



Apprentissage statistique sur données cliniques longitudinales

M2 Données Massives en Santé

le 13/01/2020

Simon Bussy

<https://simonbussy.com>

Postdoc @INSERM, UMRS 1138, Centre de Recherche des Cordeliers
Cofounder, CTO & Head of Research @Califrais

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

M2 — DMS
Machine learning
2/10

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Programme de la journée

M2 — DMS
Machine learning
2/10

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie
- ▶ Grande dimension

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie
- ▶ Grande dimension
- ▶ Données longitudinales

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie
- ▶ Grande dimension
- ▶ Données longitudinales

Programme

- I. Petit rappel théorique de l'apprentissage statistique

Programme

Rappel

Généralités

MRE

Notions

Données longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie
- ▶ Grande dimension
- ▶ Données longitudinales

Programme

- I. Petit rappel théorique de l'apprentissage statistique
- II. Particularité des données longitudinales

Programme

Rappel

Généralités

MRE

Notions

Données longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie
- ▶ Grande dimension
- ▶ Données longitudinales

Programme

- I. Petit rappel théorique de l'apprentissage statistique
- II. Particularité des données longitudinales
- III. TP sous python sur les données PBCseq...

Programme

Rappel

Généralités

MRE

Notions

Données longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme de la journée

Intitulé du cours :

Apprentissage statistique sur données cliniques longitudinales

Cadre

- ▶ Apprentissage supervisé
- ▶ Analyse de survie
- ▶ Grande dimension
- ▶ Données longitudinales

Programme

- I. Petit rappel théorique de l'apprentissage statistique
- II. Particularité des données longitudinales
- III. TP sous python sur les données PBCseq... en autonomie l'après-midi → questions à envoyer à 17h30 max à simon.bussy@gmail.com (hors bug de code!)

Programme

Rappel

Généralités

MRE

Notions

Données longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

I. Petit rappel théorique de l'apprentissage statistique

Apprentissage supervisé

- Échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \mathbb{R}$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Apprentissage supervisé

- ▶ Échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \mathbb{R}$
- ▶ $\mathcal{D}_n$: n réalisations $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}$ inconnue

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Apprentissage supervisé

- ▶ Échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \mathbb{R}$
- ▶ $\mathcal{D}_n$: n réalisations $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}$ inconnue

But: fonction de prédiction

- ▶ Construire $g : \mathcal{X} \rightarrow \mathcal{Y}$ pour prédire Y pour un nouveau $X = x \in \mathcal{X}$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Apprentissage supervisé

- ▶ Échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \mathbb{R}$
- ▶ $\mathcal{D}_n$: n réalisations $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}$ inconnue

But: fonction de prédiction

- ▶ Construire $g : \mathcal{X} \rightarrow \mathcal{Y}$ pour prédire Y pour un nouveau $X = x \in \mathcal{X}$
- ▶ Qualité mesurée par le risque intégré [3], ou erreur de généralisation, défini par $R_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[\ell(Y, g(X))]$, avec la fonction de perte $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Apprentissage supervisé

- ▶ Échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \mathbb{R}$
- ▶ $\mathcal{D}_n$: n réalisations $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}$ inconnue

But: fonction de prédiction

- ▶ Construire $g : \mathcal{X} \rightarrow \mathcal{Y}$ pour prédire Y pour un nouveau $X = x \in \mathcal{X}$
- ▶ Qualité mesurée par le risque intégré [3], ou erreur de généralisation, défini par $R_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[\ell(Y, g(X))]$, avec la fonction de perte $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$

Fonction oracle

- ▶ $g_{\mathbb{P}}^* \in \operatorname{argmin}_{g \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} R_{\mathbb{P}}(g)$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Apprentissage supervisé

- ▶ Échantillon $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ où pour tout $i \in \{1, \dots, n\}$, $x_i \in \mathcal{X} \subset \mathbb{R}^d$ et $y_i \in \mathcal{Y} \subset \mathbb{R}$
- ▶ $\mathcal{D}_n$: n réalisations $(X_i, Y_i) \stackrel{i.i.d.}{\sim} \mathbb{P}$ inconnue

But: fonction de prédiction

- ▶ Construire $g : \mathcal{X} \rightarrow \mathcal{Y}$ pour prédire Y pour un nouveau $X = x \in \mathcal{X}$
- ▶ Qualité mesurée par le risque intégré [3], ou erreur de généralisation, défini par $R_{\mathbb{P}}(g) = \mathbb{E}_{\mathbb{P}}[\ell(Y, g(X))]$, avec la fonction de perte $\ell : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}$

Fonction oracle

- ▶ $g_{\mathbb{P}}^* \in \operatorname{argmin}_{g \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} R_{\mathbb{P}}(g)$
- ▶ On cherche à l'approcher, mais elle est inconnue!

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Équivalent empirique

- Estimer $R_{\mathbb{P}}(g)$ par $R_n(g) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(X_i))$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Minimisation du risque empirique (MRE)

Équivalent empirique

- ▶ Estimer $R_{\mathbb{P}}(g)$ par $R_n(g) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(X_i))$
- ▶ LFGN: $R_n(g) \xrightarrow[n \rightarrow +\infty]{p.s.} R_{\mathbb{P}}(g)$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Équivalent empirique

- ▶ Estimer $R_{\mathbb{P}}(g)$ par $R_n(g) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(X_i))$
- ▶ LFGN: $R_n(g) \xrightarrow[n \rightarrow +\infty]{p.s.} R_{\mathbb{P}}(g)$

Algorithme d'apprentissage de MRE [4]

- ▶ Défini par

$$g_{n,\mathcal{G}} \in \operatorname{argmin}_{g \in \mathcal{G}} R_n(g)$$

avec $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Équivalent empirique

- ▶ Estimer $R_{\mathbb{P}}(g)$ par $R_n(g) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(X_i))$
- ▶ LFGN: $R_n(g) \xrightarrow[n \rightarrow +\infty]{p.s.} R_{\mathbb{P}}(g)$

Algorithme d'apprentissage de MRE [4]

- ▶ Défini par

$$g_{n,\mathcal{G}} \in \operatorname{argmin}_{g \in \mathcal{G}} R_n(g)$$

avec $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$

- ▶ $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y}) \Rightarrow$ sur-apprentissage : $R_n(g) \ll R_{\mathbb{P}}$, même quand $n \rightarrow +\infty$

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Équivalent empirique

- ▶ Estimer $R_{\mathbb{P}}(g)$ par $R_n(g) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(X_i))$
- ▶ LFGN: $R_n(g) \xrightarrow[n \rightarrow +\infty]{p.s.} R_{\mathbb{P}}(g)$

Algorithme d'apprentissage de MRE [4]

- ▶ Défini par

$$g_{n,\mathcal{G}} \in \operatorname{argmin}_{g \in \mathcal{G}} R_n(g)$$

avec $\mathcal{G} \subset \mathcal{F}(\mathcal{X}, \mathcal{Y})$

- ▶ $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y}) \Rightarrow$ sur-apprentissage : $R_n(g) \ll R_{\mathbb{P}}$, même quand $n \rightarrow +\infty$
- ▶ Choisir $\mathcal{G}$ suffisamment complexe pour approcher $g_{\mathbb{P}}^*$, mais pas trop pour éviter le sur-apprentissage

Quelques notions utiles aujourd'hui

Classification binaire

- Classification: $\mathcal{Y}$ fini ; binaire: $|\mathcal{Y}| = 2$

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Classification binaire

- Classification: $\mathcal{Y}$ fini ; binaire: $|\mathcal{Y}| = 2$
- Perte classique $\ell(y_1, y_2) = \mathbb{1}_{\{y_1 \neq y_2\}}$, ou logistique $\ell(y_1, y_2) = \log(1 + \exp(-y_1 y_2))$

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

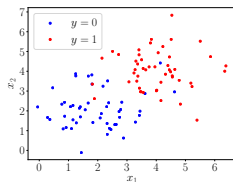
Références

Quelques notions utiles aujourd'hui

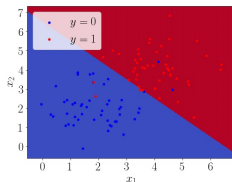
Classification binaire

- Classification: $\mathcal{Y}$ fini ; binaire: $|\mathcal{Y}| = 2$
- Perte classique $\ell(y_1, y_2) = \mathbb{1}_{\{y_1 \neq y_2\}}$, ou logistique $\ell(y_1, y_2) = \log(1 + \exp(-y_1 y_2))$
- Prédicteurs linéaires $\mathcal{G} = \mathcal{L}(\mathcal{X}, \mathcal{Y})$, soit $g_\beta : x \mapsto x^\top \beta$

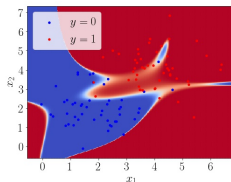
Sur-apprentissage



(a) Données simulées



(b) Oracle g_F^*



(c) g

Figure 1: Covariables arbitrairement choisies sont des produits de puissances de x_1 et x_2 , par exemple $x_1^4 x_2^2$ ou $x_1^3 x_2^3$.

<https://github.com/SimonBussy/thesis-illustrations>

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Classification binaire

- ▶ Classification: $\mathcal{Y}$ fini ; binaire: $|\mathcal{Y}| = 2$
- ▶ Perte classique $\ell(y_1, y_2) = \mathbb{1}_{\{y_1 \neq y_2\}}$, ou logistique $\ell(y_1, y_2) = \log(1 + \exp(-y_1 y_2))$
- ▶ Prédicteurs linéaires $\mathcal{G} = \mathcal{L}(\mathcal{X}, \mathcal{Y})$, soit $g_\beta : x \mapsto x^\top \beta$

Sur-apprentissage

Régularisation

- ▶ $\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2 + \operatorname{pen}(\beta)$

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Classification binaire

- ▶ Classification: $\mathcal{Y}$ fini ; binaire: $|\mathcal{Y}| = 2$
- ▶ Perte classique $\ell(y_1, y_2) = \mathbb{1}_{\{y_1 \neq y_2\}}$, ou logistique $\ell(y_1, y_2) = \log(1 + \exp(-y_1 y_2))$
- ▶ Prédicteurs linéaires $\mathcal{G} = \mathcal{L}(\mathcal{X}, \mathcal{Y})$, soit $g_\beta : x \mapsto x^\top \beta$

Sur-apprentissage

Régularisation

- ▶ $\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2 + \operatorname{pen}(\beta)$
- ▶ $\operatorname{pen}(\beta) = \lambda \|\beta\|_p = \lambda (\sum_{j=1}^d |\beta_j|^p)^{1/p}$ où $0 \leq p \leq +\infty$
et $\lambda \in \mathbb{R}^+$ est un hyper-paramètre

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme

Rappel

Généralités

MRE

Notions

**Données
longitudinales**

Problèmes

Solutions

TP python

Installation

PBCseq

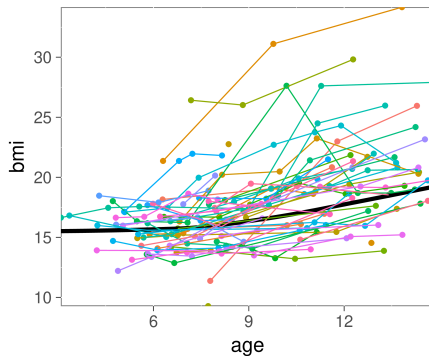
Consignes

Références

II. Particularité des données longitudinales

Problèmes des données longitudinales

Exemple



Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

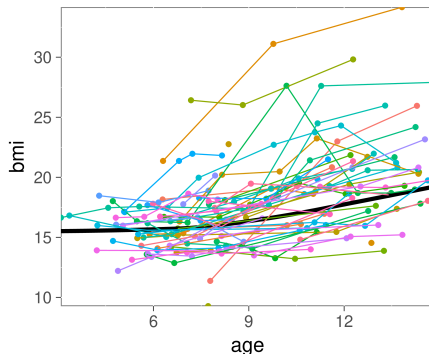
TP python

Installation
PBCseq
Consignes

Références

Problèmes des données longitudinales

Exemple

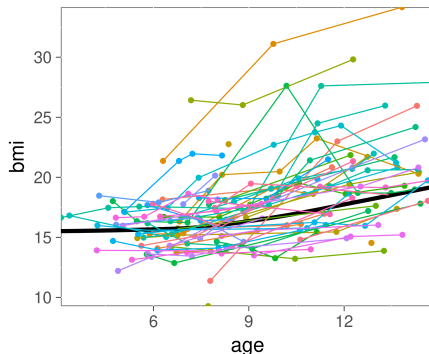


Problèmes

- Nombre de points différents par patient

Problèmes des données longitudinales

Exemple



Problèmes

- ▶ Nombre de points différents par patient
- ▶ Alignement temporel des mesures

Interpolation

- ▶ Interpoler chaque trajectoire

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Interpolation

- ▶ Interpoler chaque trajectoire
- ▶ Calcul des images sur des temps communs à tous

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Interpolation

- ▶ Interpoler chaque trajectoire
- ▶ Calcul des images sur des temps communs à tous

Extraction de features

- ▶ Afin de “représenter le signal”

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Interpolation

- ▶ Interpoler chaque trajectoire
- ▶ Calcul des images sur des temps communs à tous

Extraction de features

- ▶ Afin de “représenter le signal”
- ▶ Exemple: nombre de points, moyenne, pente, etc.

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Interpolation

- ▶ Interpoler chaque trajectoire
- ▶ Calcul des images sur des temps communs à tous

Extraction de features

- ▶ Afin de “représenter le signal”
- ▶ Exemple: nombre de points, moyenne, pente, etc.

Modèles joints

- ▶ Classe de méthodes qui modélisent de façon conjointe (vraisemblance) les données longitudinales avec la tâche prédictive

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Interpolation

- ▶ Interpoler chaque trajectoire
- ▶ Calcul des images sur des temps communs à tous

Extraction de features

- ▶ Afin de “représenter le signal”
- ▶ Exemple: nombre de points, moyenne, pente, etc.

Modèles joints

- ▶ Classe de méthodes qui modélisent de façon conjointe (vraisemblance) les données longitudinales avec la tâche prédictive
- ▶ Voir par exemple Hickey et al. [2] pour une review dans le cadre de la survie

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

IV. TP sous python

Anaconda



- ▶ Plateforme de Data Science Python/R

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

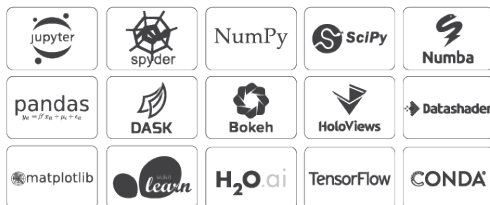
Références

Installation de l'environnement de travail

Anaconda



- Plateforme de Data Science Python/R
- Packages, librairies, outils, etc.



Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

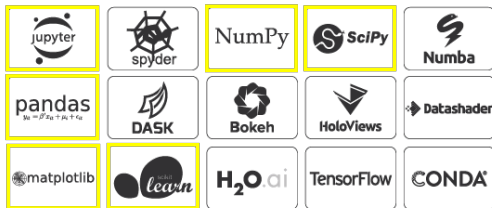
Références

Installation de l'environnement de travail

Anaconda



- Plateforme de Data Science Python/R
- Packages, librairies, outils, etc.



Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

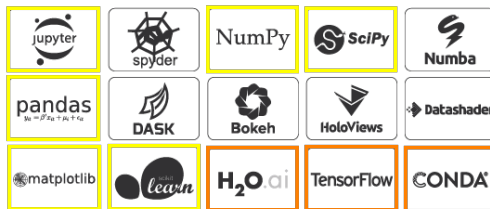
Références

Installation de l'environnement de travail

Anaconda



- Plateforme de Data Science Python/R
- Packages, librairies, outils, etc.



Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

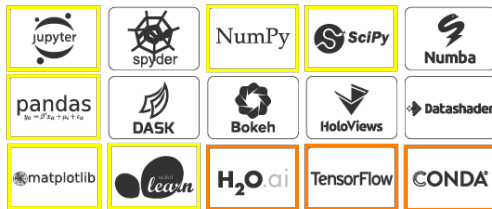
Consignes

Références

Anaconda



- ▶ Plateforme de Data Science Python/R
- ▶ Packages, librairies, outils, etc.



- ▶ <https://anaconda.com> → installer la version Python 3.7

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme

Rappel

Généralités

MRE

Notions

Données longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

► <https://openml.org/d/516> : télécharger le .csv

Les données PBCseq

- ▶ <https://openml.org/d/516> : télécharger le .csv
- ▶ Charger les données sur votre notebook jupyter

jupyter TP M2 DMS (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [5]: `!matplotlib inline
import pandas as pd
import numpy as np`

Load data

In [7]: `# reading csv file
df = pd.read_csv("./pbcseq.csv")
print("%s lines and %s columns" % df.shape)

echo 5 first lines
df.head()`

1945 lines and 19 columns

Out[7]:

	case_number	number_of_days	status	drug	age	sex	day	presence_of_asictes	pr
0	1	400	2	D- penicillamine	21464	female	no	yes	
1	1	400	2	D- penicillamine	21464	female	192	yes	
2	2	5169	0	D- penicillamine	20617	female	no	no	
3	2	5169	0	D- penicillamine	20617	female	182	no	
4	2	5169	0	D- penicillamine	20617	female	365	no	

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

- ▶ <https://openml.org/d/516> : télécharger le .csv
- ▶ Charger les données sur votre notebook jupyter

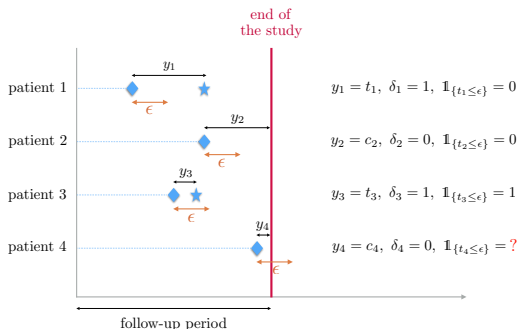
Contexte de survie

- ▶ Labels: $Y_i = \min\{T_i, C_i\}$ et $\Delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$

- <https://openml.org/d/516> : télécharger le .csv
- Charger les données sur votre notebook jupyter

Contexte de survie

- Labels: $Y_i = \min\{T_i, C_i\}$ et $\Delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$
- On se ramène à de la prédiction binaire avec le choix d'un seuil ϵ *a priori* (par exemple 1 an)



Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- Description des données

Programme

Rappel

Généralités
MRE
Notions

Données longitudinales

Problèmes
Solutions

TP python

Installation
PBCseq
Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- ▶ Description des données
- ▶ Statistiques descriptives et visualisation

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- ▶ Description des données
- ▶ Statistiques descriptives et visualisation

Extraction de features

- ▶ Utiliser par exemple `tsfresh`

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- ▶ Description des données
- ▶ Statistiques descriptives et visualisation

Extraction de features

- ▶ Utiliser par exemple `tsfresh`
- ▶ Bien d'autres méthodes sophistiquées existent.
Exemple: GP [1]

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- ▶ Description des données
- ▶ Statistiques descriptives et visualisation

Extraction de features

- ▶ Utiliser par exemple `tsfresh`
- ▶ Bien d'autres méthodes sophistiquées existent.
Exemple: GP [1]
- ▶ On se place dans un contexte de grande dimension!

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- ▶ Description des données
- ▶ Statistiques descriptives et visualisation

Extraction de features

- ▶ Utiliser par exemple `tsfresh`
- ▶ Bien d'autres méthodes sophistiquées existent.
Exemple: GP [1]
- ▶ On se place dans un contexte de grande dimension!

Régularisation

- ▶ Coder la CV pour la regression logistique pénalisée
lasso/ridge/elastic-net avec https://scikit-learn.org/stable/modules/cross_validation.html

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Consignes du TP (1/2)

Un premier pas vers le mini-mémoire

- ▶ Description des données
- ▶ Statistiques descriptives et visualisation

Extraction de features

- ▶ Utiliser par exemple `tsfresh`
- ▶ Bien d'autres méthodes sophistiquées existent.
Exemple: GP [1]
- ▶ On se place dans un contexte de grande dimension!

Régularisation

- ▶ Coder la CV pour la regression logistique pénalisée lasso/ridge/elastic-net avec https://scikit-learn.org/stable/modules/cross_validation.html
- ▶ Visualiser les courbes d'apprentissage

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

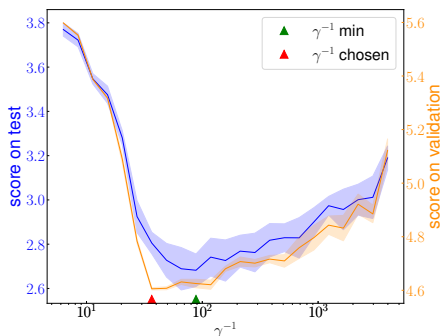
PBCseq

Consignes

Références

Régularisation

- Coder la CV pour la regression logistique pénalisée lasso/ridge/elastic-net avec https://scikit-learn.org/stable/modules/cross_validation.html
- Visualiser les courbes d'apprentissage



Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Autres algorithmes

- Utiliser par exemple les SVM, les RF, le Gradient Boosting, le deep learning (scikit-learn par ex.)

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Autres algorithmes

- ▶ Utiliser par exemple les SVM, les RF, le Gradient Boosting, le deep learning (scikit-learn par ex.)
- ▶ Comparer les tableau de contingence, F1 score, sélection de variables, interprétation, etc.

Programme

Rappel

Généralités

MRE

Notions

Données
longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Autres algorithmes

- ▶ Utiliser par exemple les SVM, les RF, le Gradient Boosting, le deep learning (scikit-learn par ex.)
- ▶ Comparer les tableau de contingence, F1 score, sélection de variables, interprétation, etc.

Pour aller plus loin

- ▶ Vérifier que les résultats dépendent du seuil (refaire tout tourner en prenant par exemple un seuil à 2 ans puis à 3 ans). Interpréter et comparer les features sélectionnées

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Autres algorithmes

- ▶ Utiliser par exemple les SVM, les RF, le Gradient Boosting, le deep learning (scikit-learn par ex.)
- ▶ Comparer les tableau de contingence, F1 score, sélection de variables, interprétation, etc.

Pour aller plus loin

- ▶ Vérifier que les résultats dépendent du seuil (refaire tout tourner en prenant par exemple un seuil à 2 ans puis à 3 ans). Interpréter et comparer les features sélectionnées
- ▶ Après le cours de survie, comparer les résultats obtenus avec les méthodes de survie

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Autres algorithmes

- ▶ Utiliser par exemple les SVM, les RF, le Gradient Boosting, le deep learning (scikit-learn par ex.)
- ▶ Comparer les tableau de contingence, F1 score, sélection de variables, interprétation, etc.

Pour aller plus loin

- ▶ Vérifier que les résultats dépendent du seuil (refaire tout tourner en prenant par exemple un seuil à 2 ans puis à 3 ans). Interpréter et comparer les features sélectionnées
- ▶ Après le cours de survie, comparer les résultats obtenus avec les méthodes de survie
- ▶ Essayer les modèles joints longitudinal-survie

Programme

Rappel

Généralités

MRE

Notions

Données

longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

Programme

Rappel

Généralités

MRE

Notions

Données longitudinales

Problèmes

Solutions

TP python

Installation

PBCseq

Consignes

Références

- [1] Simon Bussy, Raphaël Veil, Vincent Looten, Anita Burgun, Stéphane Gaïffas, Agathe Guilloux, Brigitte Ranque, and Anne-Sophie Jannot. Comparison of methods for early-readmission prediction in a high-dimensional heterogeneous covariates and time-to-event outcome framework. *BMC medical research methodology*, 19(1):50, 2019.
- [2] Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolumunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC medical research methodology*, 16(1):117, 2016.
- [3] Vladimir Vapnik. Principles of risk minimization for learning theory. In *NIPS*, pages 831–838, 1991.
- [4] Vladimir Vapnik. *Statistical learning theory*. 1998, volume 3. Wiley, New York, 1998.