

Stage de Recherche Master 2 MVA de l'ENS Cachan
Nouvelles techniques de ML pour la médecine :

**FACTEURS PRÉDICTIFS DE LA
RÉHOSPITALISATION PRÉCOCE DE
DRÉPANOCYTAIRES ADULTES POUR CRISE
VASO-OCCLUSIVE**

Simon Bussy

encadré par

Anne-Sophie Jannot, Stéphane Gaiffas, Agathe Guilloux

Présentation INSERM

5 octobre 2015

Introduction

- Contexte : CMAP et équipe INSERM n° 22 de l'HEGP ; financé par l'Institut Europlace de Finance.

Introduction

- Contexte : CMAP et équipe INSERM n° 22 de l'HEGP ; financé par l'Institut Europlace de Finance.
- HEGP : centre de référence pour le traitement des CVO

Introduction

- Contexte : CMAP et équipe INSERM n° 22 de l'HEGP ; financé par l'Institut Europlace de Finance.
- HEGP : centre de référence pour le traitement des CVO
- Taux de réhospitalisation élevé ($\sim 20\%$)

Introduction

- Contexte : CMAP et équipe INSERM n° 22 de l'HEGP ; financé par l'Institut Europlace de Finance.
- HEGP : centre de référence pour le traitement des CVO
- Taux de réhospitalisation élevé ($\sim 20\%$)
- But : construire un modèle prédictif des rechutes

- 1 Modèles
- 2 Algorithmes
- 3 Simulations
- 4 Organisation
- 5 Présentation
- 6 Preprocessing
- 7 Résultats
- 8 Conclusion

Modélisation

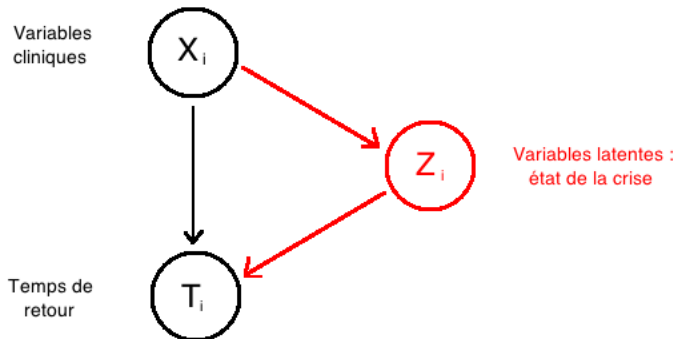
Variables
cliniques



Temps de
retour



Modélisation



Modèle de mélange

- Echantillon $(X_1, T_1), \dots, (X_n, T_n) \in \mathbb{R}^d \times \mathbb{N}^*$ *i.i.d.* $\sim \mathbb{P}$

Modèle de mélange

- Echantillon $(X_1, T_1), \dots, (X_n, T_n) \in \mathbb{R}^d \times \mathbb{N}^*$ *i.i.d.* $\sim \mathbb{P}$
- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$

Modèle de mélange

- Echantillon $(X_1, T_1), \dots, (X_n, T_n) \in \mathbb{R}^d \times \mathbb{N}^*$ *i.i.d.* $\sim \mathbb{P}$
- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$
- $\pi_0(X_i) = \mathbb{P}(Z_i = 0 | X_i) = \frac{1}{1 + e^{-X_i^\top \beta}}$

Modèle de mélange

- Echantillon $(X_1, T_1), \dots, (X_n, T_n) \in \mathbb{R}^d \times \mathbb{N}^*$ *i.i.d.* $\sim \mathbb{P}$
- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$
- $\pi_0(X_i) = \mathbb{P}(Z_i = 0 | X_i) = \frac{1}{1 + e^{-X_i^\top \beta}}$
- $\forall i \in \llbracket 1, n \rrbracket, Z_i \sim \mathcal{B}(1 - \pi_0(X_i))$

Modèle de mélange

- Echantillon $(X_1, T_1), \dots, (X_n, T_n) \in \mathbb{R}^d \times \mathbb{N}^*$ *i.i.d.* $\sim \mathbb{P}$
- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$
- $\pi_0(X_i) = \mathbb{P}(Z_i = 0 | X_i) = \frac{1}{1 + e^{-X_i^\top \beta}}$
- $\forall i \in \llbracket 1, n \rrbracket, Z_i \sim \mathcal{B}(1 - \pi_0(X_i))$
- $\forall i \in \llbracket 1, n \rrbracket, \forall t \in \mathbb{N}^*, f_{T_i | X_i}(t) = \pi_0(X_i) f_{T_i | X_i, 0}(t) + (1 - \pi_0(X_i)) f_{T_i | X_i, 1}(t)$

Modèle de mélange

- Echantillon $(X_1, T_1), \dots, (X_n, T_n) \in \mathbb{R}^d \times \mathbb{N}^*$ *i.i.d.* $\sim \mathbb{P}$
- $Z_i \in \{0, 1\}$ v.a. latente t.q. $\forall i \in \llbracket 1, n \rrbracket, \forall k \in \{0, 1\}, T_i | Z_i = k \sim \ell_k$
- $\pi_0(X_i) = \mathbb{P}(Z_i = 0 | X_i) = \frac{1}{1 + e^{-X_i^\top \beta}}$
- $\forall i \in \llbracket 1, n \rrbracket, Z_i \sim \mathcal{B}(1 - \pi_0(X_i))$
- $\forall i \in \llbracket 1, n \rrbracket, \forall t \in \mathbb{N}^*, f_{T_i | X_i}(t) = \pi_0(X_i) f_{T_i | X_i, 0}(t) + (1 - \pi_0(X_i)) f_{T_i | X_i, 1}(t)$
- $\ell_k \sim \mathcal{G}(p_k)$ de densité $\forall t \in \mathbb{N}^*, f_k(t) = p_k(1 - p_k)^{t-1}$

Inférence du modèle

- $\theta = \{p_0, p_1, \beta\} \in \mathbb{R}^{d+3}$

Inférence du modèle

- $\theta = \{p_0, p_1, \beta\} \in \mathbb{R}^{d+3}$

- $\ell_n(\theta) = \frac{1}{n} \log \prod_{i=1}^n \mathbb{P}_\theta(T_i | X_i) = \frac{1}{n} \sum_{i=1}^n \log \left[\frac{p_0(1-p_0)^{T_i-1}}{1+e^{-X_i^\top \beta_0}} + \frac{p_1(1-p_1)^{T_i-1}}{1+e^{X_i^\top \beta_0}} \right]$

Inférence du modèle

- $\theta = \{p_0, p_1, \beta\} \in \mathbb{R}^{d+3}$

- $\ell_n(\theta) = \frac{1}{n} \log \prod_{i=1}^n \mathbb{P}_\theta(T_i | X_i) = \frac{1}{n} \sum_{i=1}^n \log \left[\frac{p_0(1-p_0)^{T_i-1}}{1+e^{-X_i^\top \beta_0}} + \frac{p_1(1-p_1)^{T_i-1}}{1+e^{X_i^\top \beta_0}} \right]$

- $\hat{\theta} \in \underset{(p_0, p_1, \beta) \in]0,1[^2 \times \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^n \log \left[\frac{p_0(1-p_0)^{T_i-1}}{1+e^{-X_i^\top \beta_0}} + \frac{p_1(1-p_1)^{T_i-1}}{1+e^{X_i^\top \beta_0}} \right] + \lambda_1 \|\beta\|_1 + \frac{\lambda_2}{2} \|\beta\|_2^2 \right\}$

Régression logistique

- $Y_i = \mathbb{1}_{\{T_i \leq s\}}$; seuil s de 15 jours

Régression logistique

- $Y_i = \mathbb{1}_{\{T_i \leq s\}}$; seuil s de 15 jours
- $\ell_n(\beta) = \sum_{i=1}^n Y_i \log(g(X_i^\top \beta)) + (1 - Y_i) \log(1 - g(X_i^\top \beta))$

Régression logistique

- $Y_i = \mathbb{1}_{\{T_i \leq s\}}$; seuil s de 15 jours
- $\ell_n(\beta) = \sum_{i=1}^n Y_i \log(g(X_i^\top \beta)) + (1 - Y_i) \log(1 - g(X_i^\top \beta))$
- $\hat{\beta}_{ElasticNet} \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ - \sum_{i=1}^n Y_i \log(g(X_i^\top \beta)) + (1 - Y_i) \log(1 - g(X_i^\top \beta)) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}$

Régression logistique

- $Y_i = \mathbb{1}_{\{T_i \leq s\}}$; seuil s de 15 jours
- $\ell_n(\beta) = \sum_{i=1}^n Y_i \log(g(X_i^\top \beta)) + (1 - Y_i) \log(1 - g(X_i^\top \beta))$
- $\hat{\beta}_{ElasticNet} \in \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ - \sum_{i=1}^n Y_i \log(g(X_i^\top \beta)) + (1 - Y_i) \log(1 - g(X_i^\top \beta)) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}$
- Score : AUC

$$ROC : \mathbb{R} \rightarrow [0, 1]^2$$

$$t \mapsto \left(\mathbb{P}(s(X) \geq t | Y = 0), \mathbb{P}(s(X) \geq t | Y = 1) \right)$$

Espérance-Maximisation

- $$\ell_n^c(\theta, \mathcal{T}, \mathcal{Z}) = \frac{1}{n} \sum_{i=1}^n \log \left[Z_i \left(\log(1 - \pi_0(X_i)) + \log p_1 + (T_i - 1) \log(1 - p_1) \right) \right. \\ \left. + (1 - Z_i) \left(\log(\pi_0(X_i)) + \log p_0 + (T_i - 1) \log(1 - p_0) \right) \right]$$

Espérance-Maximisation

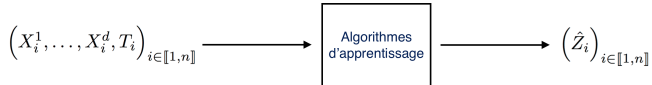
- $\ell_n^c(\theta, \mathcal{T}, \mathcal{Z}) = \frac{1}{n} \sum_{i=1}^n \log \left[Z_i \left(\log(1 - \pi_0(X_i)) + \log p_1 + (T_i - 1) \log(1 - p_1) \right) + (1 - Z_i) \left(\log(\pi_0(X_i)) + \log p_0 + (T_i - 1) \log(1 - p_0) \right) \right]$
- Etape E :
 - $\mathcal{Q}_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}} [\ell_n^c(\theta, \mathcal{T}, \mathcal{Z}) | \mathcal{T}]$
 - Remplacer Z_i par $q_i^{(l)} = \mathbb{E}_{\theta^{(l)}} [Z_i | T_i] = \frac{p_1^{(l)} (1 - p_1^{(l)})^{T_i - 1} (1 - \pi_0(X_i))}{p_0^{(l)} (1 - p_0^{(l)})^{T_i - 1} \pi_0(X_i) + p_1^{(l)} (1 - p_1^{(l)})^{T_i - 1} (1 - \pi_0(X_i))}$

Espérance-Maximisation

- $\ell_n^c(\theta, \mathcal{T}, \mathcal{Z}) = \frac{1}{n} \sum_{i=1}^n \log \left[Z_i \left(\log(1 - \pi_0(X_i)) + \log p_1 + (T_i - 1) \log(1 - p_1) \right) + (1 - Z_i) \left(\log(\pi_0(X_i)) + \log p_0 + (T_i - 1) \log(1 - p_0) \right) \right]$
- Etape E :
 - $\mathcal{Q}_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}} [\ell_n^c(\theta, \mathcal{T}, \mathcal{Z}) | \mathcal{T}]$
 - Remplacer Z_i par $q_i^{(l)} = \mathbb{E}_{\theta^{(l)}} [Z_i | T_i] = \frac{p_1^{(l)} (1 - p_1^{(l)})^{T_i - 1} (1 - \pi_0(X_i))}{p_0^{(l)} (1 - p_0^{(l)})^{T_i - 1} \pi_0(X_i) + p_1^{(l)} (1 - p_1^{(l)})^{T_i - 1} (1 - \pi_0(X_i))}$
- Etape M :
 - $\theta^{(l+1)} \in \arg \max_{\theta} \mathcal{Q}_n(\theta, \theta^{(l)})$
 - $p_0^{(l+1)} = \frac{n - \sum_{i=1}^n q_i^{(l)}}{\sum_{i=1}^n T_i (1 - q_i^{(l)})}$; $p_1^{(l+1)} = \frac{\sum_{i=1}^n q_i^{(l)}}{\sum_{i=1}^n T_i q_i^{(l)}}$
 - $\beta^{(l+1)} \in \arg \min_{\beta} \left\{ -\mathcal{Q}_n(\theta, \theta^{(l)}) + \lambda_1 \|\beta\|_1 + \frac{\lambda_2}{2} \|\beta\|_2^2 \right\}$ avec L-BGFS-B

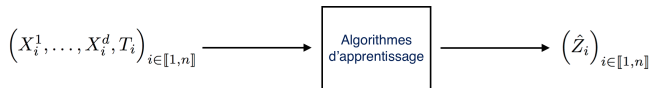
Simulations

- Idée : tester le schéma de prédiction suivant :



Simulations

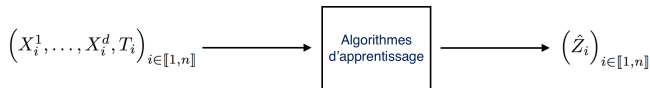
- Idée : tester le schéma de prédiction suivant :



- Pour la régression logistique : $Z_i \sim Y_i = \mathbb{1}_{\{T_i \leq s\}}$, AUC sur les $\hat{Y}_i$

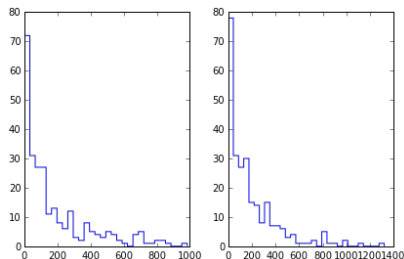
Simulations

- Idée : tester le schéma de prédiction suivant :



- Pour la régression logistique : $Z_i \sim Y_i = \mathbb{1}_{\{T_i \leq s\}}$, AUC sur les $\hat{Y}_i$
- Pour l'EM : AUC sur les $\hat{Z}_i$

Estimation des vrais paramètres



Comparaison des histogrammes de temps de retour réels et simulés après estimation

Paramètres	ρ_0	ρ_1	π
$\hat{\theta}$	0,005	0,511	0,872

Résultats de prédiction sur les vraies données

Génération des données

- On génère $Z_i \sim \mathcal{B}(1 - \pi_0)$ pour $n = 1000$ séjours avec $\pi_0 = 0,8721$

Génération des données

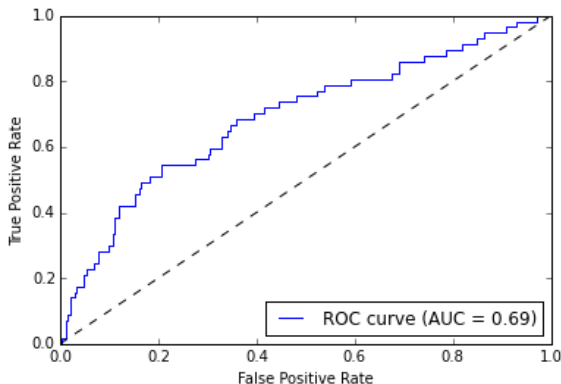
- On génère $Z_i \sim \mathcal{B}(1 - \pi_0)$ pour $n = 1000$ séjours avec $\pi_0 = 0,8721$
- Puis $T_i \sim \mathcal{G}(p_{Z_i})$ avec $p_0 = 0,005$ et $p_1 = 0,511$

Génération des données

- On génère $Z_i \sim \mathcal{B}(1 - \pi_0)$ pour $n = 1000$ séjours avec $\pi_0 = 0,8721$
- Puis $T_i \sim \mathcal{G}(p_{Z_i})$ avec $p_0 = 0,005$ et $p_1 = 0,511$
- Et enfin les données d'apprentissage :

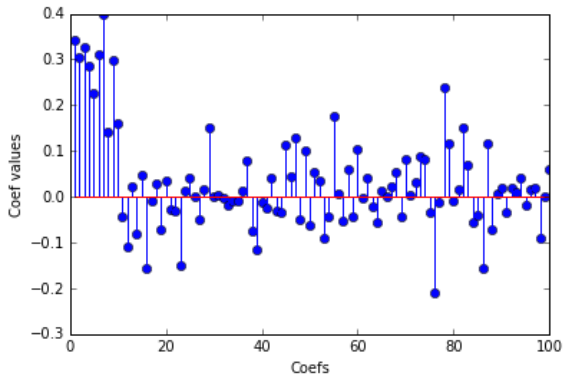
$$\forall i \in \llbracket 1, n \rrbracket, \begin{cases} \forall j \in \llbracket 1, a \rrbracket, \begin{cases} X_i^j | Z_i = 0 \sim \mathcal{N}(0, 1) \\ X_i^j | Z_i = 1 \sim \mathcal{N}(0.5, 1) \end{cases} \\ \forall j \in \llbracket a + 1, d \rrbracket, X_i^j \sim \mathcal{N}(0, 1) \end{cases}$$

Résultats : régression logistique



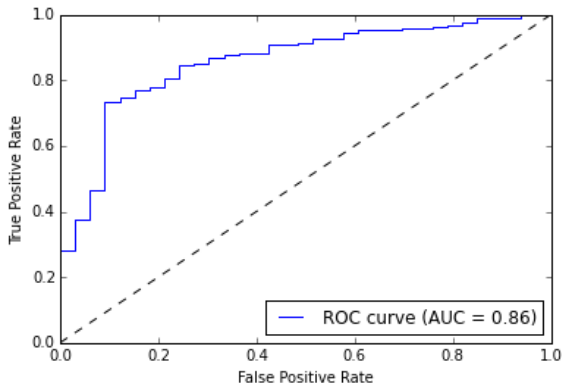
Courbe ROC pour la prédiction de la régression logistique

Résultats : régression logistique



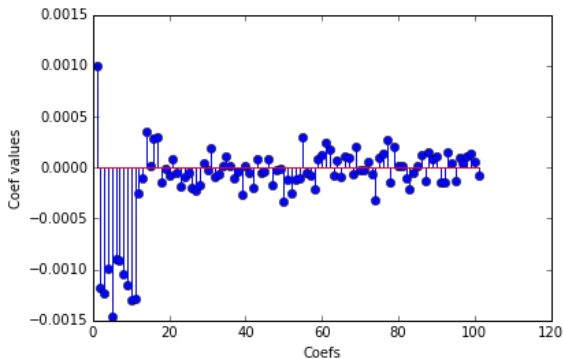
$\hat{\beta}$ appris par la régression logistique

Résultats : EM



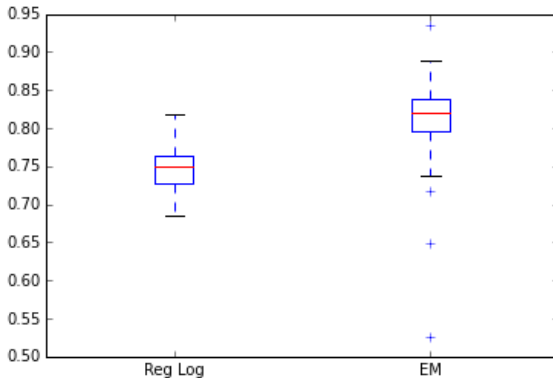
Courbe ROC pour la prédiction de l'EM

Résultats : EM



$\hat{\beta}$ appris par l'EM

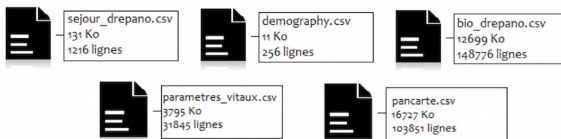
Comparaison des deux méthodes



Comparaison des AUC des 2 méthodes pour 100 tests

Organisation des données

- Données hétérogènes, différentes sources

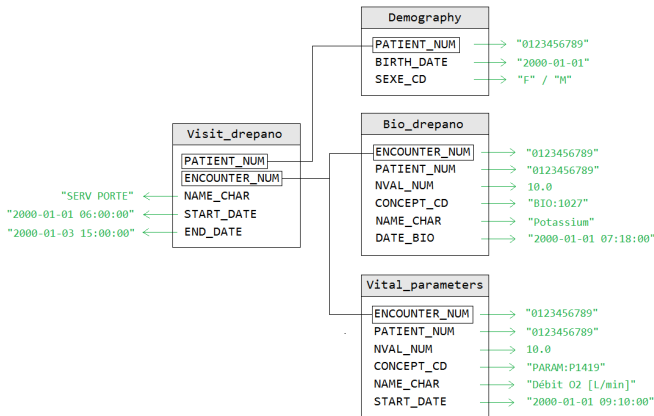


Organisation des données

- Données hétérogènes, différentes sources
- 1 séjour / patient : "choix aléatoire par classe", taux de rechute de 12,84%

Organisation des données

● Réorganisation des données



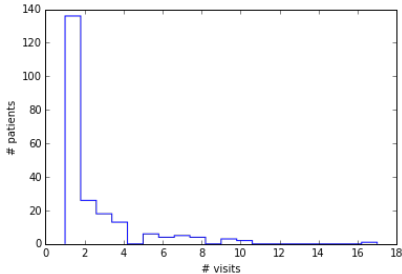
Organisation des données

● Création d'un fichier JSON

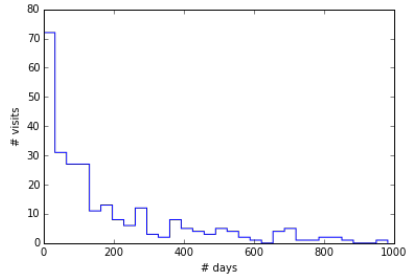
```
Patients = {
  '1' = {
    'patient_num' = '0123456789',
    'sex' = 'F/M',
    'birth_date' = '2000-01-01',
    'visits' = {
      '1' = {
        'encounter_num' = '0123456789',
        'age' = '25',
        'rea' = '0/1',
        'duration' = '13',
        'previous_visit' = 'none/10',
        'next_visit' = 'none/15',
        'details' = {
          '1' = {
            'service' = 'SERV PORTE',
            'start_date' = '2000-01-01 05:00:00',
            'end_date' = '2000-05-01 10:00:00'
          },
          '2' = {...}
        },
        'bio-infos' = {
          '1' = {
            'name_char' = 'Potassium',
            'val' = {
              '1' = {
                'nval_num' = '10.0',
                'concept_cd' = '810:1027',
                'date_bio' = '2000-01-01 11:00:00'
              },
              '2' = {...}
            }
          },
          '2' = {...}
        },
        'vital_parameters' = {
          '1' = {
            'name_char' = 'débit o2 [L/min]',
            'val' = {
              '1' = {
                'nval_num' = '10.0',
                'concept_cd' = 'PARAM:P1419',
                'date_bio' = '2000-01-01 10:30:00'
              },
              '2' = {...}
            }
          },
          '2' = {...}
        }
      },
      '2' = {...}
    },
    '2' = {...}
  },
  '2' = {...}
}
```

Présentation des données

- 218 patients, 479 visites



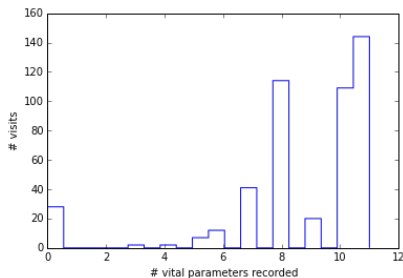
Nombre de visites par patient



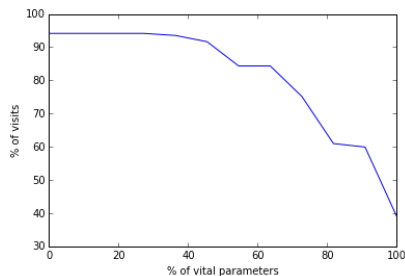
Temps inter-visites

Les paramètres vitaux

● Les paramètres vitaux



Nombre de points
par paramètre vital



Variables biologiques en commun
à tous les séjours

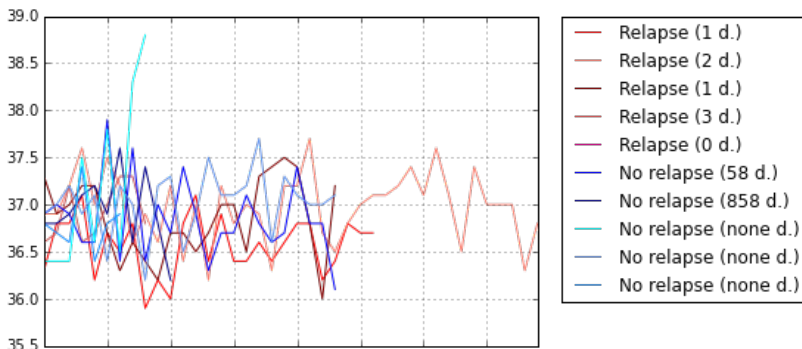
Les paramètres vitaux

- Les paramètres vitaux
- 9 variables gardées

Pression artérielle systolique [mmHg], notée PA max ;
pression artérielle diastolique [mmHg], notée PA min ;
saturation en oxygène [%] ; Douleur EVA [U] ; fréquence
respiratoire [mvt/min] ; température [°C] ; poids [kg] ;
oxygène [L/min] ; fréquence cardiaque [bpm]

Les paramètres vitaux

- Les paramètres vitaux
- 9 variables gardées
- Problème : durées $\neq$ et alignement



Evolution de la température au cours de 10 séjours

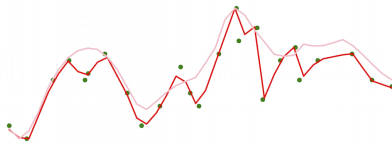
Les paramètres vitaux

- Rescaling entre $[0,1]$

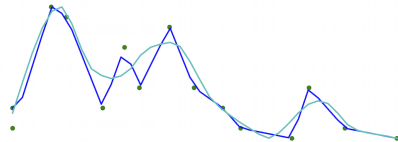
$$\forall j \in \llbracket 1, N \rrbracket, x_j = \frac{x_j - \min_k(x_k)}{\max_k(x_k) - \min_k(x_k)}$$

Les paramètres vitaux

- Rescaling entre $[0,1]$
- Interpolation : p valeurs
Puis filtre de Savitzky–Golay : p' valeurs de gradient



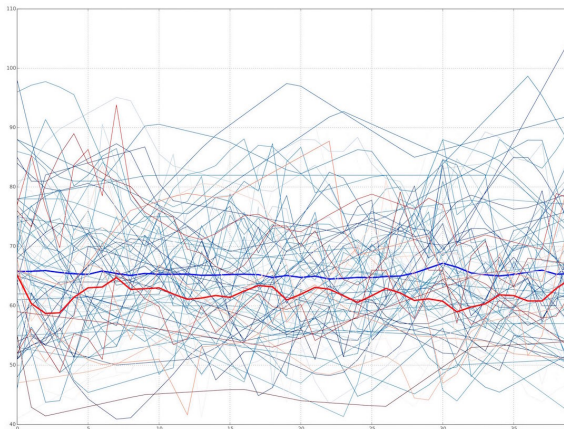
Fréquence respiratoire pour
un patient positif



Fréquence respiratoire pour
un patient négatif

Les paramètres vitaux

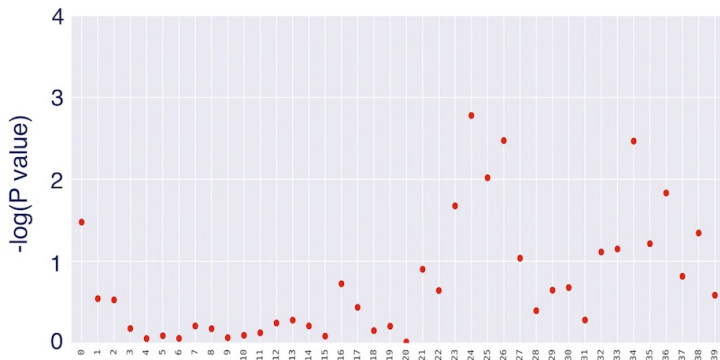
- Comportements différents en moyenne



Pression artérielle diastolique et moyennes par classe

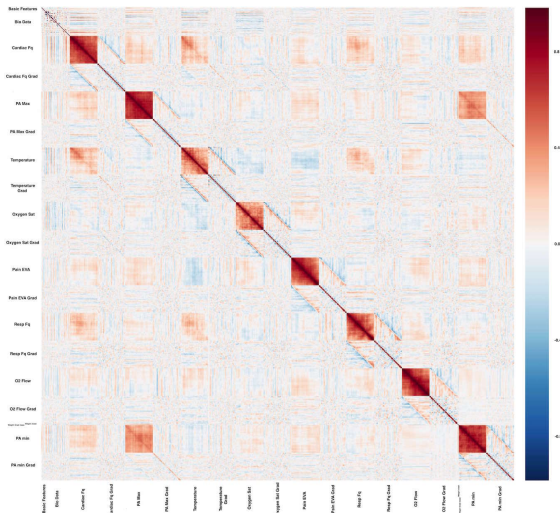
Les paramètres vitaux

- Comportements différents en moyenne
- Test de Mann-Whitney-Wilcoxon



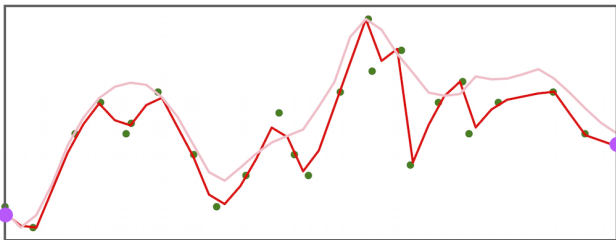
Test de Mann-Whitney-Wilcoxon pour la température

Preprocessing



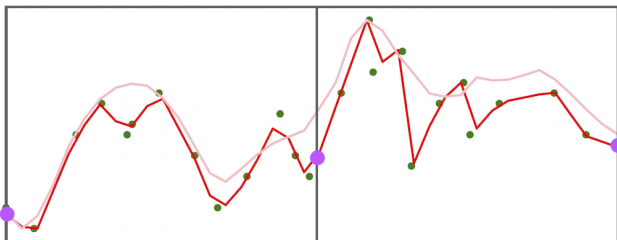
Sélection de p et p'

Procédure suivie :



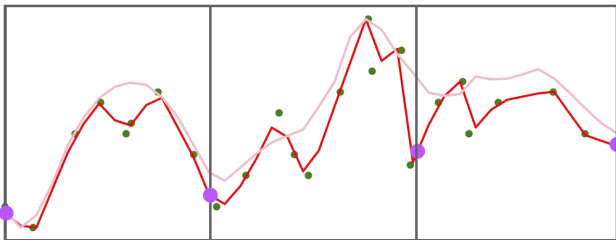
Sélection de p et p'

Procédure suivie :



Sélection de p et p'

Procédure suivie :



Sélection de p et p'

- Résultats après 50 itérations

Concept	Nombre de points	Ecart type
Fq cardiaque	28	0,3
▽ Fq cardiaque	26	0
PA max	28	0
▽ PA max	26	0
Température	30	8,4
▽ Température	18	0
Saturation O_2	34	4,8
▽ Saturation O_2	26	0
Douleur EVA	21	0,7
▽ Douleur EVA	26	0
Débit O_2	16	4,6
▽ Débit O_2	26	0
Fq respiratoire	21	5,3
▽ Fq respiratoire	26	0
PA min	28	0
▽ PA min	26	0

Sélection de p et p'

- Résultats après 50 itérations
- Dimension de l'espace : 446

Sélection de p et p'

- Résultats après 50 itérations
- Dimension de l'espace : 446
- Données manquantes remplacées par les moyennes

Benjamini-Hochberg

- Tests de Mann-Whitney-Wilcoxon : 446 *p-values*

Benjamini-Hochberg

- Tests de Mann-Whitney-Wilcoxon : 446 *p-values*
- Procédure de Benjamini-Hochberg

On conserve k features avec

$$k = \operatorname{argmax} \left\{ j : p(j) \leq \frac{\alpha j}{K} \right\}$$

Benjamini-Hochberg

- Tests de Mann-Whitney-Wilcoxon : 446 *p-values*
- Procédure de Benjamini-Hochberg
- $\alpha = 0,9$; 50 itérations

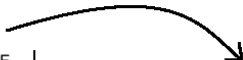
Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$

...	F_j	...
	⋮	
	12,5	
	9,2	
	3,1	
	8,7	
	⋮	

Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$

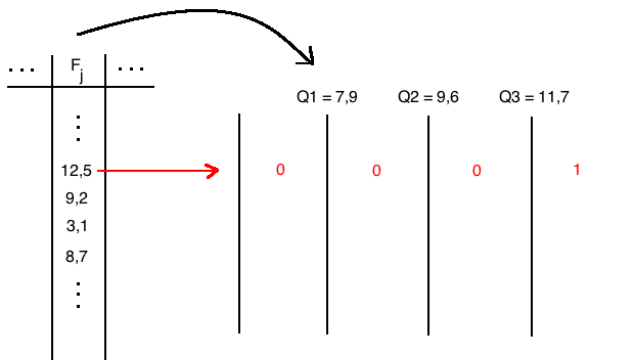


...	F_j	...
	⋮	
	12,5	
	9,2	
	3,1	
	8,7	
	⋮	

Q1 = 7,9 Q2 = 9,6 Q3 = 11,7

Binarisation

- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$



Binarisation

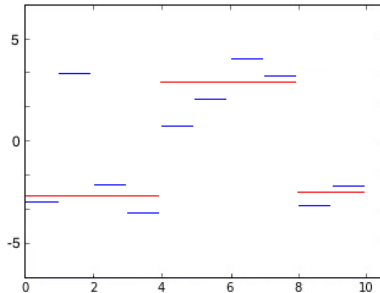
- Binarisation par quartiles ; $dim = 156 \sim 4 \times 45$

...	F_j	...		$Q1 = 7,9$	$Q2 = 9,6$	$Q3 = 11,7$	
	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
	12,5	$\rightarrow$	0	0	0	1	
	9,2	$\rightarrow$	0	1	0	0	
	3,1	$\rightarrow$	1	0	0	0	
	8,7	$\rightarrow$	0	1	0	0	
	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Binarisation

• Alternative : screening TV

$$\hat{\beta}_{TV}^k \in \underset{\beta \in \mathbb{R}^{10}}{\operatorname{argmin}} \left\{ -\sum_{i=1}^n Y_i \log(g((D_i^k)^\top \beta)) + (1 - Y_i) \log(1 - g((D_i^k)^\top \beta)) + \lambda \sum_{j=2}^{10} |\beta_j - \beta_{j-1}| \right\}$$



Association + décorrélation

- Ajout des produits cartésiens

Association + décorrélation

- Ajout des produits cartésiens
- $\dim = 11978 \sim \binom{156}{2}$

Association + décorrélation

- Ajout des produits cartésiens
- $\dim = 11978 \sim \binom{156}{2}$
- Décorrélation $|\rho| < 0,95$

Association + décorrélation

- Ajout des produits cartésiens
- $\dim = 11978 \sim \binom{156}{2}$
- Décorrélation $|\rho| < 0,95$
- Finalement, $\dim = 10574$

Association + décorrélation

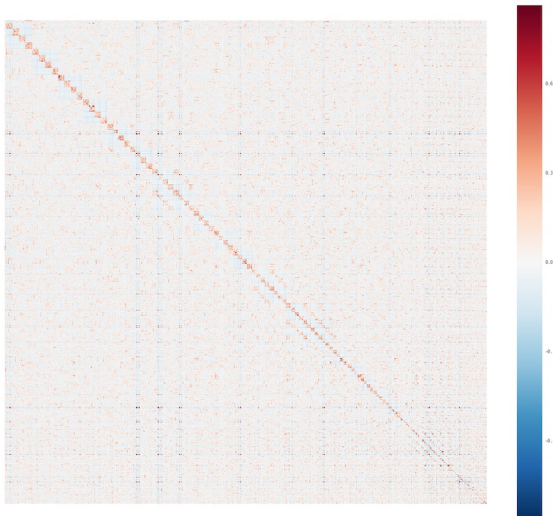
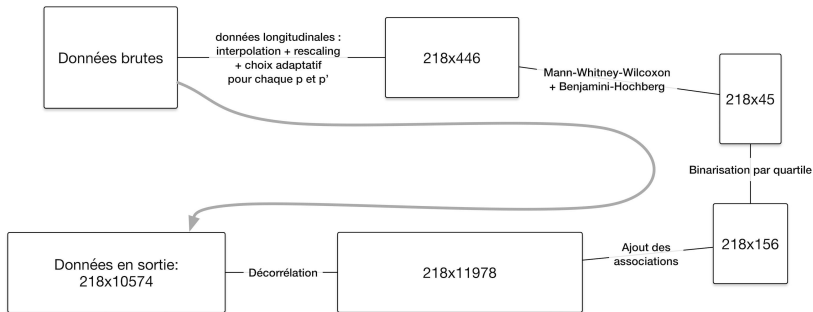
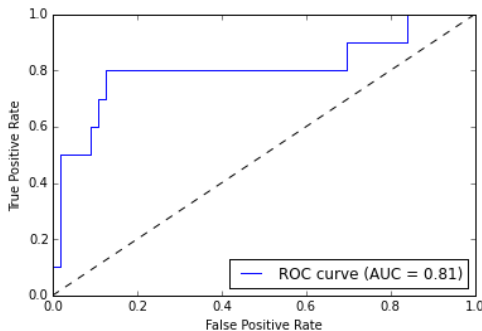


Schéma récapitulatif



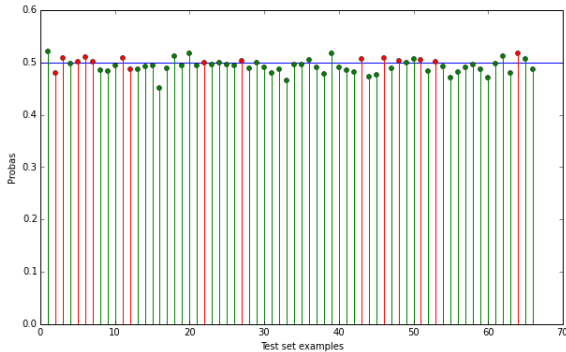
Résultats : régression ridge (dim = 10546)



Courbe ROC pour la régression ℓ_2 dans l'espace de dimension 10546

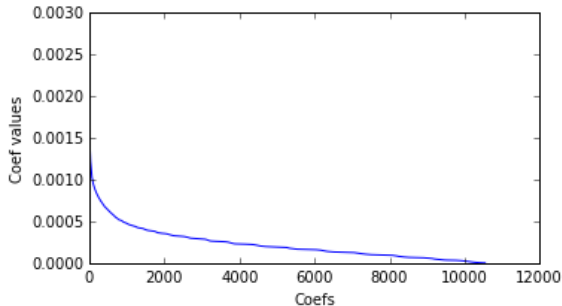
Classe	Précision	Recall	Support
0	0,96	0,77	56
1	0,38	0,80	10
Moyenne/Total	0,87	0,77	66

Résultats : régression ridge (dim = 10546)



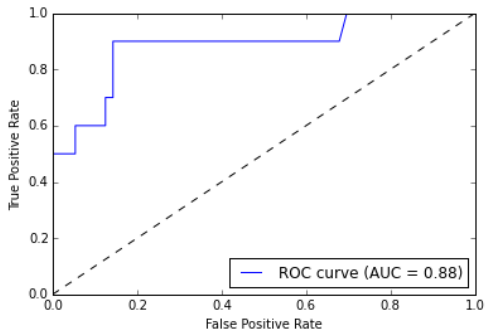
Probabilités prédites sur le jeu de test

Résultats : régression ridge (dim = 10546)



Coefficients de $|\hat{\beta}_{ridge}|$ triés

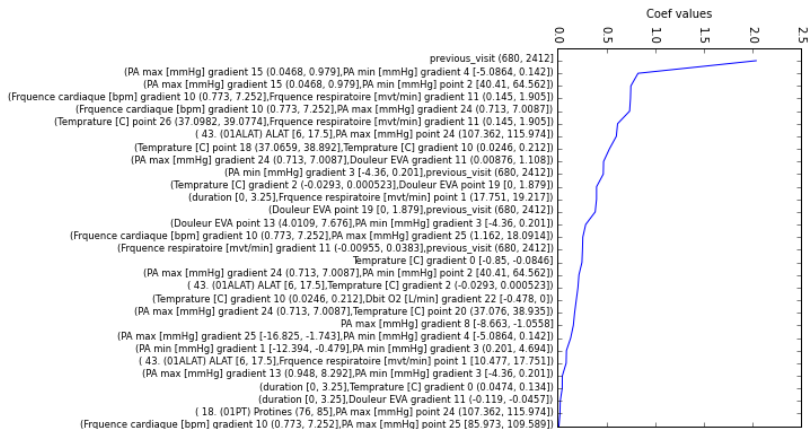
Résultats : régression lasso (dim = 10546)



Courbe ROC pour la régression ℓ_1 dans l'espace de dimension 10546

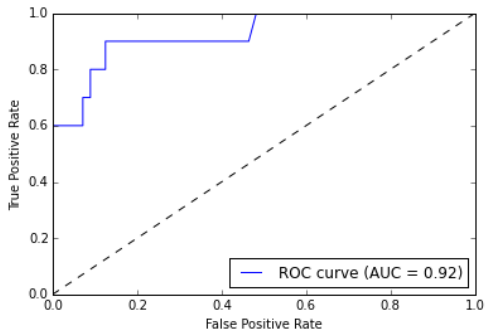
Classe	Précision	Recall	Support
0	0,98	0,82	56
1	0,47	0,90	10
Moyenne/Total	0,90	0,83	66

Résultats : Support du lasso



Coefficients de $|\hat{\beta}_{lasso}|$ non nuls triés

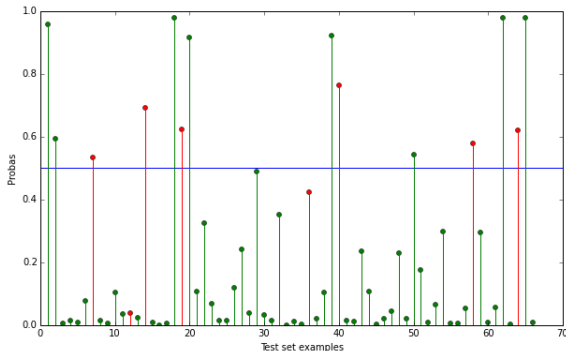
Résultats : régression ridge (dim = 30)



Courbe ROC pour la régression ℓ_2 dans l'espace de dimension 30

Classe	Précision	Recall	Support
0	0,96	0,89	56
1	0,57	0,80	10
Moyenne/Total	0,90	0,88	66

Résultats : régression ridge (dim = 30)



Probabilités prédites sur le jeu de test

Modèle censuré pour l'EM

- Echantillon $(X_1, (T_1, \delta_1)), \dots, (X_n, (T_n, \delta_n)) \in \mathbb{R}^d \times (\mathbb{N}^* \times \{0, 1\})$

Modèle censuré pour l'EM

- Echantillon $(X_1, (T_1, \delta_1)), \dots, (X_n, (T_n, \delta_n)) \in \mathbb{R}^d \times (\mathbb{N}^* \times \{0, 1\})$
- $T_i = R_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{R_i \leq C_i\}}$

Modèle censuré pour l'EM

- Echantillon $(X_1, (T_1, \delta_1)), \dots, (X_n, (T_n, \delta_n)) \in \mathbb{R}^d \times (\mathbb{N}^* \times \{0, 1\})$
- $T_i = R_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{R_i \leq C_i\}}$
- Hyp 1 : $\forall i \in \llbracket 1, n \rrbracket, R_i \perp\!\!\!\perp C_i | Z_i, X_i$

Modèle censuré pour l'EM

- Echantillon $(X_1, (T_1, \delta_1)), \dots, (X_n, (T_n, \delta_n)) \in \mathbb{R}^d \times (\mathbb{N}^* \times \{0, 1\})$
- $T_i = R_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{R_i \leq C_i\}}$
- *Hyp 1* : $\forall i \in \llbracket 1, n \rrbracket, R_i \perp\!\!\!\perp C_i | Z_i, X_i$
- *Hyp 2* : $\forall i \in \llbracket 1, n \rrbracket, C_i \perp\!\!\!\perp Z_i, X_i$

Modèle censuré pour l'EM

- Echantillon $(X_1, (T_1, \delta_1)), \dots, (X_n, (T_n, \delta_n)) \in \mathbb{R}^d \times (\mathbb{N}^* \times \{0, 1\})$
- $T_i = R_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{R_i \leq C_i\}}$
- Hyp 1 : $\forall i \in \llbracket 1, n \rrbracket, R_i \perp\!\!\!\perp C_i | Z_i, X_i$
- Hyp 2 : $\forall i \in \llbracket 1, n \rrbracket, C_i \perp\!\!\!\perp Z_i, X_i$
- $\ell_n^c(\theta, \mathcal{T}, \Delta, \mathcal{Z}) = \frac{1}{n} \sum_{i=1}^n \delta_i \left\{ Z_i [\log(1 - \pi_0(X_i)) + \log(p_1(T_i))] + (1 - Z_i) [\log(\pi_0(X_i)) + \log(p_0(T_i))] + \log(\overline{G}(T_i^-)) \right\} + (1 - \delta_i) \left\{ Z_i [\log(1 - \pi_0(X_i)) + \log(\overline{F}_1(T_i^-))] + (1 - Z_i) [\log(\pi_0(X_i)) + \log(\overline{F}_0(T_i^-))] + \log(g(T_i)) \right\}$

Modèle censuré pour l'EM

- Echantillon $(X_1, (T_1, \delta_1)), \dots, (X_n, (T_n, \delta_n)) \in \mathbb{R}^d \times (\mathbb{N}^* \times \{0, 1\})$
- $T_i = R_i \wedge C_i$ et $\delta_i = \mathbb{1}_{\{R_i \leq C_i\}}$
- Hyp 1 : $\forall i \in \llbracket 1, n \rrbracket, R_i \perp\!\!\!\perp C_i | Z_i, X_i$
- Hyp 2 : $\forall i \in \llbracket 1, n \rrbracket, C_i \perp\!\!\!\perp Z_i, X_i$
- $\ell_n^c(\theta, \mathcal{T}, \Delta, \mathcal{Z}) = \frac{1}{n} \sum_{i=1}^n \delta_i \left\{ Z_i [\log(1 - \pi_0(X_i)) + \log(p_1(T_i))] + (1 - Z_i) [\log(\pi_0(X_i)) + \log(p_0(T_i))] + \log(\bar{G}(T_i^-)) \right\} + (1 - \delta_i) \left\{ Z_i [\log(1 - \pi_0(X_i)) + \log(\bar{F}_1(T_i^-))] + (1 - Z_i) [\log(\pi_0(X_i)) + \log(\bar{F}_0(T_i^-))] + \log(g(T_i)) \right\}$
- Implémentation à terminer et à tester !

Conclusion

- De nombreuses pistes restent à explorer mais premiers résultats encourageants !

Conclusion

- De nombreuses pistes restent à explorer mais premiers résultats encourageants !
- Tester sur une cohorte extérieure

Conclusion

- De nombreuses pistes restent à explorer mais premiers résultats encourageants !
- Tester sur une cohorte extérieure
- Progrès techniques et culturels (ICML, SMILE)

Conclusion

- De nombreuses pistes restent à explorer mais premiers résultats encourageants !
- Tester sur une cohorte extérieure
- Progrès techniques et culturels (ICML, SMILE)
- Thèse à venir orientée sur les questions posées lors de ce stage : modélisation de variable longitudinales, adaptation des algorithmes d'apprentissage classiques à ces données, problèmes d'alignements, etc.