Introduction aux méthodes statistiques

Modèle

Thierry Dhorne

Université de Bretagne-Sud

Licence

Année Universitaire 2014-2015

Modèle multidimensionnel

Données

- l'information de base consiste en un tableau individus $\times$ variables
- on dispose pour chaque individu de la valeur de chacune des variables
- les individus sont représentés par les lignes
- les variables sont représentées par les colonnes

Modèle multidimensionnel
Organisation

- l'organisation générale du tableau est donc

Tableau individus $\times$ variables

	$var\ i$				
	x_1^1	x_1^2	$\dots$	x_1^i	$\dots$	x_1^p
	x_2^1	x_2^2	$\dots$	x_2^i	$\dots$	x_2^p
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$ind\ r$	x_r^1	x_r^2	$\dots$	x_r^i	$\dots$	x_r^p
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	x_n^1	x_n^2	$\dots$	x_n^i	$\dots$	x_n^p

Modèle multidimensionnel

Données

- on note x_r^i la valeur prise par la variable i sur l'individu r
- on réserve les lettres x, y, z aux fonctions des variables

→ voir dans la suite

- on réserve les lettres r, s, t,... aux individus
- et les lettres i, j, k,... aux variables
- la nature (type) d'une variable est tout à fait générale

→ voir dans la suite

Trois statistiques ?

- on peut grossièrement regrouper les méthodes statistiques en trois grands types

→ statistique explicative (et prédictive)

→ statistique extractive (et/ou générative)

→ statistique auto-projective (et prévisionnelle)

★ cette classification est imparfaite mais utile pour le diagnostic statistique

→ quelle méthode pour quel problème ?

- le tableau de données

Tableau individus $\times$ variables structuré

$$\begin{array}{ccccc}
x_1^1 & x_1^2 & \dots & x_1^i & \dots & x_1^p \\
x_2^1 & x_2^2 & \dots & x_2^i & \dots & x_2^p \\
\vdots & \vdots & & \vdots & & \vdots \\
x_r^1 & x_r^2 & \dots & x_r^i & \dots & x_r^p \\
\vdots & \vdots & & \vdots & & \vdots \\
x_n^1 & x_n^2 & \dots & x_n^i & \dots & x_n^p
\end{array}$$

→ peut être structuré en deux groupes de variables

- les variables d'intérêt (souvent une seule) (en rouge)

Tableau individus $\times$ variables structuré

$$\begin{array}{|c|c|c|c|c|c|}
\hline
x_1^1 & x_1^2 & \dots & x_1^i & \dots & x_1^p \\
x_2^1 & x_2^2 & \dots & x_2^i & \dots & x_2^p \\
\vdots & \vdots & & \vdots & & \vdots \\
x_r^1 & x_r^2 & \dots & x_r^i & \dots & x_r^p \\
\vdots & \vdots & & \vdots & & \vdots \\
x_n^1 & x_n^2 & \dots & x_n^i & \dots & x_n^p \\
\hline
\end{array}$$

- les variables annexes (en vert)

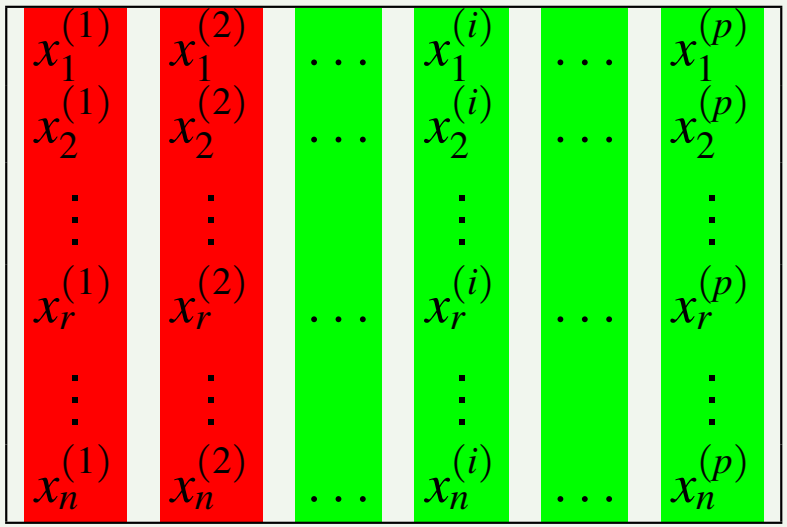
Tableau individus $\times$ variables structuré

$$\begin{array}{ccccccc} x_1^1 & x_1^2 & \dots & x_1^i & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^i & \dots & x_2^p \\ \vdots & \vdots & & \vdots & & \vdots \\ x_r^1 & x_r^2 & \dots & x_r^i & \dots & x_r^p \\ \vdots & \vdots & & \vdots & & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^i & \dots & x_n^p \end{array}$$

- on peut restructurer le tableau

→ pour des raisons de facilité

Tableau individus $\times$ variables restructuré



Statistique explicative (prédictive)
Objectif

- on peut alors expliquer et/ou prédire les variables d'intérêt à partir des variables annexes

Tableau individus $\times$ variables pour l'explication et la prédiction

$$\begin{pmatrix} y_1^1 & y_1^2 \\ y_2^1 & y_2^2 \\ \vdots & \vdots \\ y_r^1 & y_r^2 \\ \vdots & \vdots \\ y_n^1 & y_n^2 \end{pmatrix} \sim \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^q \\ x_2^1 & x_2^2 & \dots & x_2^q \\ \vdots & \vdots & \vdots & \vdots \\ x_r^1 & x_r^2 & \dots & x_r^q \\ \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^q \end{pmatrix}$$

★ le symbole $\sim$ signifie «modélisé par»

Exemple
Caractéristiques automobiles

- on dispose d'un jeu de données présentant les caractéristiques de différents véhicules

	CYL	PUI	LON	LAR	POI	VIT	HAU	PRI
citroenAX	954	50.0	3.52	1.55	706	151.0	1.35	53500
citroenZX	1124	60.0	4.07	1.70	935	161.0	1.40	69800
citroenZXb	1360	75.0	4.26	1.70	1015	165.0	1.45	85000
citroenC15	1124	60.0	3.99	1.64	890	144.0	1.80	78900
citroenXantia	1580	89.0	4.44	1.75	1170	175.0	1.38	101300
citroenXM	1998	135.0	4.71	1.79	1395	205.0	1.39	149500
citroenXMb	1998	135.0	4.96	1.79	1467	198.0	1.46	159500
citroenEvasion	1998	123.0	4.45	1.81	1510	177.0	1.71	139000
peugeot106	954	50.0	3.56	1.57	780	150.0	1.37	56700
peugeot205	1124	60.0	3.70	1.56	765	164.0	1.38	64200
peugeot306	1124	60.0	3.99	1.69	980	155.0	1.38	74900
peugeot306cab	1761	103.0	4.14	1.69	1220	182.0	1.36	138400

Exemple
Variables

- les variables fournies sont
  - la cylindrée (en cm^2)
  - la puissance (en cv)
  - la longueur (en m)
  - la largeur (en m)
  - le poids (en kg)
  - la vitesse limite (en km/h)
  - la hauteur (en m)
  - le prix (en Fr)
- pour la plupart des gens une variable est privilégiée

→ le prix

→ la seule question pertinente sur ce jeu de données est celui de la dépendance du prix par rapport aux variables techniques

Exemple
Prix des automobiles

- l'objectif consiste à expliquer le prix à partir des caractéristiques des véhicules

PRI		CYL	PUI	LON	LAR	POI	VIT	HAU
53500		954	50.0	3.52	1.55	706	151.0	1.35
69800		1124	60.0	4.07	1.70	935	161.0	1.40
85000		1360	75.0	4.26	1.70	1015	165.0	1.45
78900		1124	60.0	3.99	1.64	890	144.0	1.80
101300		1580	89.0	4.44	1.75	1170	175.0	1.38
149500	$\sim$	1998	135.0	4.71	1.79	1395	205.0	1.39
159500		1998	135.0	4.96	1.79	1467	198.0	1.46
139000		1998	123.0	4.45	1.81	1510	177.0	1.71
56700		954	50.0	3.56	1.57	780	150.0	1.37
64200		1124	60.0	3.70	1.56	765	164.0	1.38
74900		1124	60.0	3.99	1.69	980	155.0	1.38
138400		1761	103.0	4.14	1.69	1220	182.0	1.36

Expliquer le prix des automobiles

- l'objectif est donc un objectif d'explication

Explication du prix d'une automobile

prix = f(cylindrée,puissance,longueur,largeur,poids,
vitesse limite,hauteur)

→ pour comprendre

→ pour choisir

★ bien sûr la réalité est plus complexe

→ la puissance dépend de la cylindrée

→ la vitesse limite dépend de la puissance du poids et de la hauteur

- mais notre modèle est à la fois simple et pertinent

Méthodes explicatives et prédictives

- les informaticiens les appellent méthodes supervisées

→ on connaît la (les) variable(s) cible(s)

- on met au point un modèle

→ explication (validation)

- on utilise le modèle

→ prédiction

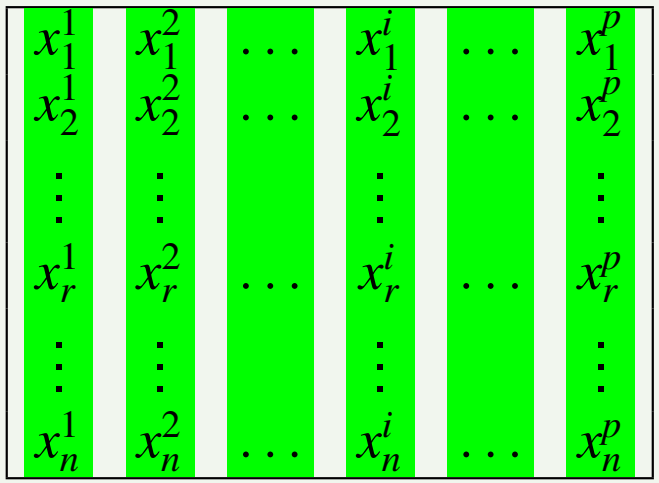
- il s'agit en particulier de méthodes régressives

→ régression linéaire, qualitative, logistique,...

- le tableau de données n'est pas structuré en groupes

→ toutes les variables sont jugées (a priori) aussi informatives

Tableau individus $\times$ variables non structuré



- mais on postule l'existence de variables potentiellement plus intéressantes

→ synthétiques, structurantes,..

Tableau individus $\times$ variables
extraction d'information

$$\begin{array}{cc} y_1^1 & y_1^2 \\ 1 & \\ 2 & \\ \vdots & \vdots \\ y_r^1 & y_r^2 \\ \vdots & \vdots \\ y_n^1 & y_n^2 \end{array} \sim \begin{array}{ccccccc} x_1^1 & x_1^2 & \dots & x_1^i & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^i & \dots & x_2^p \\ \vdots & \vdots & & \vdots & & \vdots \\ x_r^1 & x_r^2 & \dots & x_r^i & \dots & x_r^p \\ \vdots & \vdots & & \vdots & & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^i & \dots & x_n^p \end{array}$$

→ obtenues à partir des variables de départ

Exemple
Caractéristiques techniques de véhicules

- on peut reprendre l'exemple évoqué plus haut

→ en se limitant aux... variables techniques

  - la cylindrée (en cm^2)
  - la puissance (en cv)
  - la longueur (en m)
  - la largeur (en m)
  - le poids (en kg)
  - la vitesse limite (en km/h)
  - la hauteur (en m)

- une question pertinente est alors de faire la synthèse des caractéristiques techniques
  - forme ?
  - performance ?

De nouvelles variables synthétiques

F_1	F_2		CYL	PUI	LON	LAR	POI	VIT	HAU
?	?		954	50.0	3.52	1.55	706	151.0	1.35
?	?		1124	60.0	4.07	1.70	935	161.0	1.40
?	?		1360	75.0	4.26	1.70	1015	165.0	1.45
?	?		1124	60.0	3.99	1.64	890	144.0	1.80
?	?		1580	89.0	4.44	1.75	1170	175.0	1.38
?	?	$\sim$	1998	135.0	4.71	1.79	1395	205.0	1.39
?	?		1998	135.0	4.96	1.79	1467	198.0	1.46
?	?		1998	123.0	4.45	1.81	1510	177.0	1.71
?	?		954	50.0	3.56	1.57	780	150.0	1.37
?	?		1124	60.0	3.70	1.56	765	164.0	1.38
?	?		1124	60.0	3.99	1.69	980	155.0	1.38
?	?		1761	103.0	4.14	1.69	1220	182.0	1.36

Méthodes extractives et génératives

- les informaticiens les appellent méthodes non supervisées
- → on ne connaît pas la (les) variable(s) cible(s)
- on utilise un modèle ou (parfois un critère)*
- il s'agit traditionnellement de méthodes
  - factorielles
  - de classification (segmentation, clustering)

- il existe une structure (temporelle, spatiale,...) sur les individus

→ cette structure n'est pas prise en compte dans le tableau

★ les individus ne sont pas interchangeables

tableau individus $\times$ variables

$$\begin{array}{ccccccc}
x_1^1 & x_1^2 & \dots & x_1^i & \dots & x_1^p \\
x_2^1 & x_2^2 & \dots & x_2^i & \dots & x_2^p \\
\vdots & \vdots & & \vdots & & \vdots \\
x_r^1 & x_r^2 & \dots & x_r^i & \dots & x_r^p \\
\vdots & \vdots & & \vdots & & \vdots \\
x_n^1 & x_n^2 & \dots & x_n^i & \dots & x_n^p
\end{array}$$

- ce cas ne sera pas étudié dans la suite