Quelques éléments de machine learning

Aurélien Nicosia

École interdisciplinaire outils et méthodes
Cheminement 2 - Science des données : modélisation et prédiction

Apprentissage statistique (*machine learning*)

31 aout 2023

Plan

Introduction

Correlation : Qu’en savez-vous ?

Qu’est ce que la corrélation ?

Coefficient de corrélation de Pearson

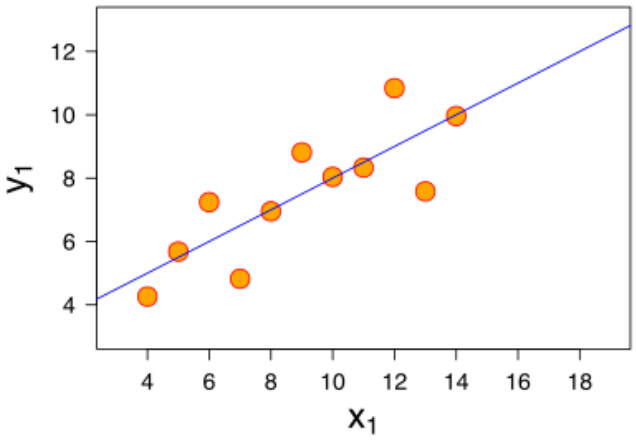
Voici la définition :

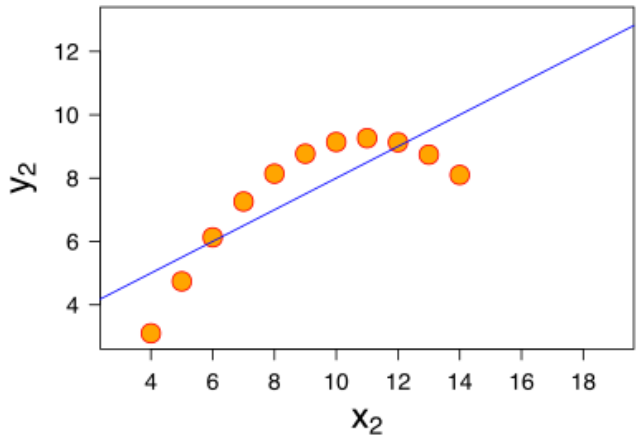
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{(n-1)s_x s_y}$$

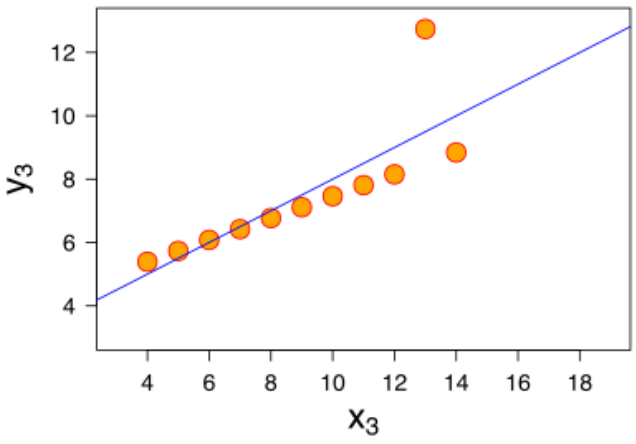
où x_i et y_i sont les valeurs des variables X et Y pour chacune des observations d'un jeu de données ($i = 1, \ldots, n$), et $\bar{x}$, $\bar{y}$, s_x et s_y sont les moyennes et écarts-types échantillonnaux des x_i et y_i respectivement.

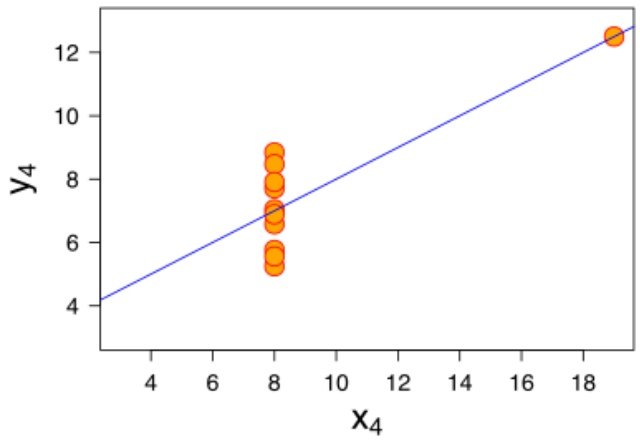
Le coefficient de corrélation r peut prendre des valeurs entre -1 et 1. Plus la valeur de r est proche de 1 en valeur absolue, plus la relation **linéaire** entre les variables est forte.

Quelques exemples

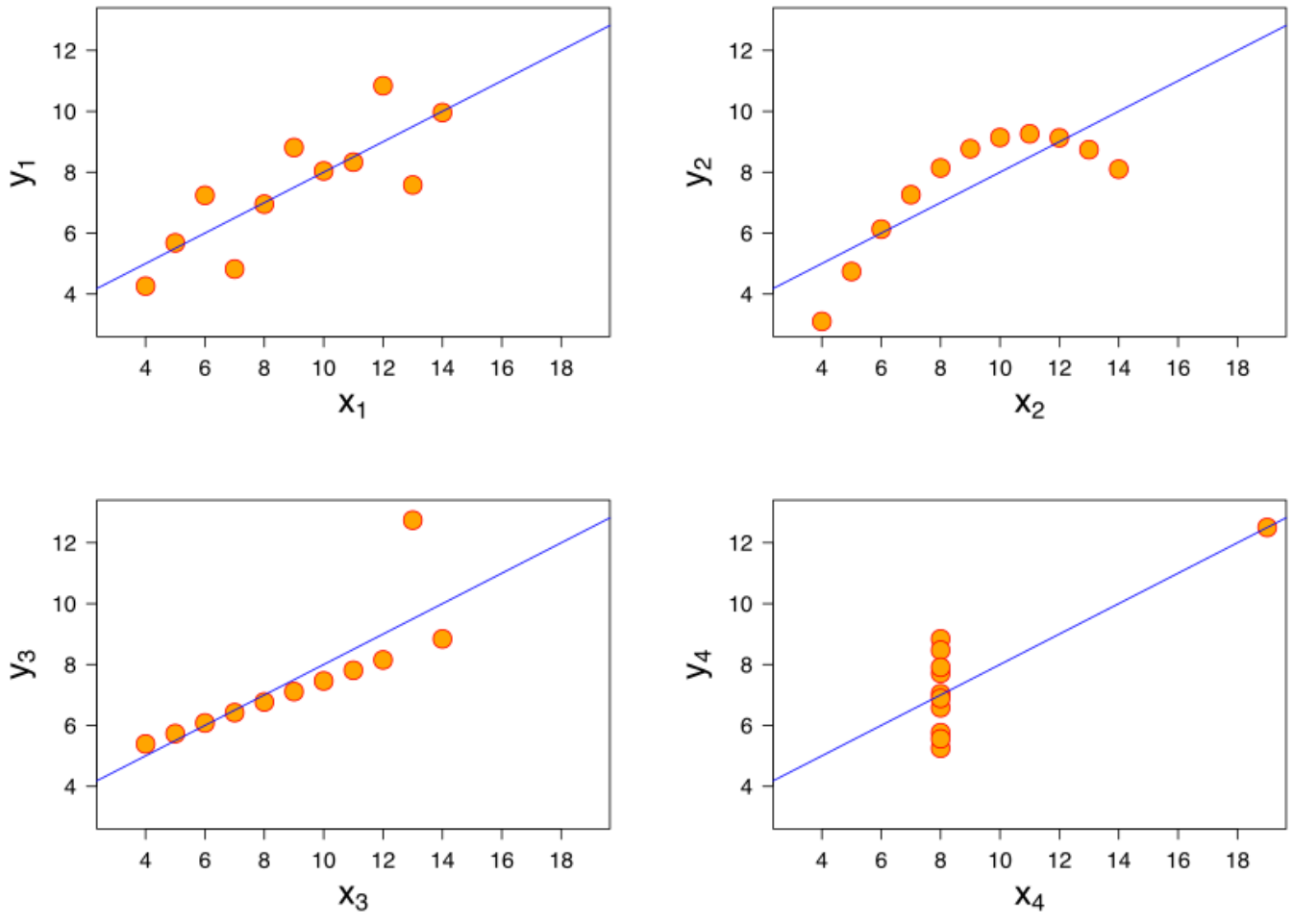








Quelques exemples

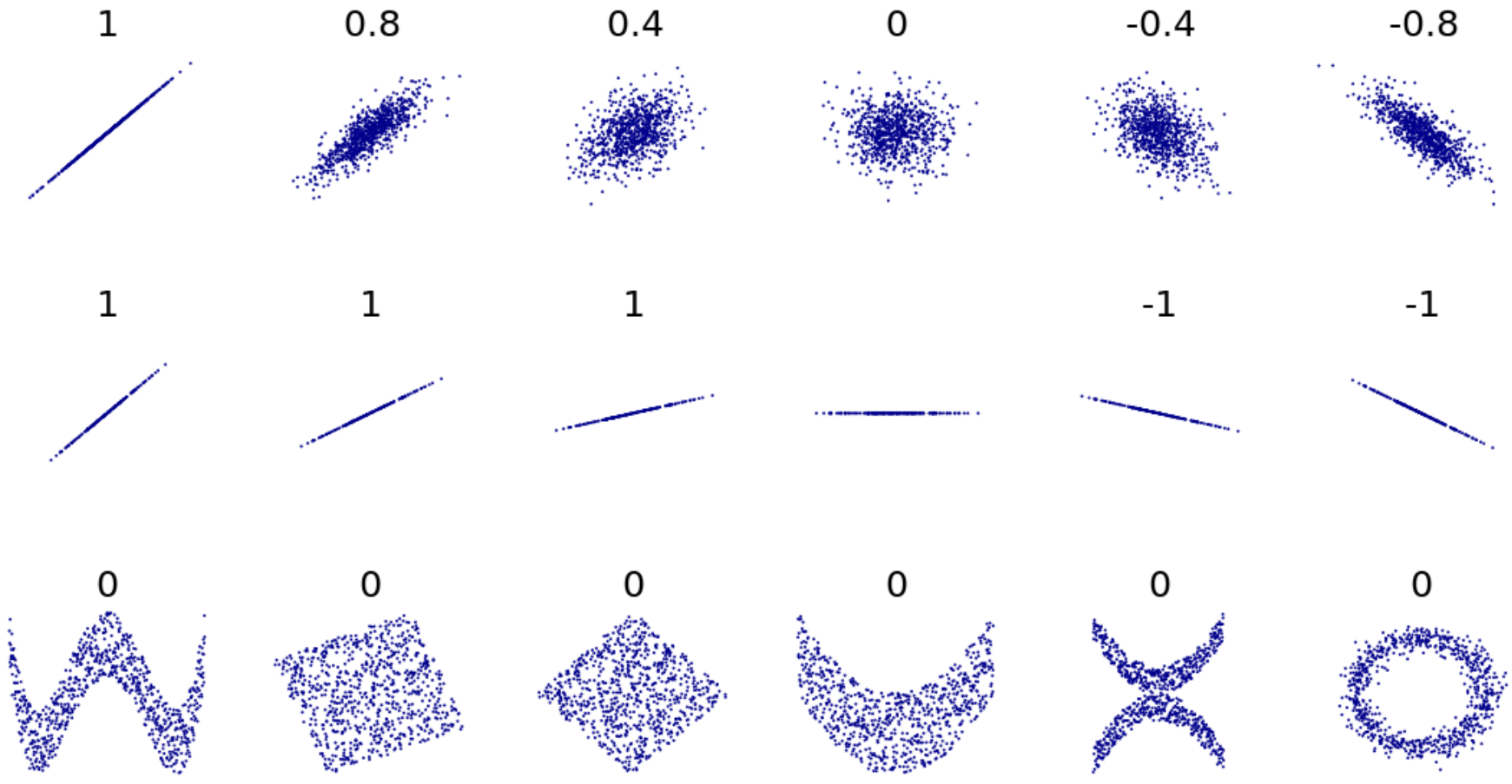


La corrélation est de 0.816 dans les quatre cas !
(voir Anscombe's quartet sur Wikipedia)

Impact d'une transformation linéaire

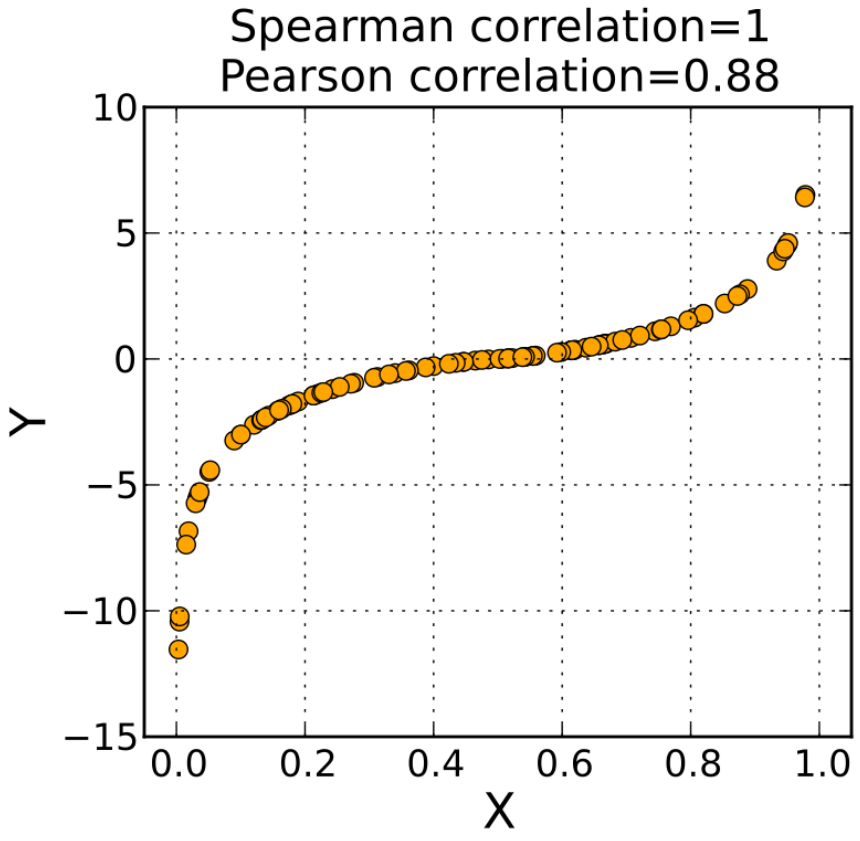
- Le coefficient de corrélation de Pearson reste inchangé lors de l'addition d'une constante, positive ou négative, à toutes les valeurs d'une variable ou même des deux variables.
- De même, la multiplication des valeurs par une constante positive n'affecte pas le coefficient.
- Par contre, si une seule des deux variables est multipliée par une constante négative, le coefficient changera de signe.

Linéarité du coefficient de corrélation de Pearson



(source : Pearson correlation coefficient sur Wikipedia)

Coefficient de corrélation de Spearman



Mesure si deux variables ont tendance à augmenter et diminuer simultanément, sans que le lien entre les deux variables ne soit nécessairement linéaire. On le calcule simplement en utilisant la formule de Pearson, mais en remplaçant les observations par leur rang.

Et il existe d'autres mesures de corrélation, notamment pour des variables catégoriques.

Transitivité de la corrélation

Exemple : Paradoxes in Film Ratings

https://www.tandfonline.com/doi/full/10.1080/10691898.2006.11910579

Résultat théorique :

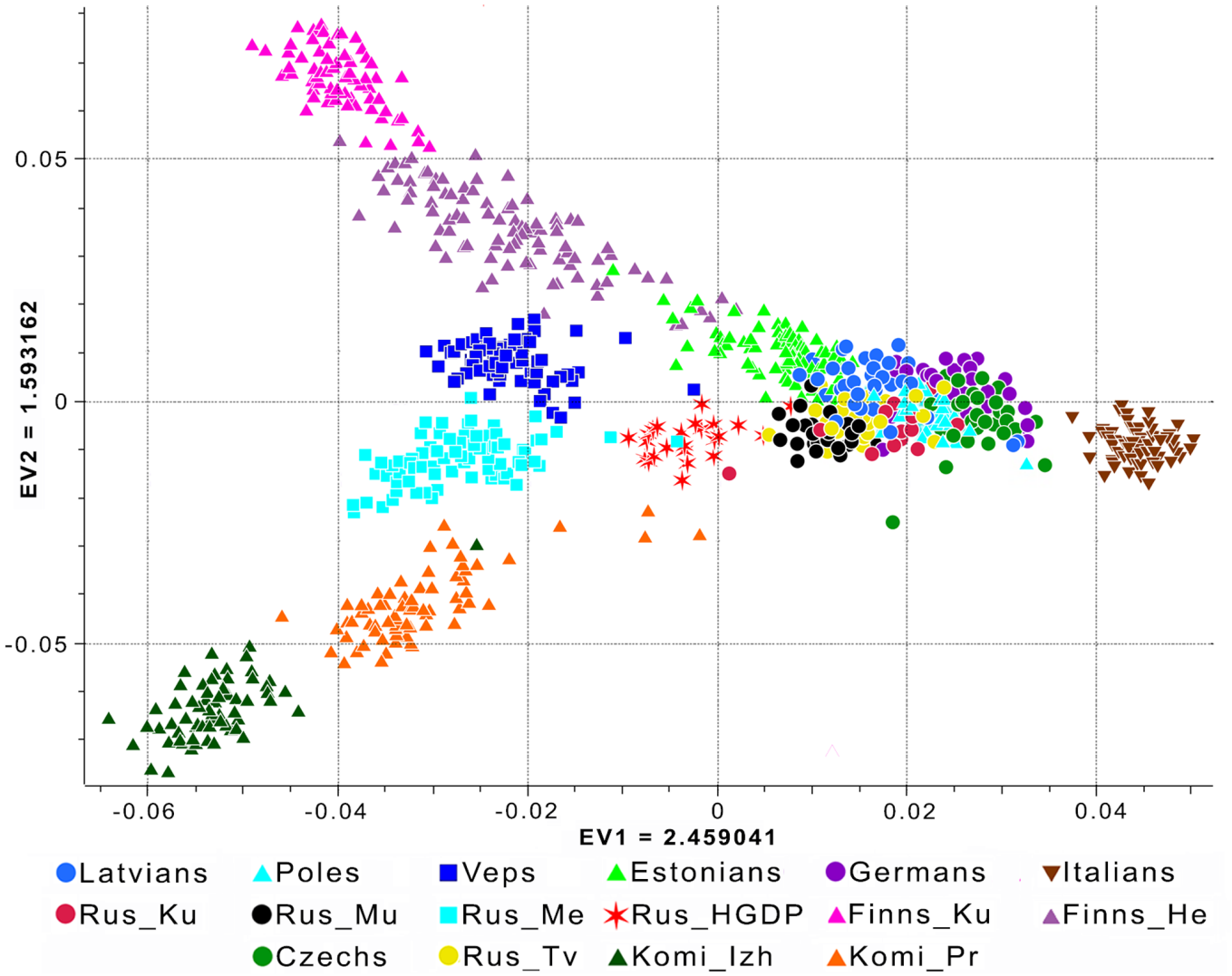
Soit X,Y,Z des variables aléatoires telles que X et Y sont corrélées positivement avec coefficient de corrélation ρ_{XY} et Y et Z sont corrélées positivement avec coefficient de corrélation ρ_{XY}. Si $\rho_{XY}^2 + \rho_{YZ}^2 > 1$ alors les variables X et Z sont également corrélées positivement.

Voir : Langford, Eric, Neil Schwertman, and Margaret Owens. *Is the Property of Being Positively Correlated Transitive ?* The American Statistician 55, no. 4 (2001) : 322–25.

http://www.jstor.org/stable/2685695.

Analyse en composantes principales

Motivation



Jeu de données : 166 000 SNPs mesurés sur des dizaines d'individus de différentes origines.

Deux exemples en science politique

- PCA and Brexit
  https://lennybronner.com/post/2019/04/13/pca-brexit.html

- Exploring the efficiency of Italian social cooperatives by descriptive and principal component analysis
  https://link.springer.com/content/pdf/10.1007/s11628-011-0131-9.pdf

ACP en bref

Une façon d'extraire la structure d'un jeu de données pour en réduire la dimensionnalité.

Pourquoi réduire la dimensionnalité ?

- Visualiser les données
- Identifier des sous-groupes dans les données
- Compresser des images, vidéos
- Simplifier les analyses
  - pour simplifier l'interprétation
  - et/ou en réponse au fléau de la dimension

Quelques détails

Données
Une matrice avec n observations, chacune étant un vecteur de p variables.

Objectif
Obtenir une représentation des données dans un espace plus restreint en conservant la plus grande quantité d'information possible.

Origine de la méthode
Harold Hotelling (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520.

Plus précisément

Espace restreint : on considère les combinaisons linéaires des variables mesurées.

Plus précisément

Espace restreint : on considère les combinaisons linéaires des variables mesurées.

Information conservée : on tente de maximiser la variabilité des données dans le nouvel espace.

Description mathématiques

Il y a plusieurs façons d'écrire l'ACP mathématiquement. On ira ici pour une présentation plus simple. On considère l'extraction des composantes l'une après l'autre.

Première composante principale

Soit un jeu de données

$$\mathbf{X} = (X_1, \ldots, X_p)^\top$$

avec matrice de covariance $\Sigma = \mathrm{var}(\mathbf{X})$.

On veut une première composante principale

$$Y_1 = \alpha_1^\top \mathbf{X} = \sum_{i=1}^{p} \alpha_{1i} X_i,$$

qui maximise $\mathrm{var}(Y_1)$.

Première composante principale (suite)

Il s'agit d'un problème d'optimisation relativement simple.
Pour que $\mathbb{V}\text{ar}(Y_1)$ soit maximale, il faut prendre

(i) $\lambda = \lambda_1$, la plus grande valeur propre de Σ ;

(ii) α_1, le vecteur propre normé correspondant.

Deuxième composante principale

On poursuit un objectif double :

(i) conserver le maximum de variation présente dans $\mathbf{X}$;

(ii) simplifier la structure de dépendance, pour faciliter l'interprétation.

Deuxième composante principale (suite)

Étant donné Y_1, la deuxième composante principale

$$Y_2 = \alpha_2^\top \mathbf{X}$$

est définie telle que

(i) $\text{var}(Y_2) = \alpha_2^\top \Sigma \alpha_2$ est maximale ;

(ii) $\alpha_2^\top \alpha_2 = 1$

(iii) $\text{cov}(Y_1, Y_2) = 0$.

On peut montrer qu'il faut alors choisir le vecteur propre normé correspondant à la deuxième plus grande valeur propre de Σ

Procédant par maximisations successives, on conclut que

$$\begin{aligned} Y_k &= \lambda_k, \text{ la } k^{\text{e}} \text{ composante principale} \\ &= \alpha_k^\top \mathbf{X}, \end{aligned}$$

où α_k est le vecteur propre normé de Σ associé à λ_k.

Matrice des composantes principales

Pour définir simultanément et de façon plus compacte les composantes principales, on pose

$$\mathbf{Y} = \mathbf{A}^\top \mathbf{X},$$

où

$$\mathbf{A} = (\alpha_1, \ldots, \alpha_p) = \begin{pmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{pp} \end{pmatrix}.$$

La matrice $\mathbf{A}$ a pour colonnes les vecteurs propres de Σ.

Note :

$$\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}_p, \quad \mathbf{A}^\top = \mathbf{A}^{-1}.$$

Interprétation des composantes

En analysant les variables qui sont grandement corrélées avec chacune des composantes principales, on peut interpréter ces composantes.

Scores des observations

La formule

$$\mathbf{Y}_i = \mathbf{A}^\top \mathbf{X}_i$$

donne les coordonnées de l'observation $\mathbf{X}_i$ dans le nouveau système d'axes.

On appelle

$$Y_{ij} = \mathbf{a}_j^\top \mathbf{X}_i = \sum_{k=1}^{p} a_{jk} X_{ik}$$

le score de $\mathbf{X}_i$ sur l'axe principal j.

Remarque

Quel est l'effet de l'ACP sur la distance entre les points ?

Remarque

Quel est l'effet de l'ACP sur la distance entre les points ?

L'ACP préserve la distance entre les points !

$$
\begin{aligned}
||\mathbf{Y}_i - \mathbf{Y}_j||^2 &= (\mathbf{Y}_i - \mathbf{Y}_j)^\top (\mathbf{Y}_i - \mathbf{Y}_j) \\
&= \{\mathbf{A}^\top (\mathbf{X}_i - \mathbf{X}_j)\}^\top \mathbf{A}^\top (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)^\top \mathbf{A}\mathbf{A}^\top (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)^\top (\mathbf{X}_i - \mathbf{X}_j) \\
&= ||\mathbf{X}_i - \mathbf{X}_j||^2,
\end{aligned}
$$

puisque $\mathbf{A}^\top = \mathbf{A}^{-1}$.

Variation expliquée par chaque CP

La trace de Pillai

$$\text{trace}(\Sigma) = \text{trace}(\Lambda) = \sum_{i=1}^{p} \lambda_i,$$

est une mesure globale de variation.

Ainsi, la proportion de variation expliquée par Y_i est

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

Estimation de Σ

Dans la pratique, la matrice Σ est inconnue.

Cependant, elle peut être estimée par

$$\hat{\Sigma} = \frac{\mathbf{S}}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^\top$$

à partir d'un échantillon aléatoire $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Un exemple

Voir le début du labo R sur l'ACP.

Point crucial

On peut faire soit :

- l'ACP de la matrice des covariances ;
- l'ACP de la matrice des corrélations.

La seconde se fait à partir des variables standardisées.

Point crucial

On peut faire soit :

- l'ACP de la matrice des covariances ;
- l'ACP de la matrice des corrélations.

La seconde se fait à partir des variables standardisées.

Elle est recommandée, à moins que les variables soient de variances semblables, ou que la différence de variabilité contienne de l'information d'intérêt.

Choix du nombre de composantes

- Dépend de l'utilisation qu'on veut en faire.
- Pour la visualisation, toujours plus facile avec 2 ou 3.
- Je présente ici 3 règles parfois utilisées.
- Il existe aussi des règles plus avancées, notamment basées sur des méthodes de rééchantillonage.

1^{re} règle

Garder autant de composantes que nécessaire pour expliquer 80% de la variation.

Pourquoi 80% ? C'est purement arbitraire !

2^{e} règle : Kaiser (1960)

Si l'ACP est effectuée sur la matrice des corrélations,

$$\text{garder } Y_k \quad \Leftrightarrow \quad \ell_k \geq 1.$$

Note : l_k est le $k^{\text{i-ème}}$ vecteur propre de l'ACP de la matrice des corrélations empirique.

Source :

H. F. Kaiser (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.

2^{e} règle : Kaiser (1960)

Autre formulation

Peu importe la matrice sur laquelle l'ACP est effectuée,

$$\text{garder } Y_k \quad \Leftrightarrow \quad \ell_k \geq \bar{\ell},$$

où

$$\bar{\ell} = (\ell_1 + \cdots + \ell_p)/p.$$

On a $\bar{\ell} = 1$ pour une matrice des corrélations.

2^{e} règle : Kaiser (1960)

Opinion divergente

Jolliffe (1972) recommande plutôt

$$\text{garder } Y_k \quad \Leftrightarrow \quad \ell_k \geq 0.7.$$

Source :

I. T. Jolliffe (1972). Discarding variables in a principal component analysis I : Artificial data. *Applied Statistics*, 21, 160–173.

3^{e} règle : Cattell (1966)

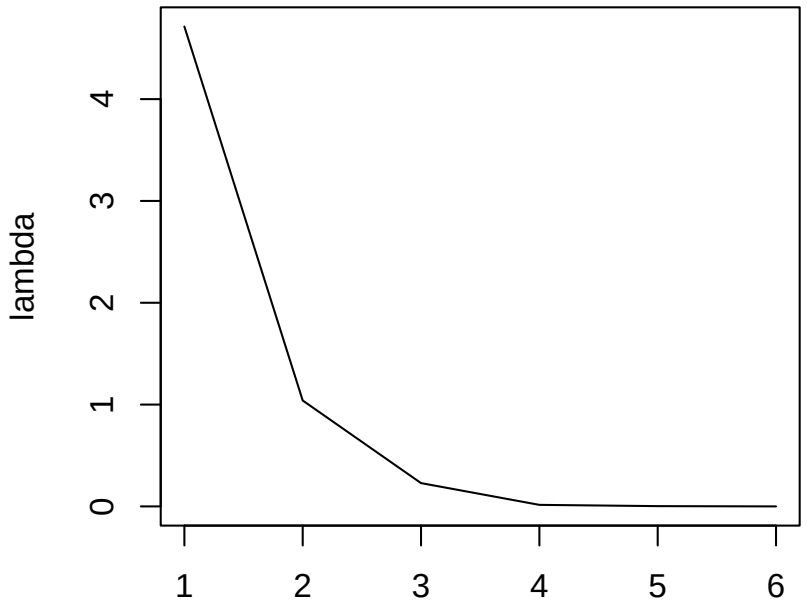
Dans le graphe des paires (k, ℓ_k),

garder les ℓ_k précédant le “pied de l’éboulis.”

Source :

R. B. Cattell (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.

Exemple :



Extensions de l'ACP classique

- ACP avec noyaux
  Pour permettre de considérer autre chose que des combinaisons linéaires des variables.
- ACP parcimonieuse
  Offre une façon de limiter le nombre de variables incluses dans une composante principale.
- ACP avec données manquantes
  Plusieurs approches, avec ou sans imputation des données manquantes.
  Dans tous les cas il faut faire très attention aux hypothèses faites par ces méthodes.

ACP en vidéo

StatQuest

Concepts de base en apprentissage supervisé

L'apprentissage supervisé

Il s'agit en fait de faire la prévision d'une ou de plusieurs variables à l'aides d'autres variables.

On dit que l'apprentissage est supervisé quand on travaille à faire cette prédiction à partir d'un jeu de données pour lequel on connait toutes les variables : celle à prédire (variable réponse) et celles à utiliser pour le faire.

Termes habituellement utilisés en apprentissage statistique :

Régression - si la variable réponse est continue.

Classification - si la variable réponse est catégorique.
(même si la régression logistique convient tout à fait ici !)

Les grandes questions

- Comment obtenir un modèle (algorithme) qui permet de bien prédire la variable d'intérêt ?
  Défis : ne pas se limiter à des relations linéaires, inclure des interactions entre des variables, avoir un modèle qui s'interprète bien, etc.

Les grandes questions

- Comment obtenir un modèle (algorithme) qui permet de bien prédire la variable d'intérêt ?
  Défis : ne pas se limiter à des relations linéaires, inclure des interactions entre des variables, avoir un modèle qui s'interprète bien, etc.

- Comment savoir quel est le meilleur modèle à utiliser pour faire des prévisions ?
  Et ce pour des nouvelles données, pas simplement celles utilisées pour faire le modèle !

Un peu de formalisme

Considérons une variable réponse numérique.

Soit Y la variable que l'on souhaite prédire, et X un ensemble de prédicteurs.

On fait l'hypothèse très générale que

$$Y = f(X) + \varepsilon$$

où $E(\varepsilon) = 0$ et $Var(\varepsilon) = \sigma^2$, une constante.

On dénote par $\hat{f}(x)$ notre prévision pour la valeur de $f(x)$.

Erreur de prévision

On peut prouver que si $Y = f(X) + \varepsilon$ où $E(\varepsilon) = 0$ et $Var(\varepsilon) = \sigma^2$, une constante, l'erreur quadratique attendue pour la prévision en x_0 s'écrit :

$$E(Y_0 - \hat{f}(x_0)^2) = \underbrace{[Biais(\hat{f}(x_0))]^2 + Var(\hat{f}(x_0))}_{\text{Erreur réductible}} + \underbrace{\sigma^2}_{\text{Erreur irréductible}}$$

L'erreur réductible peut être réduite par le choix de $\hat{f}(x_0)$ et en augmentant le nombre d'observations pour l'estimation.

L'erreur irréductible sera présente même si $\hat{f} = f$; elle est dû à l'aspect aléatoire de Y.

Compromis biais-variance

$$E(Y_0 - \hat{f}(x_0)^2) = \underbrace{[Biais(\hat{f}(x_0))]^2 + Var(\hat{f}(x_0))}_{\text{Erreur réductible}} + \underbrace{\sigma^2}_{\text{Erreur irréductible}}$$

L'erreur réductible dépend du biais et de la variance de $\hat{f}(x_0)$ en tant qu'estimateur de $f(x_0)$.

Dans le choix du $\hat{f}$, il y a généralement un compromis entre le biais et la variance de cet estimateur. C'est-à-dire qu'un estimateur avec un petit biais tend à avoir une grande variance et vice-versa.

L'idée est de choisir l'estimateur approprié pour balancer le biais et la variance de façon à ce que l'EQM (erreur quadratique moyenne, soit la partie d'erreur réductible) soit minimale.

Mesurer l'erreur de prévision

On peut évaluer la qualité de notre modèle de prévision en regardant s'il prédit bien pour les observations utilisées pour l'estimer, en utilisant

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{ou} \quad \frac{1}{n}\sum_{i=1}^{n} I[h(x_i) \neq y_i].$$

Ici, $\hat{y}_i$ dénote la valeur prédite pour la variable continue y pour l'observation i, et $h(x_i)$ dénote la valeur prédite pour une variable catégorique y pour l'observation i.

Mais, on a alors de fortes chances de sous-estimer les vraies erreurs, car $\hat{f}$ et h ont été estimés à l'aide des données, et s'adaptent donc à celles-ci.

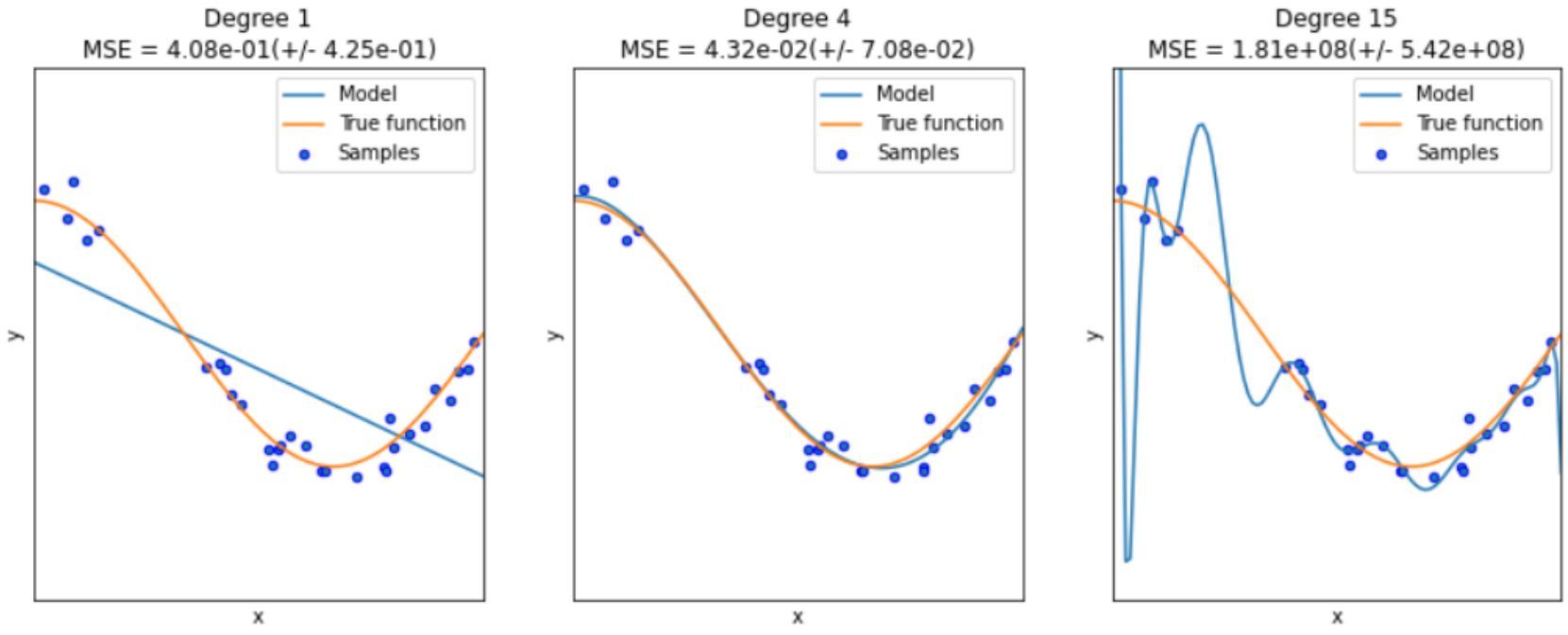
Sur-ajustement (*overfitting*)

Un modèle/classifieur plus complexe/flexible pourra mieux s'ajuster aux données observées, et mènera à une plus petite EQM, si calculée sur le jeu de données originales.

Mais si le modèle/classifieur s'adapte trop aux données observées, il risque de ne pas être aussi bon si on l'applique sur de nouvelles données.

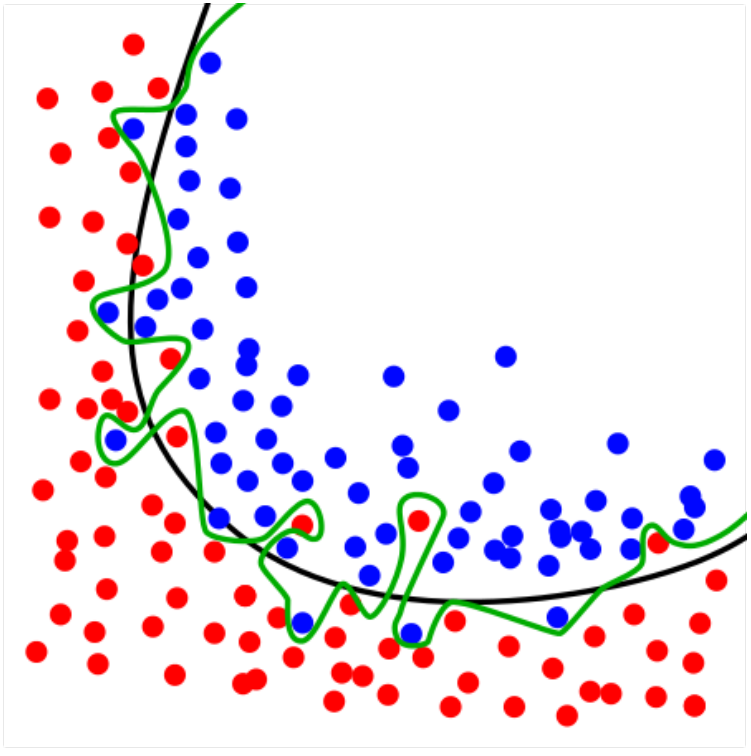
Il y a sur-ajustement si on choisit à l'aide du jeu de données d'entrainement un modèle/classifieur trop complexe/flexible, qui s'ajuste à la partie aléatoire des données de sorte que l'EQM sur un jeu de données de validation est plus grande qu'elle ne l'aurait été avec un modèle/classifieur moins complexe/flexible.

Illustration - régression



Source : `https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning`

Illustration - classification



Source : Wikipedia, overfitting

Solutions au sur-ajustement

- Utiliser un jeu de données de validation
- Utiliser la validation croisée

Utiliser un jeu de données de validation

On divise le jeu de données initial en deux parties :

1. Jeu d'apprentissage
   On estime $\hat{f}$ ou h à l'aide de ce jeu de données.
2. Jeu de validation
   On estime l'erreur quadratique moyenne espérée ou le risque du classifieur sur ce jeu de données.

Utiliser un jeu de données de validation

On divise le jeu de données initial en deux parties :

1. Jeu d'apprentissage
   On estime $\hat{f}$ ou h à l'aide de ce jeu de données.
2. Jeu de validation
   On estime l'erreur quadratique moyenne espérée ou le risque du classifieur sur ce jeu de données.

Attention :

- La qualité de l'estimation de l'erreur dépendra de la taille du jeu de validation.
- La qualité de $\hat{f}$ ou h dépendra de la taille du jeu d'apprentissage.
- Il faut faire un compromis entre ces deux aspects quand on divise le jeu de données original. Requiert en général un grand jeu de données original.

Validation croisée

Extension de l'idée du jeu de validation, mais chacune des observations sera tour à tour dans le jeu d'apprentissage et dans le jeu de validation.

1. Diviser les données en k groupes.
2. Ajuster un modèle sur $k-1$ des groupes.
   L'utiliser pour prédire les valeurs dans le dernier échantillon.
   Calculer l'erreur quadratique moyenne ou le risque de classification.
3. Répéter l'étape 2 k fois en utilisant un groupe différent comme jeu de validation à chaque fois.
4. Calculer la moyenne des k erreurs obtenues.

Choix du nombre de groupes k (1/2)

Si k est grand (disons $k = n$) :

- L'estimateur de l'erreur de prévision espérée est approximativement sans biais pour la vraie erreur de prévision espérée.
- Mais, sa variance peut être grande car les $k = n$ jeux d'apprentissage sont très corrélés.

  Note : il faut penser à la variabilité en terme de différents jeux de données observés, pas simplement en terme de répétition de la validation croisée.

Choix du nombre de groupes k (2/2)

Si k est petit (disons $k = 5$) :

- La variance de l'estimateur de l'erreur de prévision espérée est diminuée car les jeux d'apprentissage sont moins corrélés.
- L'estimateur sera toutefois biaisé, si la performance de notre modèle/classifieur varie beaucoup avec la taille d'échantillon, car on apprend sur un jeu de données de taille inférieure à n.

En pratique :
On choisit en général $k = 5$ ou $k = 10$.
Le choix de k peut dépendre du temps de calcul requis.

Laboratoire

ACP et régression

Méthode des plus proches voisins

Général

On utilise les observations les plus proches de celle d'intérêt, en termes de variables explicatives, pour prédire la variable réponse.

Peut être utilisé pour la classification ou la régression.

Méthode très simple, mais qui donne parfois de bons résultats.

Régression par la méthode des k plus proches voisins :

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_k} y_i$$

où N_k est l'ensemble des k plus proches voisins de x_0.

Compromis biais-vairance

Le choix de k aura un impact important sur le biais et la variance de l'estimateur.

Si k augmente :

- $\hat{f}(x_0)$ dépend de données plus loin de x_0, donc le biais augmente.
- $\hat{f}(x_0)$ est la moyenne d'un plus grand nombre d'observations, donc la variance diminue.

Classification par la méthode des k plus proches voisins :

$$h(x_0) = y \sum_{x_i \in N_k} (y_i = y)$$

où N_k est l'ensemble des k plus proches voisins de x_0.

C'est à dire qu'on prédit la valeur de y la plus fréquente parmi les voisins de x_0. En cas d'égalité, on peut tirer au hasard parmi les deux valeurs les plus fréquentes.

Note

Cette méthode requiert de calculer les distances entre les observations du jeu de données. La distance euclidienne est souvent utilisée, mais d'autres distances peuvent être plus appropriées dans certains cas.

Avec R

Classification : fonction `knn` dans la librairie `class`.

Régresssion : fonction `knn.reg` dans la librairie `FNN`.

Vous trouverez facilement en ligne des tutoriels sur Datacamp
- Tutorial KNN
- Cours KNN

Arbres de régression / classification

Origine

CART = Classification And Regression Trees

Livre source :
Breiman, Leo ; Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and regression trees. Monterey, CA : Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

Classification : variable dépendante catégorique
Régression : cariable dépendante continue

Variables indépendantes catégoriques ou continues.

Produire l'arbre

1. Divisions binaires en fonction de la valeur d'une variable.

   Choix des variables et des limites pour la division en fonction d'un certain critère (ex. index de Gini)

2. Chaque feuille (noeud terminal) correspond à une région à l'intérieur de laquelle on prédit la même catégorie.

   Prévision dans une feuille :
   Classification : vote à majorité (i.e. classe la plus fréquente)
   Régression : moyenne des observations de la région

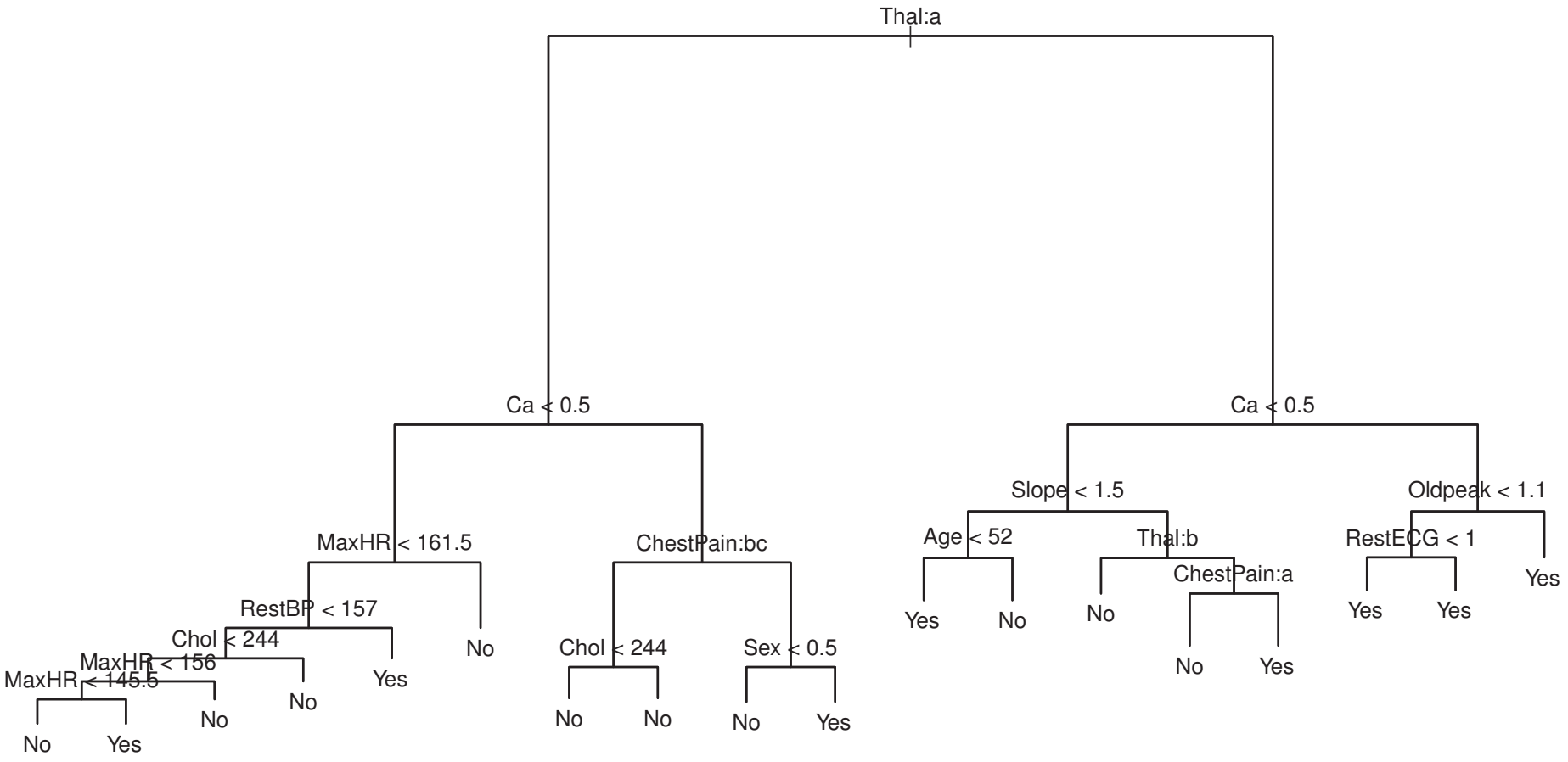
Exemple - données

303 patients atteints de douleur à la poitrine

Y : Indicateur d'un problème cardiaque (oui ou non)

13 variables indépendantes : âge, sexe,cholestérol, ...
(continues et catégoriques)

Exemple - arbre obtenu



L'algorithme

C'est un algorithme glouton, car on choisit la meilleure division à chaque niveau. On n'est pas certain d'obtenir l'arbre optimal.

À chaque étape, on utilise un critère pour choisir sur quelle variable diviser, et quelle valeur utiliser pour la division.

En R, disponible dans la librairie `rpart` et la librairie `tree`, entre autres.

Critère de coupure - Régression

Pour une variable X_j donnée et un point de coupure s donné, on obtiendra les deux régions suivantes :

$$R_1(j,s) = \{X | X_j < s\} \quad \text{et} \quad R_2(j,s) = \{X | X_j \geq s\}$$

Si X_j est une variable catégorique, on forme les régions en divisant en deux groupes les différents niveaux de la variable.

On choisira j et s de façon à minimiser la somme des carrés des erreurs

$$\sum_{i:x_i \in \mathbb{R}_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in \mathbb{R}_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

où $\hat{y}_{R_k} = \sum_{i:x_i \mathbb{R}_k(j,s)} y_i$ pour $k = 1, 2$.

Critères de coupure - Classification

Soit $\hat{p}_{mk}$ la proportion d'observations dans la région m qui font partie de la classe k.

On choisit habituellement les divisions pour minimiser un des trois critères suivants :

1. Taux d'erreur de classification :

$$E_m = 1 - \max_k(\hat{p}_{mk})$$

2. Index de Gini :

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

3. Entropie croisée :

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk})$$

Nombre de divisions

Quand arrêter de faire des divisions ?
Habituellement basé sur le critère utilisé pour choisir les divisions. On peut aussi limiter la taille des feuilles (ex. au moins 5 observations par feuille).

Importance du nombre de divisions :

Pas assez de divisions : notre modèle n'est pas assez flexible ; le biais sera donc important.

Trop de divisions : chaque prévision ne dépend que de peu d'observations ; la variance sera donc importante.

Élagage

Soit T_0 l'arbre obtenu par l'algorithme précédent.
On l'élague en enlevant des branches (divisions) dans le but d'obtenir un meilleur sous-arbre.

On utilise souvent l'élagage coût-complexité (*Cost-complexity pruning*). Étant donné un paramètre α, on choisit le sous-arbre $T \subset T_0$ tel que le critère suivant est minimal :

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

où $|T|$ dénote le nombre de feuilles dans l'arbre T.

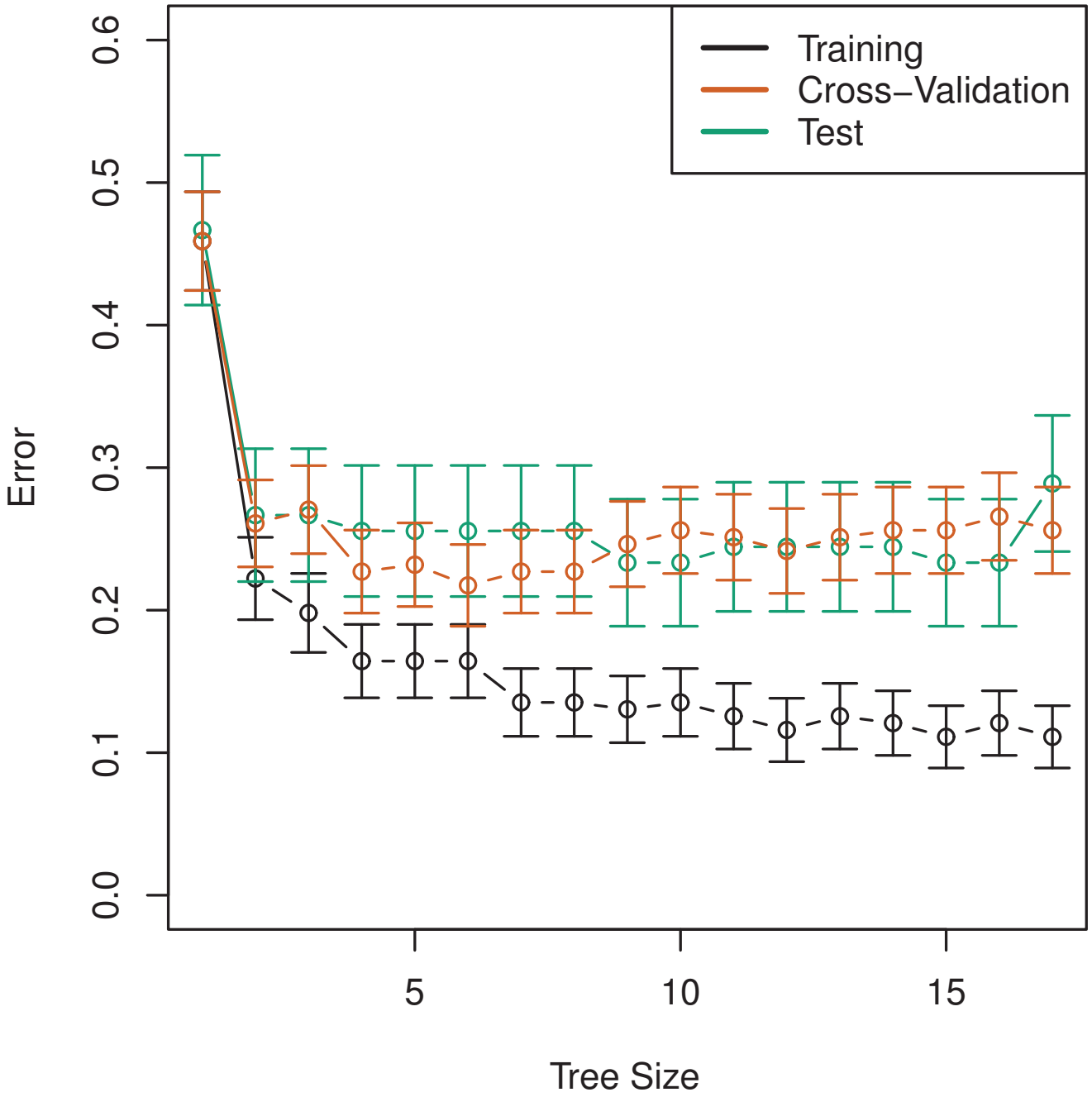
Élagage - choix de α

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

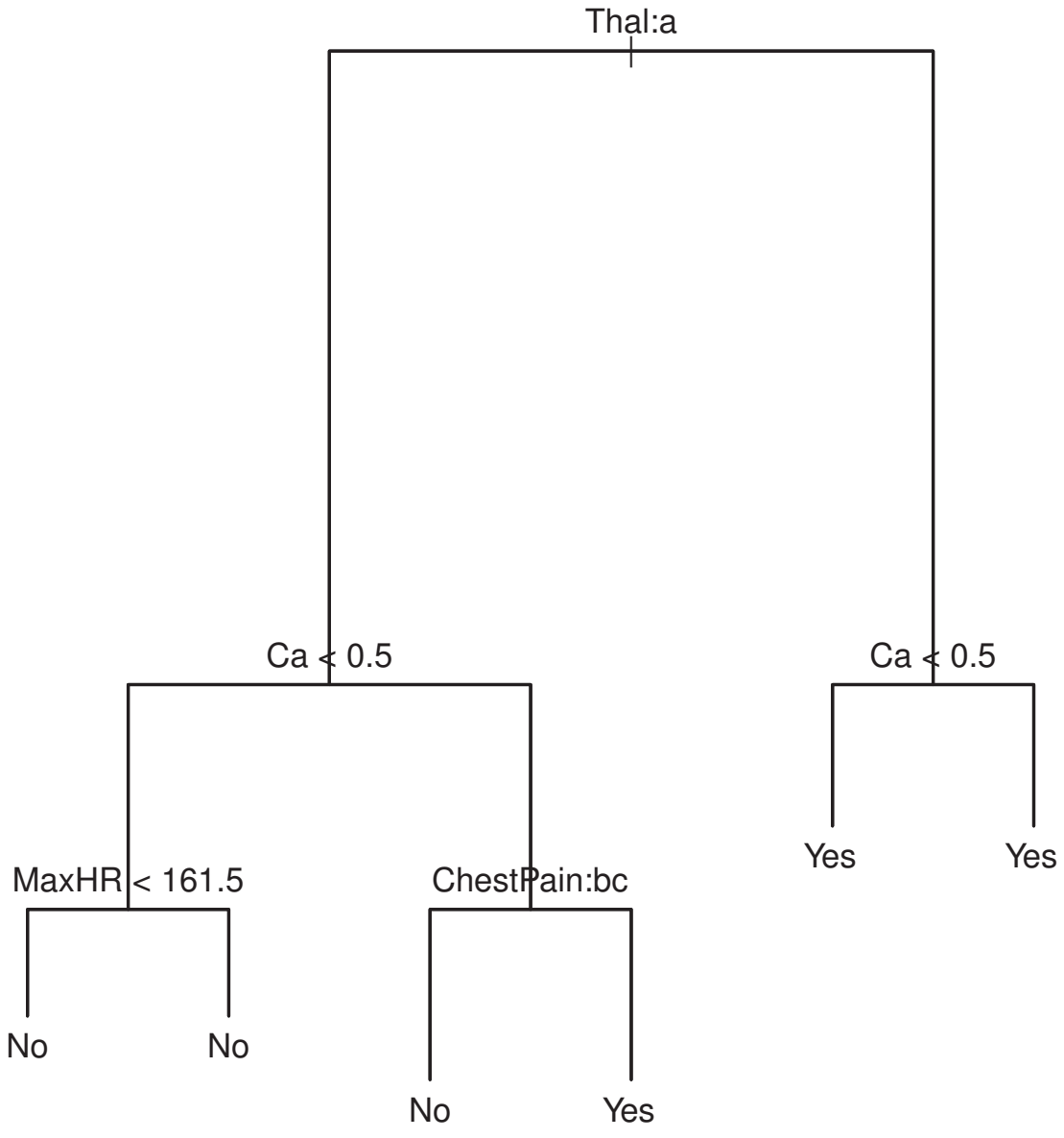
Impact de α :
Si $\alpha = 0$, on ne limite pas la taille de l'arbre.
Plus α augmente plus on favorise les petits arbres.

Choix de α : Généralement pas validation croisée.

Illustration (avec l'exemple de tout à l'heure)



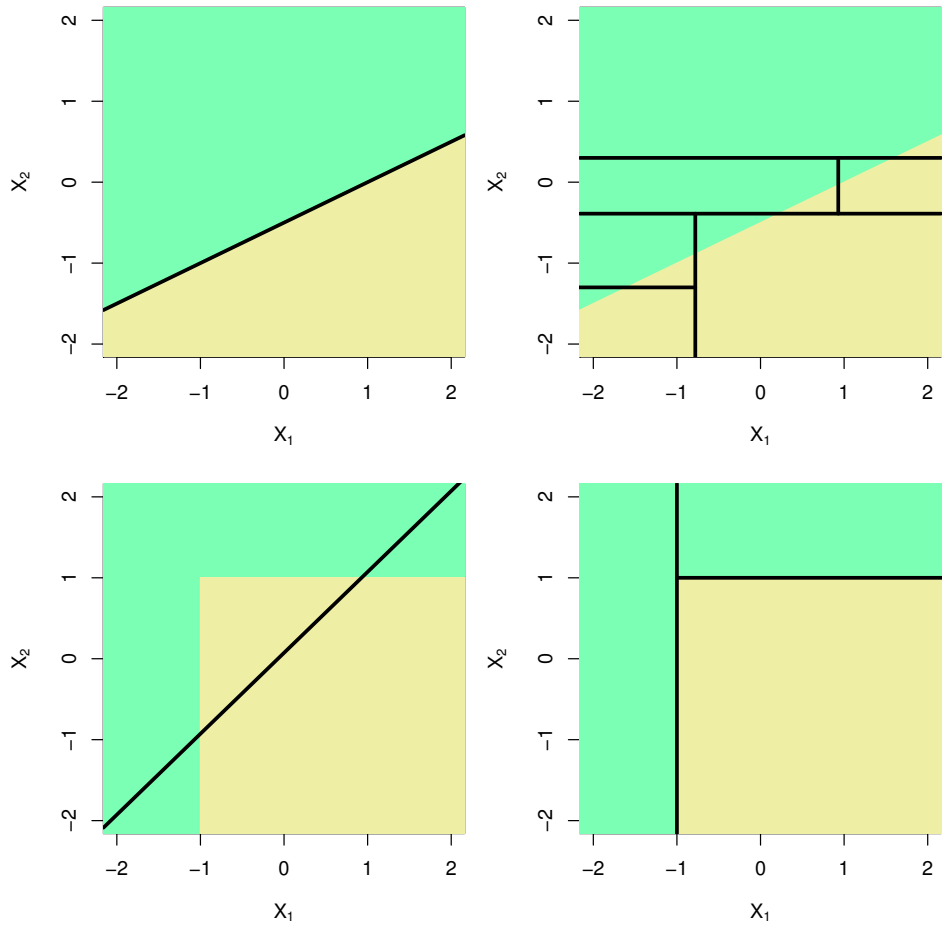
Exemple - arbre optimal post élagage



Avantages des arbres de classification

- Méthode non paramétrique :
  aucune hypothèse *a priori* sur la distribution des données
- Résistante aux données atypiques
- Facile à interpréter
- Modélise indirectement des interactions
- Sélection de variable implicite
- Modèle non linéaire

Non-linéarité du modèle



Désavantages des arbres de classification

- Besoin d'un grand nombre de données.
- Peuvent être instable.

Améliorations possibles :

- Bagging
- Forêts aléatoires
- Boosting

Autres méthodes : Très variées !

Il existe une multitude méthodes pour l'apprentissage supervisé, que ce soit pour la régression ou la classification.

L'important dans tous les cas, c'est de s'assurer de choisir un modèle approprié en faisant attention au sur-ajustement.

Stratégie souvent utilisée :

1. Choisir les hyperparamètres d'un modèle par validation croisée (ex. nombre de variables à conserver, valeur de k pour k-nn, etc.).
2. Choisir entre les différents modèles à l'aide d'un jeu de données de validation.

Méthodes plus statistiques

- Régression linéaire, logistique, Poisson, multinomiale, polynomiale, etc.
- Régression pénalisée (régularisée), LASSSO
- Régression non-paramétrique
- Régression linéaire locale
- Régression avec splines
- *Generalized additive models*
- Etc.

Méthodes plus ML

- Classifieur bayesien naif
- Forêts aléatoires
- Machine à vecteur de supports (SVM)
- Réseaux de neurones
- Apprentissage profond (CNN, RNN, GAN, etc.)
- Etc.

Source

Certaines figures de cette présentation viennent du livre *An Introduction to Statistical Learning with Applications in R.*

Laboratoire

Laboratoire Arbre
Aller plus loin :
- Module Chap 5 DataCamp
- StatQuest

Classification non-supervisée (clustering)

Objectif et méthodes

On veut grouper des observations en un certain nombre de groupes homogènes, sans connaitre ce nombre de groupes à l'avance, ni l'appartenance des individus aux groupes.

Plusieurs méthodes :

- Classification hiérarchique
- Classification par regroupements
- Classification basée sur des modèles
- Classification basée sur une densité

Aujourd'hui : k-moyennes

Des exemples

- Bail, C. A. (2008). The configuration of symbolic boundaries against immigrants in Europe. American Sociological Review, 73(1), 37-59. ici
- Discovering diverse mechanisms of migration : The Mexico–US Stream 1970– 2000. Population and Development Review. ici

Ces deux liens viennent d'un plan de cours pour Machine Learning for Social Sciences par Jorge Cimentada, trouvé ici.

Méthode des k-moyennes

Idée générale

Soit un ensemble d'observations $x_1, \ldots, x_n$.
Soit $C_1, \ldots, C_K$ les ensembles des index des observations dans chacun des groupes.
Par ex. $C_3 = \{2,5\} \implies$ observations 2 et 5 dans le groupe 3

On veut une partition des observations, c'est-à-dire

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,2,\ldots,n\}$
- $C_k \cap C_{k'} = \emptyset$ pour tout $k \neq k'$

Idée générale

Soit un ensemble d'observations $x_1, \ldots, x_n$.
Soit $C_1, \ldots, C_K$ les ensembles des index des observations dans chacun des groupes.
Par ex. $C_3 = \{2,5\} \implies$ observations 2 et 5 dans le groupe 3

On veut une partition des observations, c'est-à-dire

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,2,\ldots,n\}$
- $C_k \cap C_{k'} = \emptyset$ pour tout $k \neq k'$

On va essayer de minimiser la variation à l'intérieur de chaque groupe.

Mesure de la variabilité intra-groupe :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Détails techniques

Mesure de la variabilité intra-groupe :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

On veut donc choisir les $C_1, \ldots, C_K$ qui minimisent

$$\sum_k \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Détails techniques

Mesure de la variabilité intra-groupe :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

On veut donc choisir les $C_1, \ldots, C_K$ qui minimisent

$$\sum_k \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

C'est un problème d'optimisation très difficile.
On peut toutefois approximer la solution assez facilement.

Un algorithme (celui de Lloyd)

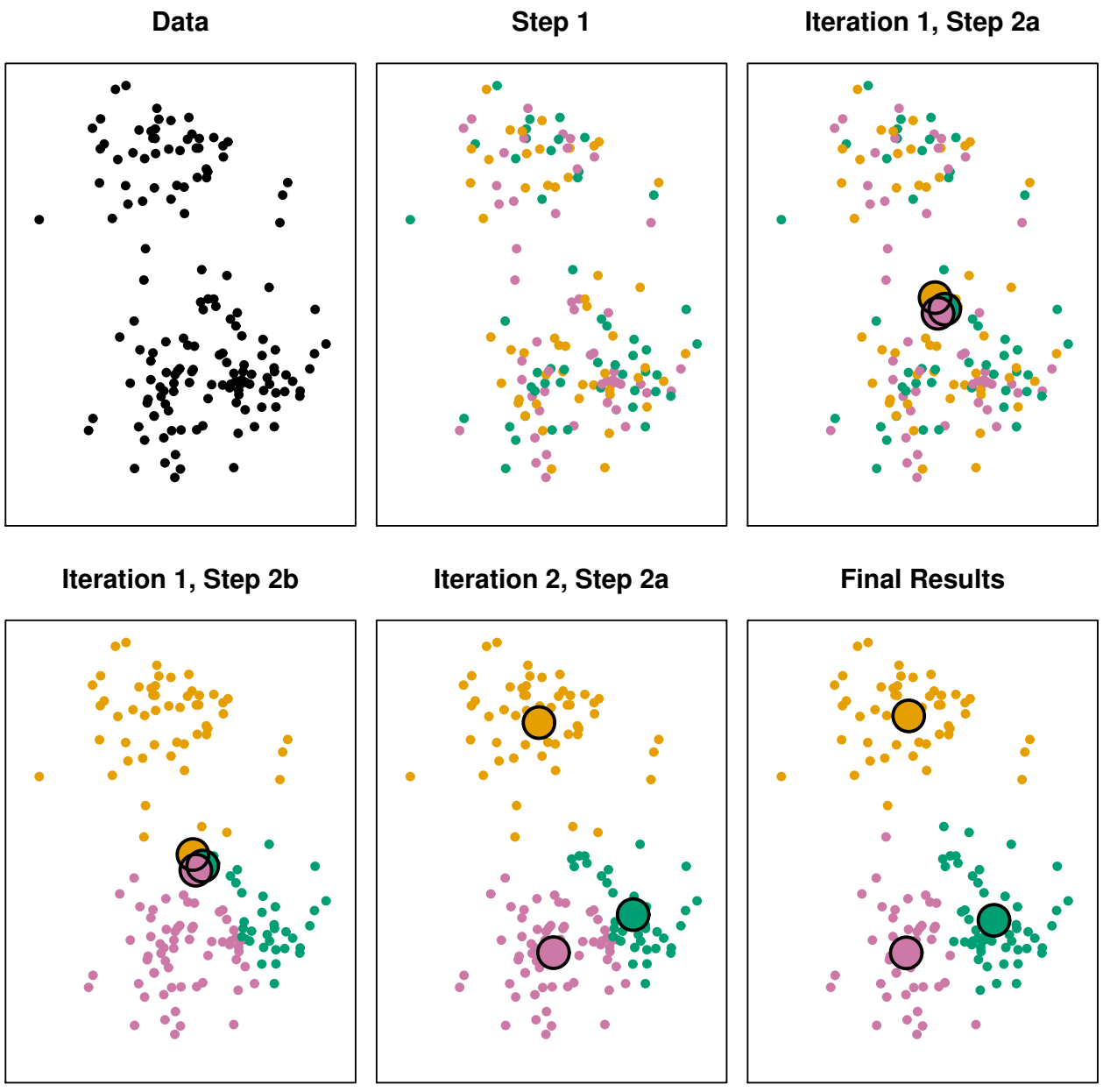
1. Assigner aléatoirement chaque observation à un des k groupes.
2. Calculer la moyenne des observations dans chacun des groupes.
3. Assigner chaque observation au groupe duquel elle est le plus proche.
4. Répéter les étapes 2 et 3 jusqu'à convergence.

Un algorithme (celui de Lloyd)

1. Assigner aléatoirement chaque observation à un des k groupes.
2. Calculer la moyenne des observations dans chacun des groupes.
3. Assigner chaque observation au groupe duquel elle est le plus proche.
4. Répéter les étapes 2 et 3 jusqu'à convergence.

Note : À l'étape 1, on peut aussi simplement choisir au hasard k observations pour être les centres des k groupes. On ira directement à l'étape 3 pour la première étape.

Exemple


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

En vidéo

StatQuest

Convergence

On peut prouver assez facilement que l'algorithme convergera nécessairement vers un minimum local.

Pour s'assurer que ce minimum local est un minimum global, on répète habituellement l'algorithme avec plusieurs valeurs de départ différentes. (par exemple avec l'option `nstart` dans la fonction `kmeans` de R)

Exemple

Choix du nombre de groupes

Il faut décider dès le départ du nombre de groupes désiré pour utiliser l'algorithme k-moyennes.

On fait habituellement la classification avec plusieurs valeurs différentes de k.

On doit ensuite choisir la valeur de k.
Il existe plusieurs méthodes pour ce faire.

Hypothèses implicites

En minimisant la somme des carrés des distances à l'intérieur des groupes, la méthode des k-moyennes fait implicitement les hypothèses suivantes :

- Les variables suivent des lois normales
- Les variances de toutes les variables sont similaires
- Les groupes sont de mêmes tailles

On peut l'appliquer dans d'autres cas, mais les résultats risquent de ne pas être aussi bons.
Voir ces quelques exemples :
`https://www.r-bloggers.com/2015/01/k-means-clustering-is-not-a-free-lunch/`.

Autres caractéristiques

- Sensible aux valeurs extrêmes
- Uniquement pour des variables continues
- Dépend de la standardisation des variables

Variante : K-modes

Partitionne les observations en k groupes pour que la somme des distances entre les observations et le mode de leur groupe soit minimisée.

Utile pour des données catégoriques.

R : fonction `kmodes` de la librairie `klaR`

Variante : K-medoids

On ajoute la contrainte que le centre d'un groupe soit une des observations de ce groupe.

Utile par exemple si je veux regrouper des personnes autour d'un centre qui est aussi une personne.

Peut se faire à partir d'une matrice de dissimilarités.

R : fonction `pam` de la librairie `cluster`.

Choix de k

Aperçu

Il existe une panoplie de mesures pour

- évaluer la qualité d'une classification non supervisée
- comparer deux classifications entre elles

On peut évidemment utiliser ces méthodes pour choisir le nombre de groupes.

La librairie `NbClust` implémente 30 mesures pour choisir le nombre de classes. Elle peut être utilisée avec plusieurs méthodes de classification, mais seulement celles implémentées dans la librairie.

D'autres librairies implémentent certains autres critères, par exemple

- `cclust`
- `clusterSim`
- `clv`
- `clValid`

Ultimement, ça devient presque plus un art qu'une science...

Mesures de la qualité d'une classification

On distingue en général deux types de mesures :

- Mesures internes
  e.g. Index de Dunn, Connectivité, Largeur de silhouette,...
- Mesures de stabilité
  Stabilité si on enlève une des variables (voir clValid), si on modifie le jeu de données utilisé,...

Index de Dunn

On maximise l'index suivant :

$$D = \frac{min_{i \neq j}\, d(C_i, C_j)}{max_k\, d'(C_k)}$$

Le numérateur donne la distance minimale entre deux groupes. Le dénominateur donne la distance intra-groupe maximale (diamètre) de tous les groupes.

L'index de Dunn cherche donc à créer des groupes denses et bien séparés.

Connectivité

On veut minimiser l'index suivant :

$$Conn(C) = \sum_{i=1}^{n} \sum_{j=1}^{L} X_{i,nn_i(j)}$$

où $nn_i(j)$ donne l'index du j-ième plus proche voisin de i et

$$X_{i,nn_i(j)} = \begin{cases} 1/j, & \text{si } i \text{ et } nn_i(j) \text{ ne sont pas dans le même groupe} \\ 0, & \text{sinon} \end{cases}$$

La connectivité mesure donc si les points proches sont dans le même groupe.

La valeur de L doit être choisie. (10 par défaut dans la fonction `clValid`)

Largeur de silhouette

La silhouette de l'observation i mesure la confiance dans le choix du groupe pour l'observation i :

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

où
a_i est la distance moyenne entre l'observation i et les autres observations de son groupe

b_i est la distance moyenne entre l'observation i et les observations du groupe le plus proche de i (parmi ceux auxquels il n'appartient pas)

On souhaite maximiser la silhouette moyenne des observations.
On peut aussi regarder la distribution des silhouettes à l'intérieur des groupes pour trouver des groupes clairement séparés.