MA431 : Data

Données et échantillons

D. Barcelo

Grenoble INP ESISAR

2022/2023

Introduction

De la statistique à la data science

- Avant les années 70 : Statistique avec échantillon représentatif (30 individus sur 10 variables)
- les années 70 : Science exploratoire des données (analyse des données)
- les années 80 : modèles statistiques non paramétriques et début des réseaux de neurones,
- les années 90 : début du data mining (fouille de données), émergence du statistical learning,
- les années 00 : le nombre de variables explose (supérieur à 10^4), on parle d'apprentissage statistique
- les années 10 : le nombre d'individus explose, la data science arrive.

Environnement

R ou Python

- Toutes les méthodes d'apprentissage sont implémentées en R (packages),
- R langage interprété, les temps d'exécution peuvent être très long,
- Python et la librairie Scikit-learn dispose des principales méthodes d'apprentissage,
- Python plus rapide que R,
- De manière générale, R pratique pour modéliser et interpréter, Python pour modéliser efficacement et effectuer des prévisions.

Classification ou régression ?

- Les données sont collectées avant l'analyse.
- On observe p variables $X = (X_1, \ldots, X_p)$ sur n individus.
- Objectif : construction d'un modèle de prédiction d'une variable Y.
  1. si Y est quantitative, on parle de régression.
  2. si Y est qualitative, on parle de classification.

Apprentissage statistique

1. Extraction des données.
2. Exploration des données.
3. Traitement des valeurs manquantes.
4. Partition des données pour validation du modèle.
5. Construction du modèle à partir d'une base d'apprentissage.
6. Validation sur une base test.
7. Comparaison de différents modèles.
8. Choix du meilleur modèle.
9. Utilisation sur de nouvelles données.

Les bases de données

- Temps de préparation d'une base de données très important.
- Parfois utilisation de bases existantes.
- Il existe des bases publiques de données d'attaque :
  1. NSL-KDD (évolution de KDD99),
  2. CTU-13 (botnets),
  3. UNSW-NB15 (Académie des forces de défense australienne),
  4. CICDS18 (network traffic),
  5. etc. .

Données manquantes

Comment traiter les données manquantes des bases de données ?

Données manquantes

On peut supprimer des données :

1. On ne conserve que les individus "complets" (risque),
2. On supprime la variable avec des données manquantes du jeu de données.

Ou bien compléter la base :

1. en remplaçant par la dernière valeur,
2. en remplaçant par la moyenne ou la médiane,
3. en utilisant une méthode d'apprentissage supervisé : le kNN (k plus proches voisins)
4. en effectuant une régression linéaire locale.

Supervisé ou non supervisé

L'apprentissage statistique peut être

- **supervisé** :
  Y discrète ou qualitative : Classification.
  Y continue : Régression.
- **non supervisé** :
  Y discrète : Custering.

Apprentissage non supervisé

Clustering

Le clustering (ou *classification automatique*) permet de regrouper des individus dans des classes (clusters) non définies à priori. Il s'agit d'un apprentissage automatique non supervisé. Les classes sont déterminées au cours de l'algorithme. Elles regroupent des individus ayant des caractéristiques similaires et séparent ceux qui ont des caractéristiques différentes.

Apprentissage supervisé

Classification

La classification (ou *classement*) permet d'affecter des individus à des classes existantes à priori en fonction de ses caractéristiques. Il s'agit d'un apprentissage supervisé. Le résultat de la classification est un algorithme permettant d'affecter chaque individu à la meilleure classe.

Régression

La régression est la recherche d'un modèle pour prévoir les valeurs d'une variable continue.
On recherche une fonction minimisant les erreurs d'approximation commises.

Apprentissage supervisé

Pour réaliser un apprentissage supervisé, on a besoin d'au moins deux jeux de données : un pour l'apprentissage et pour le test.

Base d'apprentissage

Jeu de données utilisé pour ajuster les paramètres du classifieur ou du modèle.
Objectif : obtenir un modèle qui se généralise bien à des données inconnues.
Souvent de taille importante, attention à pouvoir généraliser le modèle et à éviter le surapprentissage.

Apprentissage supervisé

Base test

Jeu de données indépendant de la base d'apprentissage mais qui possède la même distribution de probabilité des valeurs des variables.
Objectif : évaluer la validité du modèle entrainé sur la base d'apprentissage.
Si le modèle s'adapte peu à la base test mais beaucoup à la base d'apprentissage, il y a un risque de surapprentissage.

Exemple

Exemple de construction d'une base d'apprentissage et d'une base test.

Erreurs dans une régression

Erreurs dans une régression :

- Erreur de prévision,
- Mesurer les écarts entre valeur réelle et prévision,
- Mesure d'un écart quadratique
- Indicateurs spécifiques au modèle

Erreurs dans une classification

Matrice de confusion :

Classe prévue Classe réelle	Positif	Négatif
Positif	TP	FN
Négatif	FP	TN

Erreurs dans une classification

A partir de la matrice de confusion, on peut calculer différents indicateurs :

- le taux de vrais positifs (ou sensibilité) : $\frac{TP}{TP + FN}$
- le taux de vrais négatifs (ou spécificité) : $\frac{TN}{FP + TN}$
- le taux de faux positifs : $\frac{FP}{FP + TN}$
- le taux de faux négatifs : $\frac{FN}{TP + FN}$
- la précision : $\frac{TP + TN}{TP + FN + FP + TN}$

Qualité d'une classification

Le coefficient de corrélation de Matthews (MCC

Il mesure la qualité des classifications binaires.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Très efficace pour évaluer la qualité de la classification, c'est un coefficient de corrélation entre les classes prédites et les classes réelles.

On a : $-1 \leqslant MCC \leqslant 1$.

- MCC=1 prédiction parfaite
- MCC=0 équivaut à une prédiction aléatoire
- MCC=-1 prédiction opposée

Courbe ROC

Courbe ROC

- Visualiser le pouvoir discriminant d'un modèle.
- Courbe Receiver Operating Characteristic
- Représente la sensibilité (taux de vrais positifs) en fonction de 1-spécifité (taux de faux positifs)

Courbe ROC

Courbe ROC

- Visualiser le pouvoir discriminant d'un modèle.
- Courbe Receiver Operating Characteristic
- Représente $\alpha(s)$ en fonction de $1 - \beta(s)$

$$\alpha(s) \approx \text{Proportion de vrais positifs au score supérieur à } s$$

$$\beta(s) \approx \text{Proportion de vrais négatifs au score supérieur à } s$$

$$1 - \beta(s) \approx \text{Proportion de faux positifs au score supérieur à } s$$

Courbe ROC

Courbe ROC

- Pour un seuil $s = 1$:
- Pour un seuil $s = 0$:

- Modèle parfait :
- Modèle aléatoire :

Courbe ROC

Courbe ROC

- Pour un seuil $s = 1$: Ni vrais positifs ni faux positifs donc point $(0\,;0)$
- Pour un seuil $s = 0$:

- Modèle parfait :
- Modèle aléatoire :

Courbe ROC

Courbe ROC

- Pour un seuil $s = 1$: Ni vrais positifs ni faux positifs donc point $(0\,;0)$
- Pour un seuil $s = 0$: Tous les vrais positifs et tous les faux positifs donc point $(1\,;1)$
- Modèle parfait :
- Modèle aléatoire :

Courbe ROC

Courbe ROC

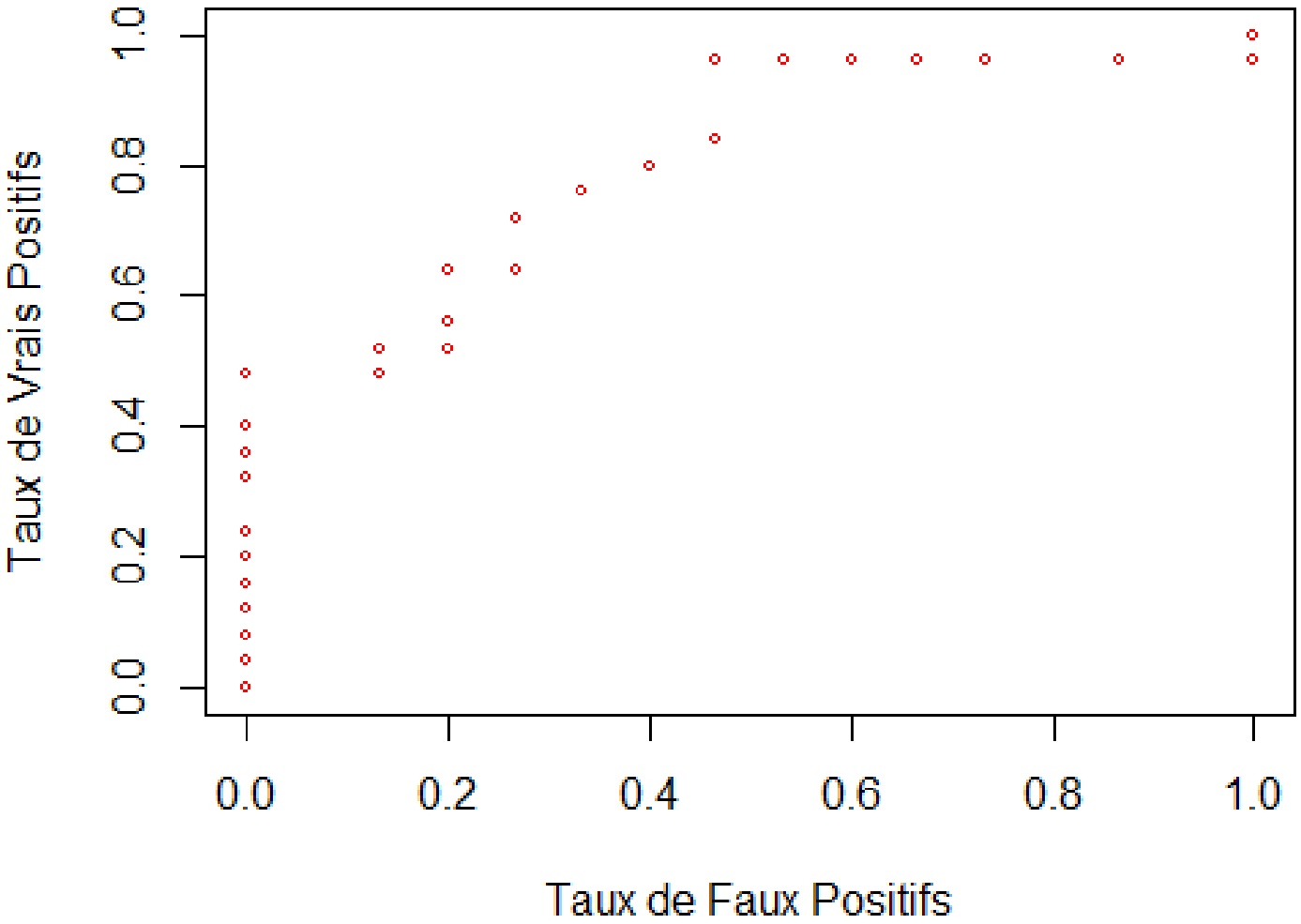
- Pour un seuil $s = 1$: Ni vrais positifs ni faux positifs donc point $(0\,;0)$
- Pour un seuil $s = 0$: Tous les vrais positifs et tous les faux positifs donc point $(1\,;1)$
- Modèle parfait : Tous les vrais positifs et aucun faux positifs
- Modèle aléatoire :

Courbe ROC

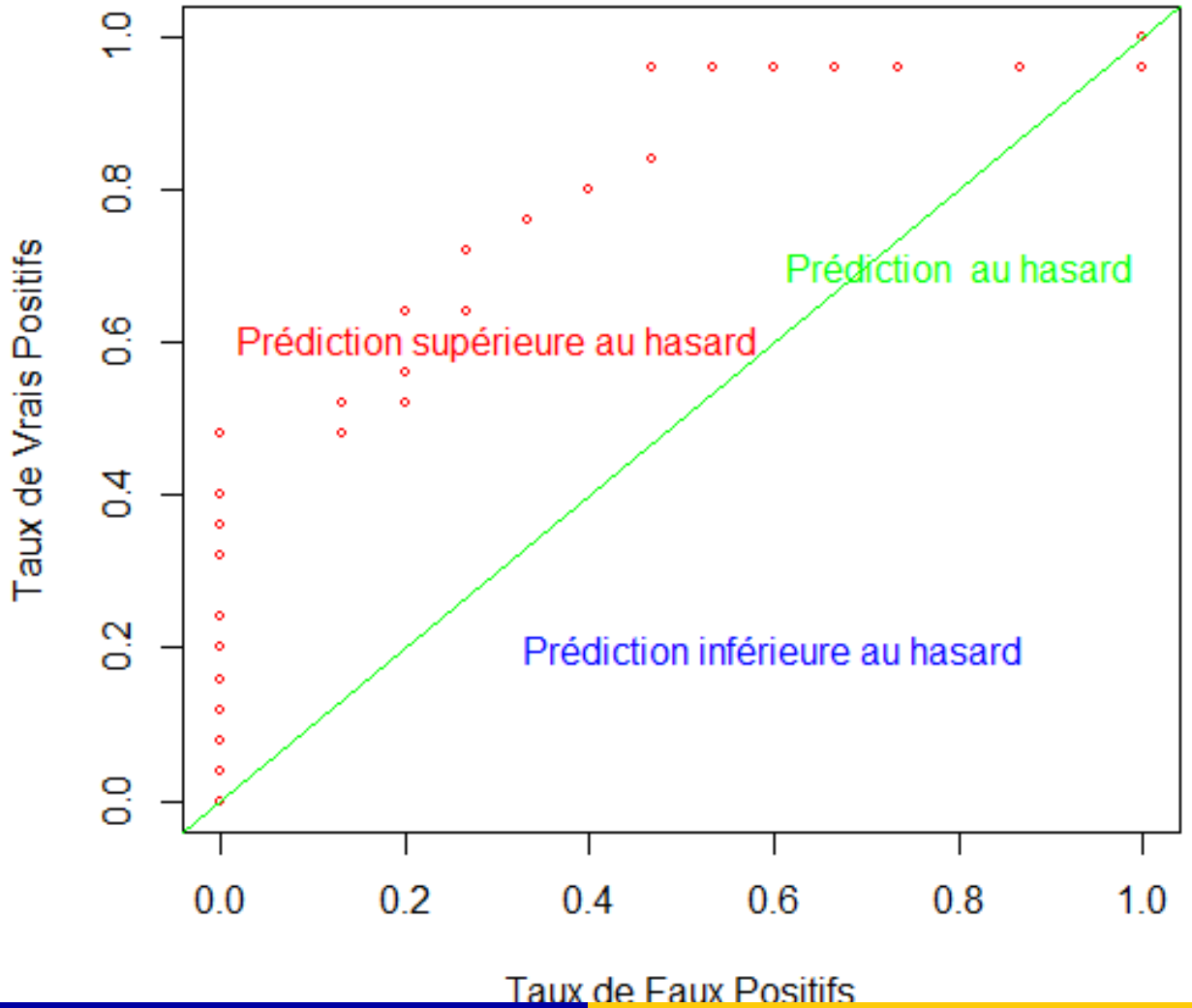
Courbe ROC

- Pour un seuil $s = 1$: Ni vrais positifs ni faux positifs donc point $(0\,;0)$
- Pour un seuil $s = 0$: Tous les vrais positifs et tous les faux positifs donc point $(1\,;1)$
- Modèle parfait : Tous les vrais positifs et aucun faux positifs
- Modèle aléatoire : Autant de vrais positifs que de faux positifs

Courbe ROC


Taux de Vrais Positifs
0.0
0.2
0.4
0.6
0.8
1.0
Taux de Faux Positifs
0.0
0.2
0.4
0.6
0.8
1.0

Courbe ROC


Taux de Vrais Positifs
0.0
0.2
0.4
0.6
0.8
1.0
Prédiction au hasard
Prédiction supérieure au hasard
Prédiction inférieure au hasard
Taux de Faux Positifs

Courbe ROC

Critère AUC

Pour comparer différentes modélisations logistiques : critère AUC

- Area Under the Curve
- Modèle performant qui sépare les vrais positifs des faux positifs :

- AUC=0,5 :
- AUC<0,5 :
- Comparaison de deux modèles :

Courbe ROC

Critère AUC

Pour comparer différentes modélisations logistiques : critère AUC

- Area Under the Curve
- Modèle performant qui sépare les vrais positifs des faux positifs : AUC proche de 1
- AUC=0,5 :
- AUC<0,5 :
- Comparaison de deux modèles :

Courbe ROC

Critère AUC

Pour comparer différentes modélisations logistiques : critère AUC

- Area Under the Curve
- Modèle performant qui sépare les vrais positifs des faux positifs : AUC proche de 1
- AUC=0,5 : autant tirer à pile ou face les affectations
- AUC<0,5 :
- Comparaison de deux modèles :

Courbe ROC

Critère AUC

Pour comparer différentes modélisations logistiques : critère AUC

- Area Under the Curve
- Modèle performant qui sépare les vrais positifs des faux positifs : AUC proche de 1
- AUC=0,5 : autant tirer à pile ou face les affectations
- AUC<0,5 : autant tirer à pile ou face les affectations
- Comparaison de deux modèles :

Courbe ROC

Critère AUC

Pour comparer différentes modélisations logistiques : critère AUC

- Area Under the Curve
- Modèle performant qui sépare les vrais positifs des faux positifs : AUC proche de 1
- AUC=0,5 : autant tirer à pile ou face les affectations
- AUC<0,5 : autant tirer à pile ou face les affectations
- Comparaison de deux modèles : comparaison des AUC