Apprentissage supervisé

Master MLDS - 2015/2016
Université Paris Descartes
Lazhar.labiod@parisdescartes.fr

Déroulement du cours

- 6 Séances de cours/ TP (3h/Séance)
- Présentation de plusieurs modèles/algorithmes pour l'apprentissage supervisé
- Application de ces algorithmes sur différents jeux de données réelles
- TP sous R
- Evaluation
  - Projet sur données réelles

Objectif du cours

- Initier les étudiants aux méthodes d'apprentissage supervisé et leurs mises en application sous R.
- à l'issue du cours, l'étudiant doit
  - Connaître le principe de base et les limites des différentes méthodes
  - Savoir les mettre en œuvre
  - Savoir interpréter les résultats
  - Savoir choisir la méthode la plus adaptée à l'objectif de l'étude et à la nature des données

Quelques Références

- **Larry Wasserman (2004).** All of statistics - A concise course in Statistical Inference Springer Texts in Statistics. http://www.stat.cmu.edu/~larry/all-of-statistics/
- **Trevor Hastie, Robert Tibshirani et Jerome Friedman (2009).** The Elements of Statistical Learning Springer Series in Statistics. http://statweb.stanford.edu/~tibs/ElemStatLearn/
- **Christopher M. Bishop (2009).** Pattern recognition and machine learning. Springer. http://research.microsoft.com/en-us/um/people/cmbishop/prml/
- **Breiman L., Friedman J., Olshen R., Stone C.**, *Classification and Regression Trees*, Chapman & Hall, 1984, 1993, 1998.
- **Tufféry S.**, *Data Mining et statistique décisionnelle*, Technip, 2005.
- **Hardle W., Simar L.**, *Applied Multivariate Statistical Analysis*, Springer, 2003.
- **Saporta G.**, *Probabilit és, ADD et statistique*, Technip, 1990.

Ressources

- Cours
  - Machine Learning, Andrew Ng (Stanford University) : https://www.coursera.org/course/ml
  - WikiStat : http://wikistat.fr/
- Logiciels
  - R + RStudio : http://www.rstudio.com/
  - Python + scikit-learn : http://scikit-learn.org/
- Jeux de donnees
  - UC Irvine Machine Learning Repository : http://archive.ics.uci.edu/ml/
- Challenges industriels
  - Kaggle : https://www.kaggle.com/
  - Datascience.net : https://datascience.net/
- Conferences
  - ICML : http://icml.cc/
  - NIPS : http://nips.cc/

Plan du cours

- **Introduction**
  - Définitions
  - Données
  - Démarche
  - Méthodes
  - Evaluation et comparaison de méthodes
  - Exemples sous R
- **Régression linéaire simple**
- **Analyse Discriminante (AFD, LDA)**
- **KNN**
- **Régression Logistique**
- **Arbre de décision (CART)**
- **Séparateurs à Vastes Marges (SVM)**
- **Projet**

Généralités - Rappels

Qu'est-ce que l'apprentissage automatique

- **Arthur Samuel (1959)** : "Domaine d‘ étude qui donne aux ordinateurs la capacité d‘apprendre sans être explicitement programmés".

- **Tom M. Mitchell (1997):** "On dit qu'un programme informatique apprend de l‘ expérience E par rapport a un type de tâches T et une mesure de performance P, si sa performance aux tâches de T, telle que mesurée par P, s‘ améliore avec l‘ expérience E".
- En quelques mots
  - 1. observations d’un phénomène
  - 2. construction d’un modèle de ce phénomène
  - 3. prévisions et analyse du phénomène grâce au modèle
  - le tout automatiquement (sans intervention humaine)

Notations

- x : variables explicatives (espace associé X)
- y : variable à expliquer (espace associé Y)
- un modèle f : une fonction de X dans Y
- f(x) est la prédiction/prévision du modèle pour l'entrée x
- l'ensemble des données à partir desquelles on construit le modèle est l'ensemble d'apprentissage
- collisions Français et Anglais :
  - Français — Anglais
  - Classification — Clustering
  - Classement — Classification ou ranking
  - Discrimination — Classification

Apprentissage supervisé

- **Objectifs** : A partir d'un ensemble d'observation X= {*x*1,*x*2,.., *xn*} et de mesures Y= {*yi*}, on cherche à estimer les dépendances entre l'ensemble *X* et *Y*.

- **Exemple** : on cherche à estimer les liens entre les habitudes alimentaires et le risque d'infarctus. *xi* est un patient décrit par p caractéristiques concernant son régime et *yi* une catégorie (risque, pas risque).

- On parle **d'apprentissage supervisé** car les *yi* permettent de guider le processus d'estimation

- **Exemples de méthodes** : Méthode du plus proche voisin, réseaux de neurones, Séparateurs à Vastes Marges etc..
- **Exemples d'applications** : détection de fraude, marketing téléphonique, changement d'opérateurs téléphonique etc…

Apprentissage supervise

- Ce sont des méthodes prédictives
- **Classement** : la variable a expliquer (ou cible , réponse, dépendante) est *qualitative*

- on parle aussi de **classification** (en anglais) ou **discrimination**
- **Prédiction** : la variable a expliquer est *quantitative*
- on parle aussi de **régression**

  - exemple : le prix d'un appartement (en fonction de sa superficie, de l' étage et du quartier)

- **Scoring** : classement applique a une problématique d'entreprise (variable a expliquer souvent binaire)
- chaque individu est affecté a une classe (risque ou non risque, par exemple) en fonction de ses caractéristiques

Apprentissage supervisé vs non-supervisé

- ***Apprentissage supervisé (Classement)*** : la variable expliquée est connue pour les n individus. k classes sont construites suivant les modalités des p+1 variables en présence afin d'affecter une valeur à la variable expliquée lorsque celle-ci est inconnue (*prédicteur*).

- • **Remarque** : la différence entre *classification* et *régression* vient du type du caractère Y :
  - ***Classification*** :Y est qualitative ou quantitative discrète.
  - ***Régression*** :Y est quantitative continue.
  - Les deux cas utilisent des méthodes quelque peu différentes.

- ***Apprentissage non-supervisé (Classification)*** : pas de variable expliquée. k classes sont construites suivant les modalités des p variables explicatives afin de représenter les données.

Quelques exemples

- Reconnaissance de la parole
- Identification d'empreinte digitale
- Identification de séquences ADN
- Credit scoring
  - prédire l'achat d'un produit ou service
  - prédire les impayés ou la fraude
  - prédire en temps réel les impayés
  - prédire le départ du client vers un concurrent
- En médecine : diagnostic (bonne santé: oui / non) en fonction du dossier du patient et des analyses médicales
- Courriels : spam (oui / non) en fonction des caractéristiques du message (fréquence des mots…)

Données

Présentation des données

- n individus, p variables (*variables explicatives* ou *co-variables*), 1 variable qualitative (*variable expliquée* ou *label*).
- A chaque individu i sont associées p valeurs *x1i , . . ., xpi* , correspondant aux valeurs prises par les variables explicatives X1i , . . . ,*Xpi* , et une valeur *yi*, correspondant à la valeur prise par la variable expliquée Yi. Les couples (Xi, Yi), (i=1,..,n) sont supposés indépendants.
- • Présentation sous forme de *tableau* à double entrée, avec en général la variable expliquée sur la première colonne :

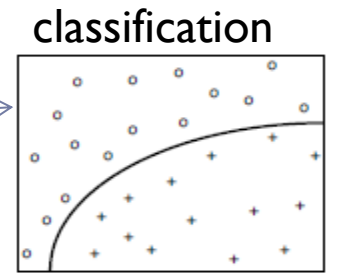
Individu \ Variable	Y	X^1	...	X^p
1	y_1	x_1^1	...	x_1^p
2	y_2	x_2^1	...	x_2^p
...	...	...	...	...
n	y_n	x_n^1	...	x_n^p

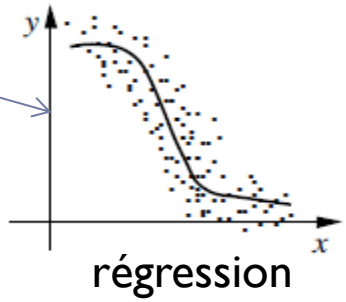
Objectif

Une variable Y à expliquer

- Le tableau de données contient une variable particulière Y qui doit être expliqué (ou prédite) en termes des autres variables Xi
  - Y est une variable catégorielle => problème de classification supervisée (discrimination)
  - Y est une variable numérique => problème de régression

Individu \ Variable	Y	X^1	…	X^p
1	y_1	x_1^1	…	x_1^p
2	y_2	x_2^1	…	x_2^p
…	…	…	…	…
n	y_n	x_n^1	…	x_n^p





Deux aspects : descriptive vs prédictive

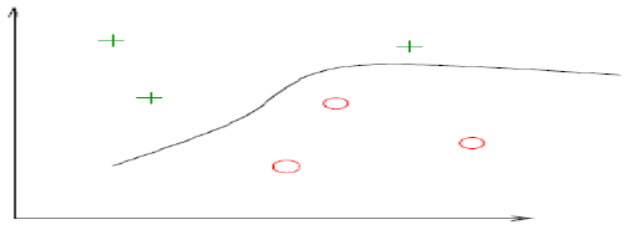
- Il y a deux aspects principaux:
  - - Un aspect descriptif: le modèle tente d'expliquer Y (la réponse ou la cible, la variable dépendante) en termes de variables aléatoires *xi* et leurs réalisations (les données disponibles)
  - - Un aspect prédictif: le modèle tente de prédire la valeur de Y en termes de variables aléatoires Xi; données dans le vecteur d'observations $x = [x1, x2, ..., xp]$ T
- La valeur prédite fournie par le modèle sera notée *y*
  - Label yi connu pour toutes les données *xi*.
  - **Problème** : prévoir une sortie y pour toute nouvelle entrée $x = (x1, . . . , xp)$.
  - **Méthode** : Apprendre un *prédicteur* f sur les données connues afin d'assigner le label $y=f(x)$ à toute nouvelle entrée x.

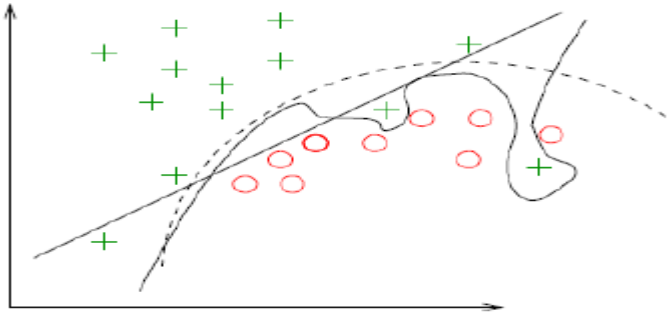
On cherche un modèle fiable

- L'objectif de la classification supervisée est principalement de définir des **règles permettant de classer** des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets. Les méthodes s'étendent souvent à des variables Y quantitatives (régression).

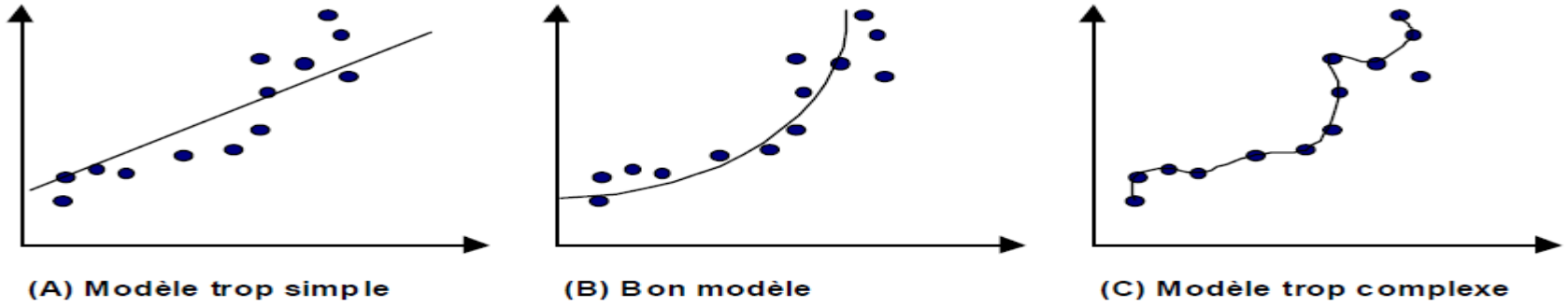
- On dispose au départ d'un **échantillon dit d'apprentissage** dont le classement est connu. Cet échantillon est utilisé pour l'apprentissage des règles de classement.

- Il est nécessaire d'étudier la **fiabilité** de ces règles pour les comparer et les appliquer, évaluer les cas de **sous apprentissage** ou de **sur apprentissage** (**complexité du modèle**). On utilise souvent un deuxième échantillon indépendant, dit de **validation ou de test**.



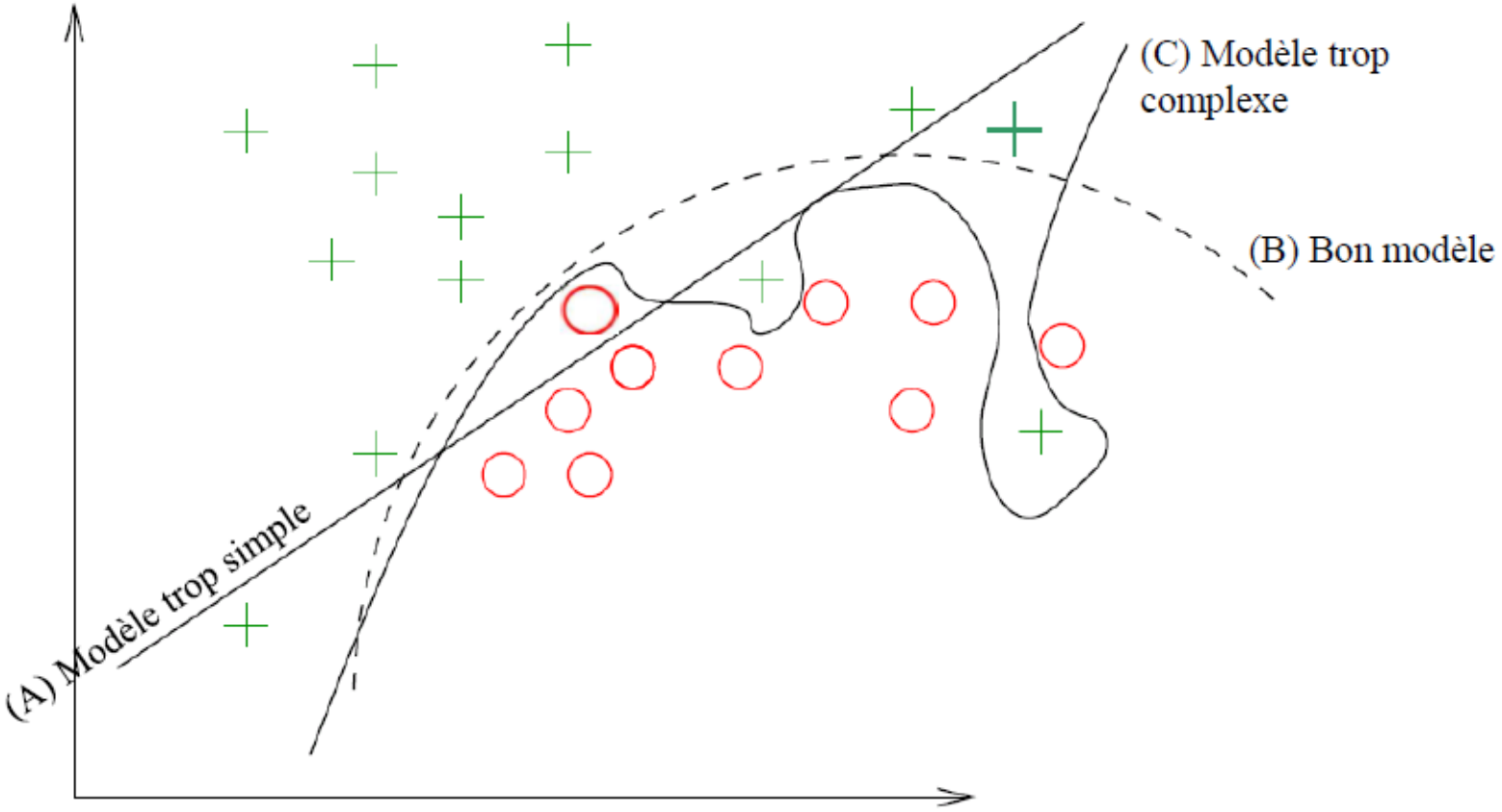


Sur-apprentissage en régression

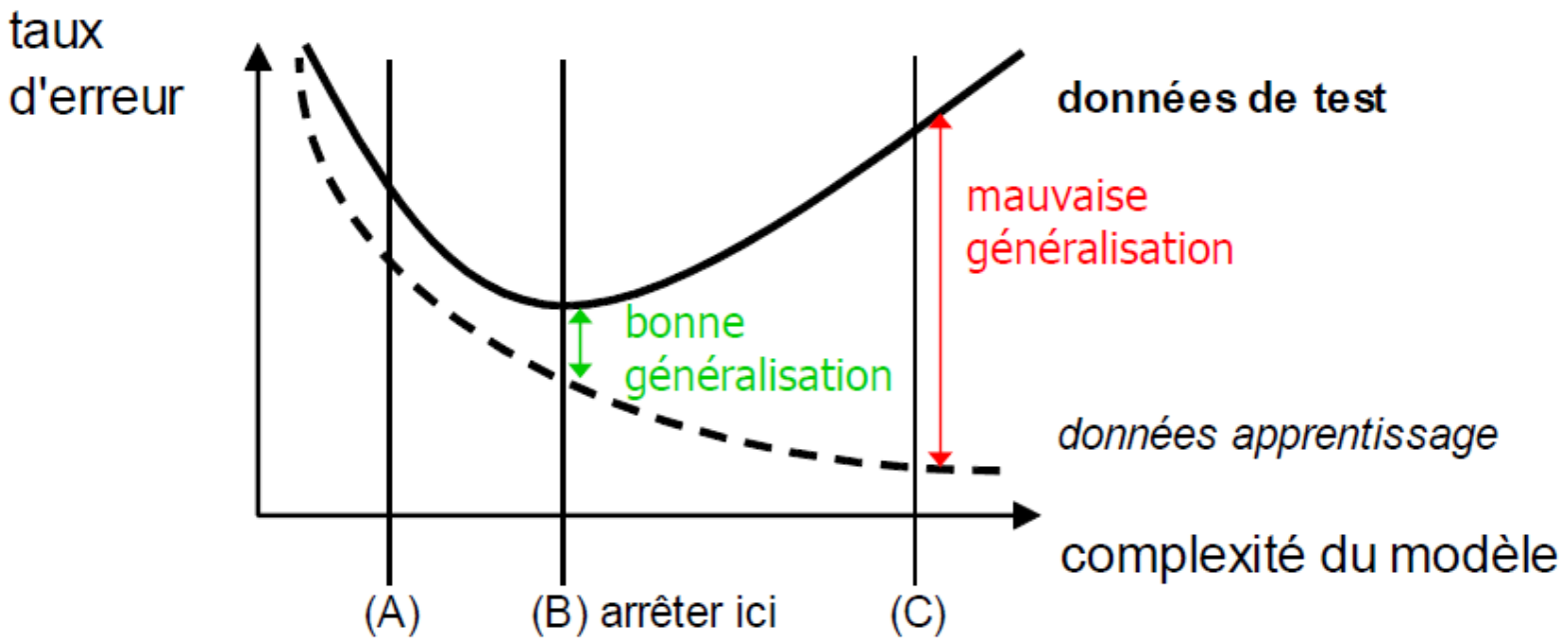


- Un modèle trop poussé dans la phase d'apprentissage :
  - épouse toutes les fluctuations de l' échantillon d'apprentissage,
  - détecte ainsi de fausses liaisons,
  - et les applique a tort sur d'autres échantillons
- On parle de sur-apprentissage ou sur-ajustement

Sur-apprentissage en classification



Taux d' erreur en fonction de la complexité du modèle



Démarche

Quatre étapes

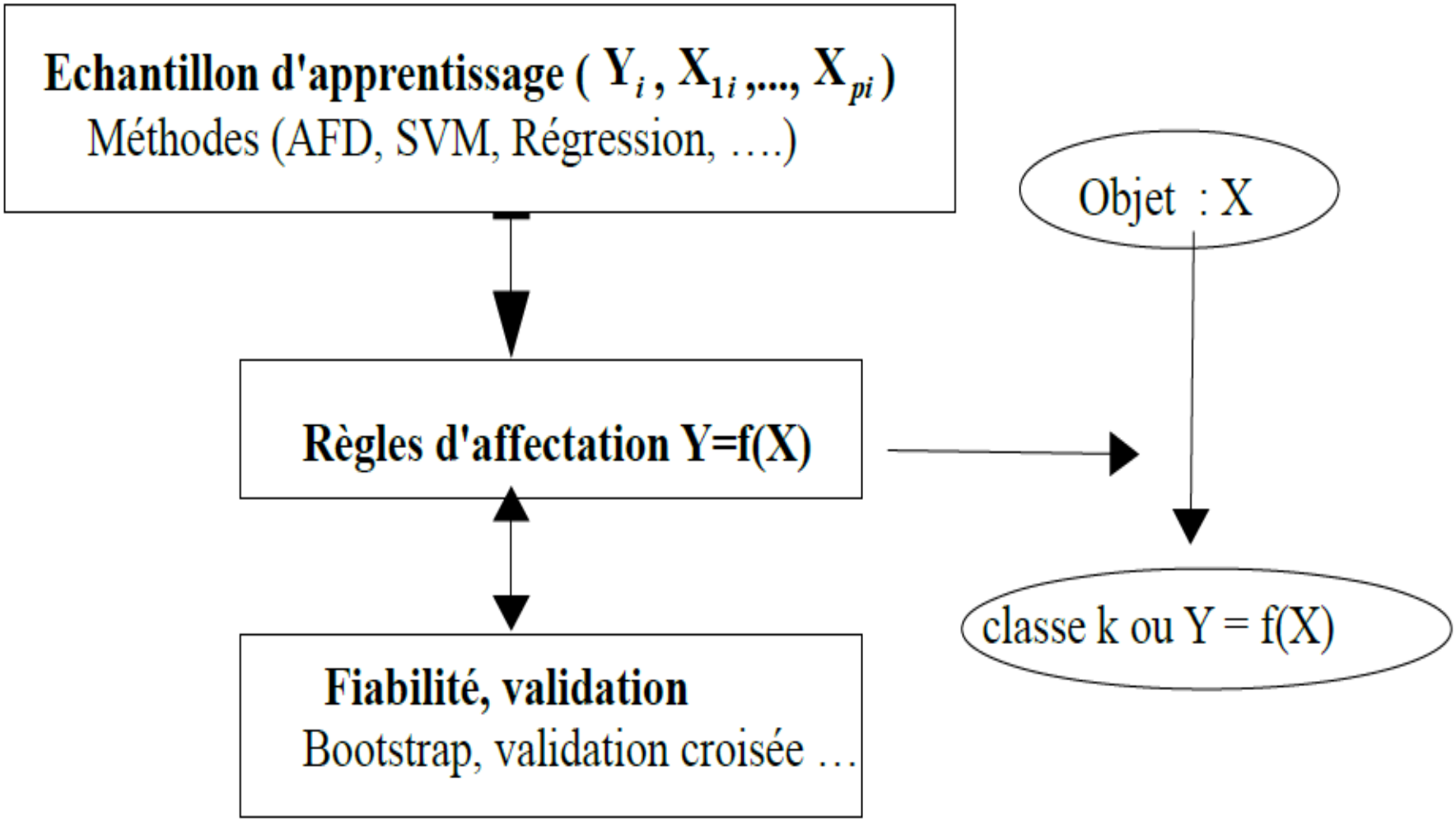
- Apprentissage : **construction du modèle** sur un 1er échantillon pour lequel on connait la valeur de la variable a expliquer

- Test : **vérification du modèle** sur un 2d échantillon pour lequel on connait la valeur de la variable a expliquer, que l'on compare a la valeur prédite par le modèle
  - si le résultat du test est insuffisant (d' après la *matrice de confusion* ou la courbe *ROC*), on recommence l'apprentissage

- **Validation du modèle** sur un 3e échantillon, éventuellement « out of time », pour avoir une idée du taux d'erreur non biaise du modèle

- **Application du modèle** a l'ensemble de la population
  - Application du meilleur modèle pour les nouveaux cas: {(xi,?)} (yi inconnue) = Phase de production

Démarche


Echantillon d'apprentissage (Y_i , X_{1i} ,..., X_{pi})
Méthodes (AFD, SVM, Régression, ….)
Règles d'affectation Y=f(X)
Fiabilité, validation
Bootstrap, validation croisée …
Objet : X
classe k ou Y = f(X)

Démarche

- On dispose de différentes **stratégies d'apprentissage** :

- • **Règle majoritaire** : à tout objet, on associe la classe k telle que P(k) est maximale.

- • **Règle du maximum de vraisemblance** : à tout objet on associe k telle que P(d/k) maximale.

- • **Règle de Bayes** : à tout objet on associe k telle que P(k/d) maximale.

Méthodes

Quelques méthodes classiques

- Analyse discriminante linéaire
  - Résultat explicite P(Y/ X1,… ,Xp) sous forme d'une formule
  - Requiert des Xi continues et des lois Xi/Y multi-normales et homoscédastiques (attention aux individus hors norme)
  - Optimale si les hypothèses sont remplies

- Régression logistique
  - Sans hypothèse sur les lois Xi/Y, Xi peut être discret, nécessaire absence de colinéarité entre les Xi
  - Méthode très souvent performante
  - Méthode la plus utilisée en scoring

- Arbres de décision
  - Règles complètement explicites
  - Traitent les données hétérogènes, éventuellement manquantes, sans hypothèses de distribution
  - Détection d'interactions et de phénomènes non linéaires
  - Mais moindre robustesse

Quelques méthodes classiques

- K plus proche voisins (KNN)
  - choix de K crucial : CV, AUC, éch. test…
  - plus n est grand, plus on peut se permettre de prend un K grand

- Séparateurs à Vastes Marges (SVM)
  - Efficace dans de nombreux domaines d'application
  - Utilise l'astuce noyau
  - Un bon paramétrage est nécessaire

Evaluation et comparaison de méthodes

Evaluation des méthodes de Classification

- Taux de bon classement: le taux d'erreur doit être le plus bas possible, et l'aire sous la courbe ROC la plus proche possible de 1

- La robustesse : la méthode être le moins sensible possible aux fluctuations aléatoires de certaines variables et aux valeurs manquantes, ne pas dépendre de l'échantillon d'apprentissage utilisé et bien se généraliser a d'autres échantillons

- La concision : les règles du modèle doivent être les plus simples et les moins nombreuses possible

Evaluation des méthodes de Classification

- Des résultats explicites : les règles du modelé doivent être accessibles et compréhensibles

- La diversité des types de données manipulées : toutes les méthodes ne sont pas aptes a traiter les données qualitatives, discrètes, continues et… manquantes

- La rapidité de calcul du modèle : un apprentissage trop long limite le nombre d'essais possibles

- Les possibilités de paramétrage : dans un classement, il est parfois intéressant de pouvoir pondérer les erreurs de classement, pour signifier, par exemple, qu'il est plus grave de classer un patient malade en « non-malade » que l'inverse

Comparaison de méthodes

- Comment comparer différentes méthodes de classification supervisée?
  - On utilisera un échantillon test : le taux de bon classement sera évalué sur un échantillon test n'ayant pas servi à estimer les règles de classement (découpage éch. existant en 2/3 apprentissage 1/3 test)

- la validation croisée (cross validation - CV) Leave One Out
  - la prédiction de la classe du ième individu est obtenue sans utiliser cet individu pour estimer les paramètres du modèle
  - la validation croisée K-fold où l'échantillon d'apprentissage est découpé en K partie, chaque partie servant tour à tour d'échantillon test (leave one out = n-fold)

- un critère de choix de modèles (BIC) pour les méthodes probabilistes

Comparaison de méthodes

- Comment comparer différentes méthodes de classification supervisée ?

- Les pièges à éviter
- erreur apparente : comparer des méthodes en terme de taux de bon classement sur l'échantillon d'apprentissage ayant servi à estimer les paramètres favorisera toujours les méthodes les plus complexes.

- Dans la classification des données déséquilibrées , par exemple, intrusions sur un réseau ou la détection de fraudes financières, on s'intéresse seulement à la classe minoritaire. Un taux de bon classement élevé ne signifie pas que toute intrusion est détectée.

- Par exemple, si on a 1% d'intrusions. On obtient 99% de taux de bon classement (sans rien faire).

Matrice de Confusion

		Prédit		Total
		Y=0	Y=1	
Réel	Y=0	VN	FP	N
	Y=1	FN	VP	P
Total		N1	P1	n

- Positif : relatif à une modalité de référence (Z = 1, malade, achat…)
- s : seuil tel que Y = 1 si p(Y = 1)>= s
- Se(s) : sensibilité (taux de vrais positifs)
- 1-Sp(s) = : 1-spécificité (taux de faux positifs)

$$Se(s) = \frac{VP}{VP + FN}. \qquad 1 - Sp(s) = 1 - \frac{VN}{VP + FP}.$$

Taux de bon classement (accuracy)

$$accuracy = \frac{VP + VN}{VP + FN + VN + FN}.$$

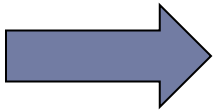
$$erreur = 1 - \frac{VP + VN}{VP + FN + VN + FN}. \qquad erreur = 1 - accuracy$$

Sensibilité et Spécificité

- Sensibilité = capacité à diagnostiquer les malades parmi les malades

- Spécificité = capacité à reconnaître les non-malades parmi les non-malades

- 1 - Spécificité = risque de diagnostiquer un malade chez les non-malades.



Trouver un compromis acceptable entre forte sensibilité et forte spécificité.

Mesures de précision et de rappel

$$p = \frac{VP}{VP + FP}. \qquad r = \frac{VP}{VP + FN}.$$

- **Précision p :** est le nombre d'exemples positifs correctement classés divisé par le nombre total d'exemples qui sont classés comme positifs.
- **Rappel r :** est le nombre d'exemples positifs correctement classés, divisé par le nombre total d'exemples positifs réels dans l'ensemble de test.

Un exemple

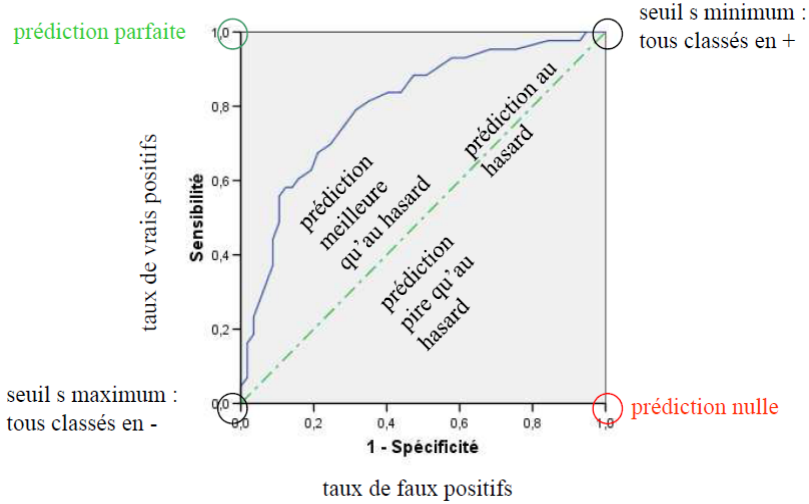
	Classified Positive	Classified Negative
Actual Positive	1	99
Actual Negative	0	1000

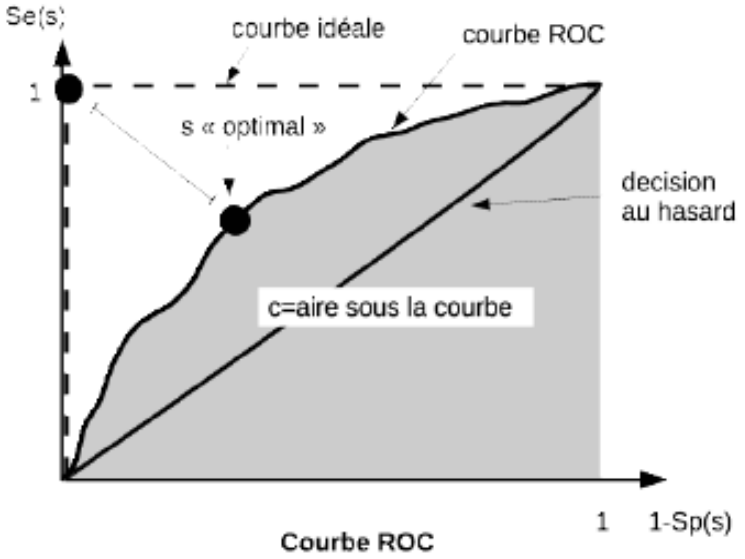
- Cette matrice de confusion donne
  p = précision de 100% et
  rappel r = 1%

- Parce que un seul exemple positif a été classé correctement et pas d'exemples négatifs mal classés.

- Remarque: la précision et le rappel mesurent uniquement la classification de la classe positive.

Courbe ROC





- sur l'axe Y : sensibilité = Se(s)
- sur l'axe X : 1 - spécificité = 1 - Sp(s)
- proportion y de vrais positifs en fonction de la proportion x de faux
- positifs, lorsque l'on fait varier le seuil s du score
- Aire AUC sous la courbe
- AUC = 0,5 ⇒ modèle pas meilleur qu'une notation aléatoire

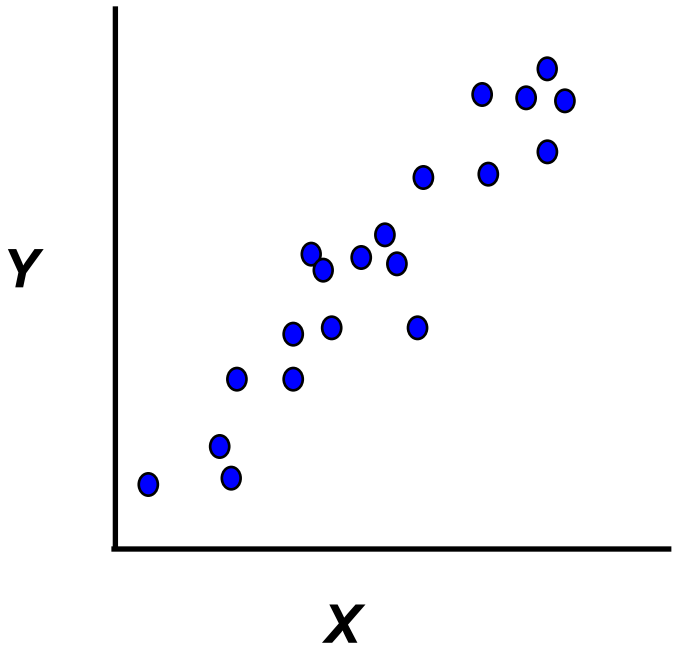
Quelques exemples sous R

Régression linéaire simple

Introduction

Etude de la relation entre deux variables quantitatives:

Nuage de points:



-description de l'association linéaire: corrélation, régression linéaire simple

- explication / prédiction d'une variable à partir de l'autre: modèle linéaire simple

La régression linéaire simple

1. Le modèle

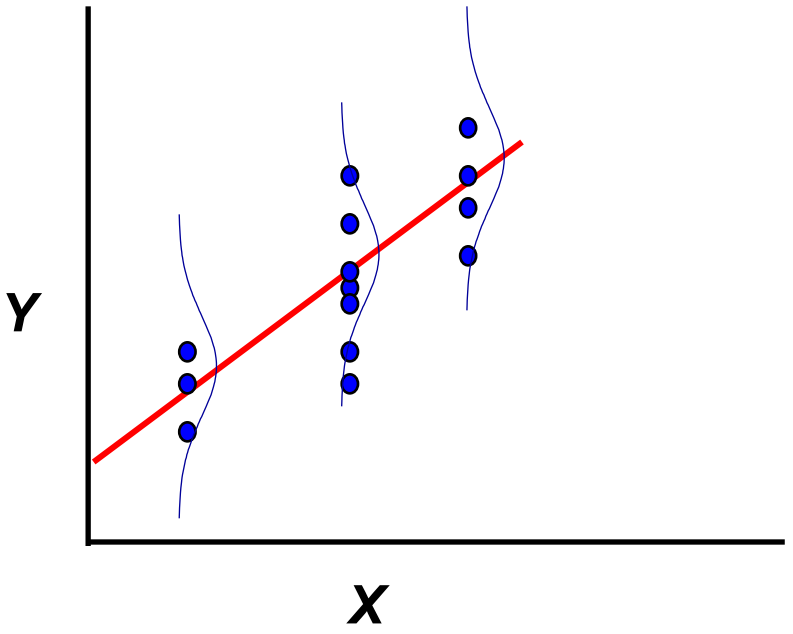
On suppose: $y = f(x) = a + bx$

Modèle: $Y_i = a + bX_i + e_i$ avec, pour $X = x_i$, Y_i : $N(a+bx_i, s)$

X = variable explicative (« indépendante »), contrôlée

Y = variable expliquée (dépendante), aléatoire

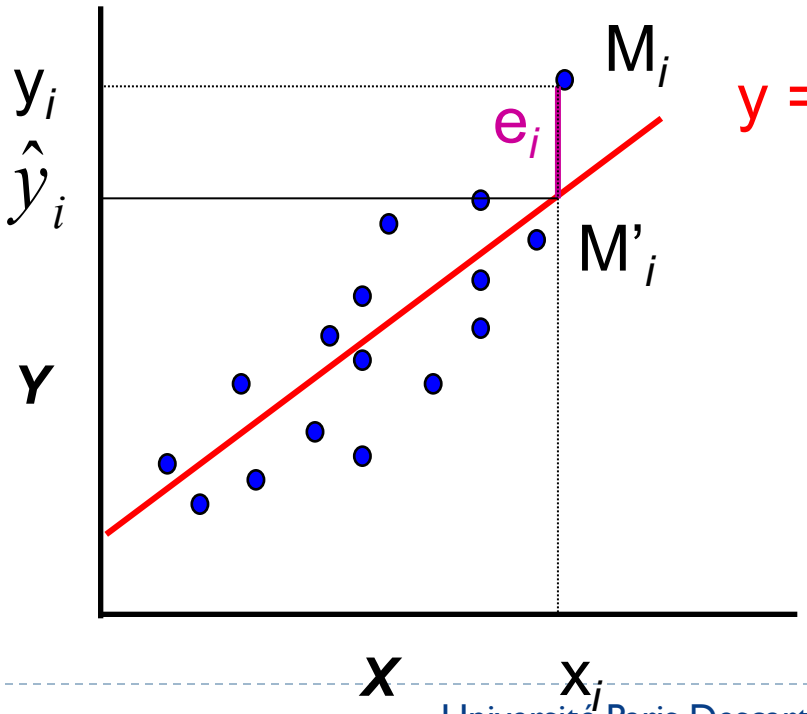
Relation de causalité ≠ interdépendance



La régression linéaire simple

2. L'estimation des paramètres

a? b?
Méthode d'estimation: les moindres carrés:



$y = a+bx$

$e_i = y_i - (a + bx_i)$

$\sum e_i^2$ minimale

La régression linéaire simple

2. L'estimation des paramètres

Méthode des moindres carrés

On cherche le minimum de $\sum_{i=1}^{n}(y_i-(a+bx_i))^2 = E(a,b)$

$$\begin{cases} \dfrac{\partial E}{\partial a} = \sum_{i=1}^{n} 2(y_i-(a+bx_i))(-1) = 0 \quad (1) \\ \dfrac{\partial E}{\partial b} = \sum_{i=1}^{n} 2(y_i-(a+bx_i))(-x_i) = 0 \quad (2) \end{cases}$$

La régression linéaire simple

2. L'estimation des paramètres

Méthode des moindres carrés

$$(1) \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (a + bx_i) = na + b\sum_{i=1}^{n} x_i$$

$$n\bar{y} = na + nb\bar{x}$$

$$a = \bar{y} - b\bar{x}$$

La régression linéaire simple

2. L'estimation des paramètres

Méthode des moindres carrés

$$n(\mathrm{cov}(x,y)+\bar{x}\bar{y})-(\bar{y}-b\bar{x})n\bar{x}-bn(s_x^2+\bar{x}^2)=0$$

$$\mathrm{cov}(x,y)=bs_x^2 \qquad b=\frac{\mathrm{cov}(x,y)}{s_x^2}$$

Si y = a+bx alors $\hat{b}=\dfrac{\mathrm{cov}(x,y)}{s_x^2}$ et $\hat{a}=\bar{y}-b\bar{x}$

On peut alors prédire y pour x compris dans l'intervalle des valeurs de l'échantillon:

$$\hat{y}_i=\hat{a}+\hat{b}x_i$$

La régression linéaire simple

3. Qualité de l'ajustement

On a supposé: $Y_i = a + bX_i + e_i$ avec

pour $X = x_i$, Y_i : $N(a+bx_i, s)$

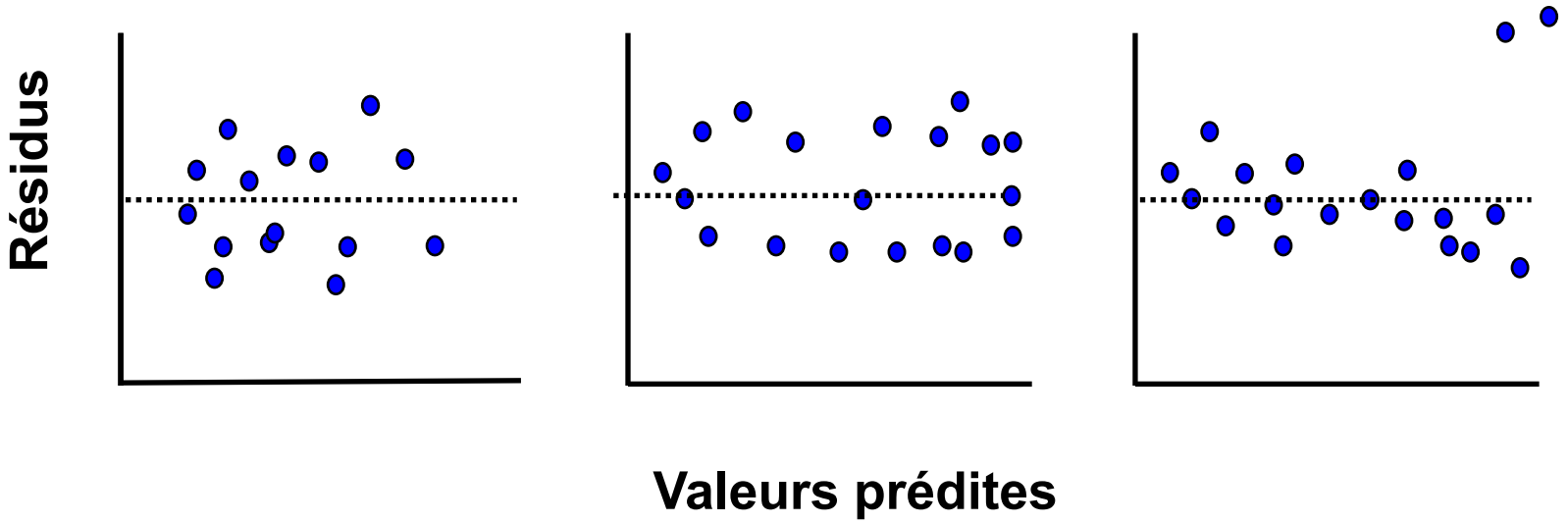
- distribution normale des erreurs
- variance identique (homoscédasticité)
- indépendance: $\text{cov}(e_i, e_j) = 0$
- linéarité de la relation

Test *a posteriori* : étude du nuage de points/ du graphe des résidus

La régression linéaire simple

3. Qualité de l'ajustement

Normalité de l'erreur

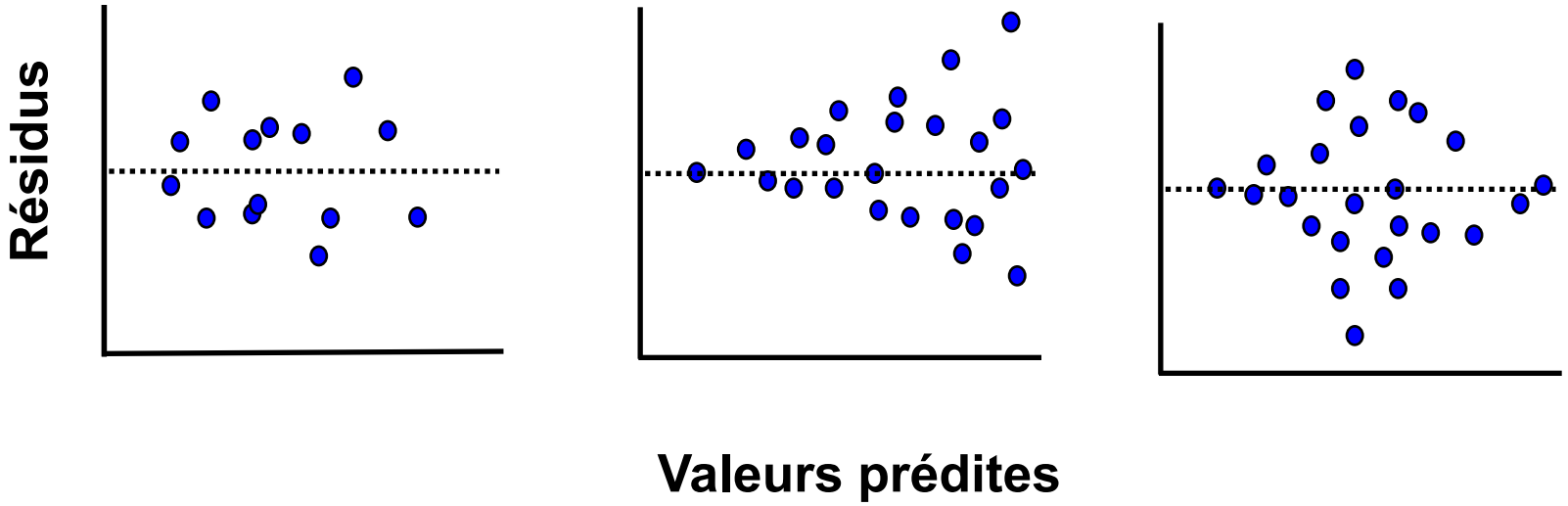


Questions à se poser: structure de l'erreur?
Valeurs extrêmes: ont-elles un sens biologique?
Influencent-elles l'estimation des paramètres?

La régression linéaire simple

3. Qualité de l'ajustement

Homoscédasticité

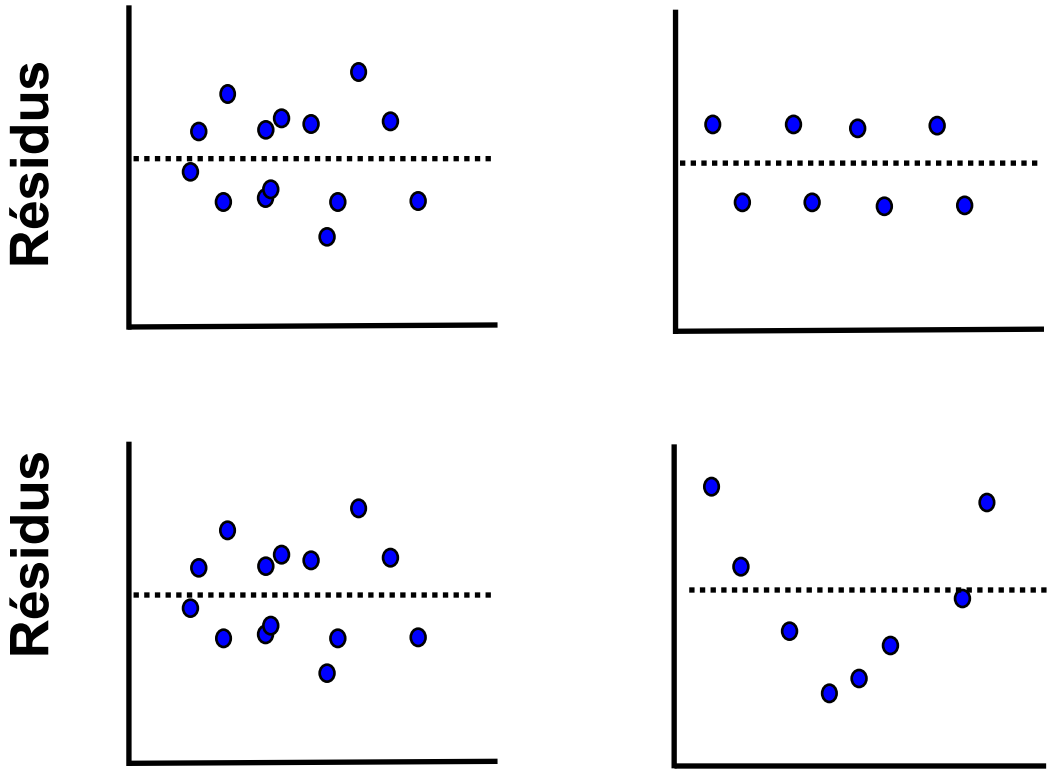


Possibilité de transformation: attention aux transformations *ad hoc*

La régression linéaire simple

3. Qualité de l'ajustement

Indépendance entre erreurs, linéarité



Structure de l'erreur?

Relation non linéaire?

4. Coefficient de détermination

Décomposition de la variation

Quelle part de la variabilité de Y est expliquée par la relation linéaire avec X?

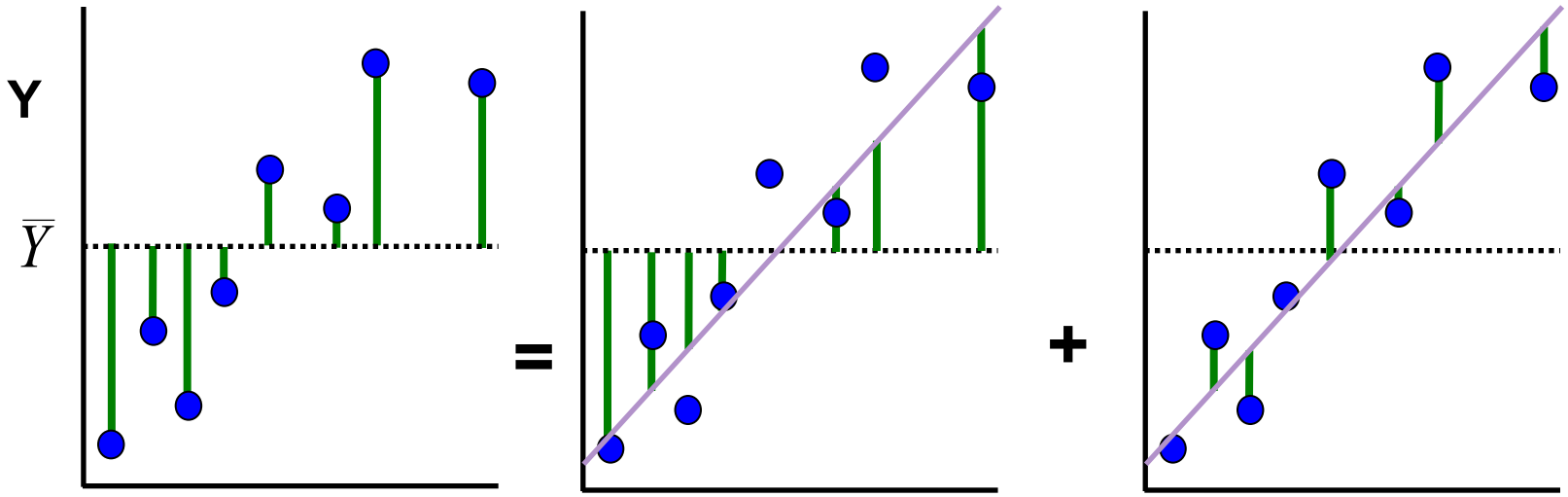
Variabilité? Somme des Carrés des Ecarts SCE:

$$SCE_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = ns_y^2$$

La régression linéaire simple

4. Coefficient de détermination

Décomposition de la variation



***SCE* Totale** **= *SCE* reg.lin. (Expliquée)** **+ *SCE* hors reg.lin. (erreur)**

$$\sum_{i=1}^{N}(Y_i-\overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2 + \sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2$$

La régression linéaire simple

4. Coefficient de détermination

La décomposition de la SCE permet d'estimer la part de SCE de Y expliquée par la régression:

$$r^2 = \frac{SCE_{reg.lin.}}{SCE_T}$$ Coefficient de détermination

$0 \leq r^2 \leq 1$