distribution de **chaque** variable entre les deux groupes.

La **différence standardisée** est l'indicateur usuel : on la souhaite inférieure à 0,1 en valeur absolue.

L'erreur la plus fréquente

On juge parfois le score de propension à sa capacité à **prédire** le traitement — son AUC.

C'est un contresens. Un score qui prédit parfaitement le traitement signifie qu'il n'existe **aucun** chevauchement : traités et non traités sont parfaitement séparables, donc incomparables.

Le score doit équilibrer, pas prédire. Un excellent classifieur fait ici un mauvais score de propension.

Le diagnostic de support commun

Tracer les distributions du score dans les deux groupes. Elles doivent **se chevaucher** largement.

Les zones sans chevauchement doivent être écartées de l'analyse — et cela change la population sur laquelle porte le résultat, ce qu'il faut dire.

La limite de fond

Ce que la méthode ne peut pas faire

Elle ne traite que l'observable

Appariement et pondération reposent entièrement sur l'**indépendance conditionnelle** : toutes les variables qui affectent à la fois le traitement et le résultat doivent être **observées**.

Si la motivation détermine la participation à la formation **et** l'emploi, et qu'elle n'est pas mesurée, aucune sophistication du score n'y remédie.

Ces méthodes ne créent pas d'identification : elles exploitent celle qu'on suppose.

Pourquoi elles restent utiles

Elles conviennent lorsqu'on dispose de données administratives riches, où les déterminants principaux sont mesurés.

Elles rendent aussi **transparent** ce qu'une régression fait implicitement : la régression pondère elle aussi, mais sans montrer ses poids ni son support commun. Le score de propension oblige à regarder les deux.

Définition

Les chapitres 6 à 9 changent de stratégie : au lieu de supposer l'observable suffisant, ils exploitent une **source de variation extérieure** — le temps, un seuil, un instrument.

C'est un argument d'identification plus solide, quand une telle source existe.

Ce qu'il faut retenir

Cinq points

1. L'appariement **exact** échoue par la dimension.
2. Le **score de propension** $e(X) = P(T=1 \mid X)$ ramène le problème à une dimension.
3. Trois usages : appariement, stratification, **pondération** par l'inverse de la probabilité.
4. On vérifie l'**équilibre** des covariables, pas la capacité prédictive du score.
5. Tout repose sur l'**indépendance conditionnelle** : la méthode ne traite que l'observable.

La suite

Le chapitre 6 exploite une autre source de comparaison : non plus des individus semblables, mais **le même individu avant et après**, corrigé de ce qui aurait bougé de toute façon.