Analyse de Données – ID

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – FRE 2729
Site de l'Ecole Polytechnique de l'université de Nantes

Polytech'Nantes
École d'ingénieurs de l'université de Nantes

UNIVERSITÉ DE NANTES

lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Le programme

Volume horaire (approximatif)

- 10 séances de cours/TD
- 6 séances de TP
- 1 mini-projet
- 1 examen écrit

Le site du cours

- http://madoc.univ-nantes.fr/course/view.php?id=15157
- en création ...
- les supports de cours/TD, les sujets de TP
- les informations sur le mini-projet
- un forum pour les questions-réponses

Le plan

objective	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	
	1	0	0	0	0	4854	0	0	0	0	0	0
	1	1	0	0	0,94	5504	0	0	0,99	0	0	0
	1	0,97	0	1	0	11178	0	0	0	0	0	0
	1	0,97	0	0	0	2628	0	0	0	0	0	0
	0	0	0	0,99	0	3712	0	0	0	0,97	0	0
	1	0,96	0,99	1	0	5213	0	0	0	0,97	0,97	0
	1	0	0	0	0	5483	0,05	0,41	0	0	0	0
	0	0	0	0,99	0	7926	0	0	0	0,96	1	0
	1	0,93	0,98	1	0	6743	0	0	0	0	0	0
	1	1	0	0	0	6449	0	0	0	0	0	0
	1	0	0	0	0,94	3547	0	0	0	0	0	0
	0	0	0	0	0	2324	0	0	0	0,96	0	0
	1	0	0	0	0	1924	0	0	0	0	0	0
	1	0,99	0	1	0	6292	0,01	0,05	0	0	0,99	0
	0	0	0	1	0	1353	0	0	0	0	0	0
	1	0	0	1	0,97	11079	0	0	0	0	0	0
	1	0	0	0	0	4810	0	0	0	0	0	0
	0	0	0	0	0	1991	0	0	0	0	0	0
	1	0,96	0,99	0	0	1592	0	0	0	0,98	0	0

Le point de départ : des données

un exemple : webmining, des milliers de visiteurs, des centaines de caractéristiques

Le plan

objective	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	
1	0	0	0	0	4854	0	0	0	0	0	0	
1	1	0	0	0,94	5504	0	0	0,99	0	0	0	
1	0,97	0	1	0	11178	0	0	0	0	0	0	
1	0,97	0	0	0	2628	0	0	0	0	0	0	
0	0	0	0,99	0	3712	0	0	0	0,97	0	0	
1	0,96	0,99	1	0	5213	0	0	0	0,97	0,97	0	
1	0	0	0	0	5483	0,05	0,41	0	0	0	0	
0	0	0	0,99	0	7926	0	0	0	0,96	1	0	
1	0,93	0,98	1	0	6743	0	0	0	0	0	0	
1	1	0	0	0	6449	0	0	0	0	0	0	
1	0	0	0	0,94	3547	0	0	0	0	0	0	
0	0	0	0	0	2324	0	0	0	0,96	0	0	
1	0	0	0	0	1924	0	0	0	0	0	0	
1	0,99	0	1	0	6292	0,01	0,05	0	0	0,99	0	
0	0	0	1	0	1353	0	0	0	0	0	0	
1	0	0	1	0,97	11079	0	0	0	0	0	0	
1	0	0	0	0	4810	0	0	0	0	0	0	
0	0	0	0	0	1991	0	0	0	0	0	0	
1	0,96	0,99	0	0	1592	0	0	0	0,98	0	0	

1. Rappels d'algèbre linéaire

les outils mathématiques

Le plan

objective	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	
1	0	0	0	0	4854	0	0	0	0	0	0	
1	1	0	0	0,94	5504	0	0	0,99	0	0	0	
1	0,97	0	1	0	11178	0	0	0	0	0	0	
1	0,97	0	0	0	2628	0	0	0	0	0	0	
0	0	0	0,99	0	3712	0	0	0	0,97	0	0	
1	0,96	0,99	1	0	5213	0	0	0	0,97	0,97	0	
1	0	0	0	0	5483	0,05	0,41	0	0	0	0	
0	0	0	0,99	0	7926	0	0	0	0,96	1	0	
1	0,93	0,98	1	0	6743	0	0	0	0	0	0	
1	1	0	0	0	6449	0	0	0	0	0	0	
1	0	0	0	0,94	3547	0	0	0	0	0	0	
0	0	0	0	0	2324	0	0	0	0,96	0	0	
1	0	0	0	0	1924	0	0	0	0	0	0	
1	0,99	0	1	0	6292	0,01	0,05	0	0	0,99	0	
0	0	0	1	0	1353	0	0	0	0	0	0	
1	0	0	1	0,97	11079	0	0	0	0	0	0	
1	0	0	0	0	4810	0	0	0	0	0	0	
0	0	0	0	0	1991	0	0	0	0	0	0	
1	0,96	0,99	0	0	1592	0	0	0	0,98	0	0	

2. Analyse en composantes principales

comment réduire la taille de l'espace des caractéristiques ?

Le plan

objective	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	
1	0	0	0	0	4854	0	0	0	0	0	0	
1	1	0	0	0,94	5504	0	0	0,99	0	0	0	
1	0,97	0	1	0	11178	0	0	0	0	0	0	
1	0,97	0	0	0	2628	0	0	0	0	0	0	
0	0	0	0,99	0	3712	0	0	0	0,97	0	0	
1	0,96	0,99	1	0	5213	0	0	0	0,97	0,97	0	
1	0	0	0	0	5483	0,05	0,41	0	0	0	0	
0	0	0	0,99	0	7926	0	0	0	0,96	1	0	
1	0,93	0,98	1	0	6743	0	0	0	0	0	0	
1	1	0	0	0	6449	0	0	0	0	0	0	
1	0	0	0	0,94	3547	0	0	0	0	0	0	
0	0	0	0	0	2324	0	0	0	0,96	0	0	
1	0	0	0	0	1924	0	0	0	0	0	0	
1	0,99	0	1	0	6292	0,01	0,05	0	0	0,99	0	
0	0	0	1	0	1353	0	0	0	0	0	0	
1	0	0	1	0,97	11079	0	0	0	0	0	0	
1	0	0	0	0	4810	0	0	0	0	0	0	
0	0	0	0	0	1991	0	0	0	0	0	0	
1	0,96	0,99	0	0	1592	0	0	0	0,98	0	0	

3. Classification automatique

comment réduire la taille de l'espace des individus ?

comment regrouper automatiquement les individus en classes ?

Espace vectoriel

Définition

Un triplet $(E, +, .)$ est un $\mathbb{K}$-*espace vectoriel* si

- $(E, +)$ est un groupe abélien
- et si la multiplication externe : $\mathbb{K} \times E \to E, (\lambda, x) \mapsto \lambda x$ vérifie les propriétés suivantes, $\forall \lambda, \mu \in \mathbb{K}, \forall x, y \in E$:
  - $\lambda(x + y) = \lambda x + \lambda y$
  - $(\lambda + \mu)x = \lambda x + \mu x$
  - $(\lambda\mu)x = \lambda(\mu x)$
  - $1x = x$

En pratique, on essaie d'abord de montrer qu'un espace donné est un sous-espace vectoriel d'un espace vectoriel connu.

Sous-espace vectoriel

Caractérisation :

Pour montrer que F est un sous-espace vectoriel de E, il suffit de prouver que :

- F n'est pas l'ensemble vide,
- F est stable, i.e., $\forall \lambda \in \mathbb{K}, \forall u, v \in F, u + \lambda v \in F$.

Exercice

Combinaison linéaire

Définition

Soit E un $\mathbb{K}$-e.v., soit $n \in \mathbb{N}^*$ et soit $v_1, \ldots, v_n$ une famille finie de vecteurs de E. On appelle combinaison linéaire (c.l.) de la famille $v_1, \ldots, v_n$ tout vecteur v de E s'écrivant de la façon suivante :

$$v = \sum_{i=1}^{n} \alpha_i v_i,$$

où $\alpha_1, \ldots, \alpha_n$ sont des scalaires.

Proposition à démontrer

L'ensemble de tous les vecteurs c.l. des vecteurs $v_1, \ldots, v_n$ est un s.e.v. de E. Ce s.e.v., noté note $Vect(v_1, \ldots, v_n)$, est appelé *s.e.v. engendré par* $v_1, \ldots, v_n$.

Famille

Définition

La famille $v_1, \ldots, v_n$ est génératrice de E ssi $E = Vect(v_1, \ldots, v_n)$

Famille libre

La famille $v_1, \ldots, v_n$ est dite *libre* lorsque $\forall \alpha_1, \ldots, \alpha_n \in \mathbb{K}$,

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \Rightarrow \forall i \in \{1, \ldots, n\}, \alpha_i = 0.$$

Famille liée

La famille $v_1, \ldots, v_n$ est dite *liée* lorsque l'un au moins de ses vecteurs s'écrit comme c.l. des autres.

Exercice

Base d'un espace vectoriel

Définition

Soit $n \in \mathbb{N}^*$ et soit $B = (v_1, \ldots, v_n)$ une famille de vecteurs de E.
B est une *base* de E ssi B est libre et génératrice de E.

Proposition

Soit $n \in \mathbb{N}^*$ et soit $B = (v_1, \ldots, v_n)$ une famille de vecteurs de E.
B est une base de E ssi $\forall v \in E$, $\exists! \alpha_1, \ldots, \alpha_n \in \mathbb{K}$ tels que

$$v = \sum_{i=1}^{n} \alpha_i v_i.$$

Les coefficients $\alpha_1, \ldots, \alpha_n$ sont appelés les *composantes* du vecteur v dans la base B.

Dimension d'un espace vectoriel

Définition

E est de dimension finie s'il existe dans E au moins une famille génératrice finie de E.

Proposition

Soit E un $\mathbb{K}$-e.v. de dimension finie. Alors

- E admet au moins une base finie.
- Toutes les bases de E ont même cardinal, noté $\dim E$ et appelé *dimension* de E.
- Une famille libre de $\dim E$ éléments est une base de E.

Pour montrer qu'une famille de n vecteurs forme une base d'un e.v. de dimension n, il suffit de prouver que la famille est libre.

Dimension d'un espace vectoriel

Exercices

Sous-espaces vectoriels particuliers

Soient G_1 et G_2 deux s.e.v. de E.

Proposition

$G_1 + G_2 = \{x \in E | \exists g_1 \in G_1, g_2 \in G_2 : x = g_1 + g_2\}$ est un s.e.v de E.

Définition : sous-espace en somme directe

G_1 et G_2 sont dits en *somme directe* si tout vecteur de $G_1 + G_2$ s'écrit de façon unique comme la somme d'un vecteur de G_1 et d'un vecteur de G_2.

Proposition

G_1 et G_2 sont en *somme directe* ssi $G_1 \cap G_2 = \{0_E\}$.

Sous-espaces vectoriels particuliers

Soient G_1 et G_2 deux s.e.v. de E.

Définition : sous-espaces supplémentaires

G_1 et G_2 sont dits en *supplémentaires* dans E si tout vecteur de E s'écrit de façon unique comme la somme d'un vecteur de G_1 et d'un vecteur de G_2.
On note alors $E = G_1 \oplus G_2$.

Proposition

G_1 et G_2 sont *supplémentaires* dans E ssi $\forall x \in E$,
$\exists g_1 \in G_1, g_2 \in G_2$ tels que $x = g_1 + g_2$ et $G_1 \cap G_2 = \{0_E\}$.