3

La réduction de dimensions

- Qu’est-ce que c’est ?
- Sélection d’attributs

FACULTÉ DE
GESTION,
ÉCONOMIE
& SCIENCES

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

La réduction de dimensions

- Dimension = autre nom des attributs
- Réduction de dimensions
  - Simplifier le jeu de données
  - Simplifier les modèles / les prédictions
  - Diminuer le nombre d'informations à collecter
  - Big Data : Arriver à traiter un jeu de données trop volumineux

Dimensions
(ici, des gènes codants)

Observations (ici, des patients)

ID	G1	G2	…	G20000	Maladie
I1	O	O		N	O
I2	O	N		N	N
I3	O	N		N	O
…					…
In	N	N		O	N

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Quelques applications de la réduction de dimensions

- Agriculture et élevage : tests génétiques bons marché
  - Vise max. 10 gènes ciblés (au lieu du génome entier)
  - Permet d'identifier :
    - si un animal va produire + de viande, + de lait, + de bébés,…
    - si une plante va pousser vite, produire beaucoup, résister aux nuisibles,…

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Quelques applications de la réduction de dimensions

- Médecine
  - Quelles informations minimum pour visualiser l'évolution d'une maladie donnée ?
  - Quels symptômes/mesures biologiques caractérisent le mieux une maladie ?
  - Permet :
    - De limiter les tests à faire passer au patient
    - De limiter le nombre de questions à poser (consultation ou essai clinique)

Image : © youcef naimi

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Deux techniques

ID	G1	G2	…	G20000	Maladie
I1	O	O		N	O
I2	O	N		N	N
I3	O	N		N	O
…					…
In	N	N		O	N

Sélection d'attributs

Ici, on a gardé seulement les 3 attributs les plus significatifs

ID	G3	G26	G13520	Maladie
I1	O	N	O	O
I2	O	O	O	N
I3	N	N	N	O
…				…
In	N	O	O	N

Génération d'attributs (technique ACP : Analyse en Composantes Principales)

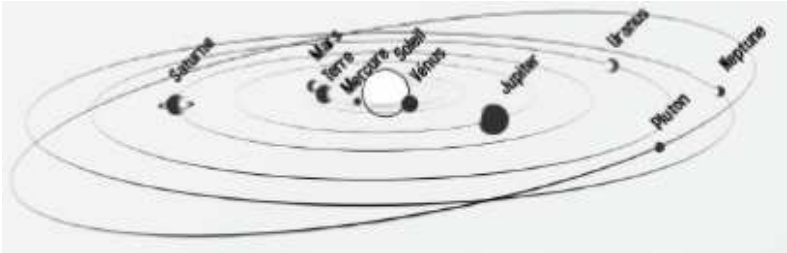
Ici, on a généré les attributs A1 et A2

ID	A1	A2	Maladie
I1	0,25	1,68	O
I2	3,5	2,9	N
I3	1,2	3,1	O
…			…
In	0,9	1,5	N

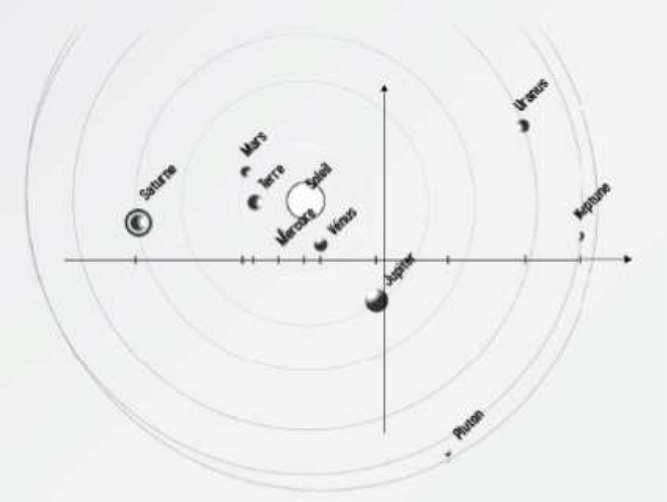
FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Vulgarisation de l'ACP

Système solaire en 3D



Système solaire en 2D



Système solaire en 1D

Illustrations : **Data Science : fondamentaux et études de cas**

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Aller plus loin sur l'ACP…

Pour aller plus loin (pour les amoureux des stats), chapitre 15 de :

Data Science : fondamentaux et études de cas
Machine Learning avec Python et R

Auteur(s): **Lutz, Michel**
Biernat, Eric
Editeur: **Eyrolles**
Année de Publication: 2015
pages: 311
ISBN: 978-2-212-14243-3

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

La sélection d'attributs

- Plus facile à interpréter ☺
- Fonctionnement
  - Calculer un score par attribut
  - Conserver les attributs avec le meilleur score
  - Supprimer les autres

Penser à vérifier que les scores (confiance, sensibilité, CA) ne se sont pas dégradés après le filtrage

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

Scores pour la sélection d'attributs

Plusieurs manières de mesurer « l'efficacité » d'un attribut :

- χ^2 : indique le degré de dépendance statistique entre l'attribut et la classe
- **Gain** : mesure la réduction d'entropie apportée par cet attribut
- **Gini** : mesure du niveau d'inégalité pour une variable et sur une population donnée. Il varie entre 0 (égalité parfaite) et 1 (inégalité extrême).
- …

FACULTÉ DE GESTION, ÉCONOMIE & SCIENCES

TP d'application sur l'insuffisance cardiaque

- 299 patients
- 12 informations par patient (age, fumeur?,diabète?, créatinine,…)
- Classe : décès
- Quels informations surveiller en priorité ?