PRAMA

Syllabus

1. Présentation générale

Les méthodes d'apprentissage statistique, ou machine learning, permettent de résoudre opérationnellement nombre de problématiques de prévision ; elles sont parfois adossées à d'autres modèles (physiques, économiques, etc.) et de l'expertise métier.

Le cours PRAMA (*PRAtique des Méthodes d'Apprentissage*) vise à doter les étudiants des connaissances de base sur l'apprentissage et leur mise en pratique.

Objectifs pédagogiques :

A la fin du module, les étudiants seront capables de :

- Comprendre les grands enjeux de l'apprentissage (coût et risque, biais-variance, etc.).
- Savoir mettre en œuvre les méthodes suivantes :
    - Régression régularisée
    - GAM
    - CART
    - Bagging (dont random forest)
    - Gradient boosting
    - SVM
    - Réseaux de neurones

Les méthodes rencontrées seront mises en œuvre sur différents jeux de données, à l'aide des logiciels R ou Python.

Choix du format pédagogique

Dans l'idée d'accroître l'autonomie, le cours PRAMA a adopté un format hybride :

- Des documents de cours (vidéos, polycopiés et exemples) à consulter en autonomie.
- Des QCM d'auto-évaluations à distance permettant de percevoir sa compréhension des concepts lus et vus.
- Une mise en application directe du cours sur des données issues d'un challenge.
- Un projet sur un jeu de données métier pour mettre en pratique les concepts et méthodes.

2. Modalités pédagogiques

Programme

- Séquence 1 : Apprentissage statistique
- Séquence 2 : Régression régularisée
- Séquence 3 : Noyaux de lissage & plus proches voisins
- Séquence 4 : Régression spline & modèle GAM
- Séquence 5 : Arbres de régression et de décision
- Séquence 6 : Bagging
- Séquence 7 : Boosting
- Séquence 8 : SVM
- Séquence 9 : Réseaux de neurones

Cours

Afin d'appréhender les notions abordées dans le cours, il faut :

- Consulter les vidéos et transparents associés en ligne.
- Lire les exemples traités (sous R et Python).

Les ressources figurent sur *educnet*.

QCM

A l'issue des séquences de cours, il faut répondre à des QCM en ligne qui permettent de vérifier la compréhension des principaux concepts. Les dates limites sont les suivantes :

- Séquences 1, 2, 3 et 4 : 13 mars 2022.
- Séquences 5, 6 et 7 : 10 avril 2022.
- Séquences 8 et 9 : 1er mai 2022.

Challenges

Les étudiants constitueront un binôme **au sein de leur petite classe (PC)**. Ce binôme sera invariant pour tous les rendus.

2 réponses au challenge (synthèse courte & code) seront à rendre par binôme :

- Rendu n°1 : à partir au moins de la régression linéaire et de la régression régularisée.
- Rendu n°2 : à partir au moins des random forests et du gradient boosting.

Les rendus devront être envoyés à l'enseignant de PC **au plus tard** aux dates mentionnées dans le calendrier figurant dans ce document.

Seront fournis les codes (au format R ou Python) ainsi que les rapports (au format pdf ou html**)**, avec comme format de nom pour les fichiers :

Challenge_*nom étudiant 1*_*nom étudiant 2*_*(…)*

Projet

Les étudiants constitueront un groupe de 5 étudiants **au sein de leur PC**.

Afin de mener le projet dans le cadre de cours, il faudra :

- Choisir un jeu de données (si possible en lien avec votre spécialité) et déterminer une problématique.
- Mettre en œuvre les techniques de prévision vues dans le cours sur ce jeu de données.
- Rédiger un rapport présentant la démarche et les résultats obtenus.
- Présenter les résultats dans une soutenance courte (10 minutes).

Seront fournis les codes (au format R ou Python) ainsi que les rapports (au format pdf), avec comme format de nom :

Projet_*nom étudiant 1*_*nom étudiant 2*_*(…)*

Charge de travail

La charge de travail estimée pour le cours est détaillée dans le tableau suivant :

Contenu	**Vidéos**	**Poly**	**Exemples**	**QCM**
Séquence 1 : Apprentissage statistique	45 mn	60 mn	30 mn	15 mn
Séquence 2 : Régression régularisée	45 mn	60 mn	30 mn	15 mn
Séquence 3 : Noyaux de lissage	45 mn	60 mn	30 mn	15 mn
Séquence 4 : Régression spline & GAM	45 mn	60 mn	30 mn	15 mn
Séquence 5 : CART	45 mn	60 mn	30 mn	15 mn
Séquence 6 : Bagging	45 mn	60 mn	30 mn	15 mn
Séquence 7 : Gradient boosting	45 mn	60 mn	30 mn	15 mn
Séquence 8 : SVM	45 mn	60 mn	30 mn	
Séquence 9 : Réseaux de neurones	45 mn	60 mn	30 mn	15 mn
Total	**405 mn**	**540 mn**	**270 mn**	**120 mn**

Par ailleurs la charge de travail estimée par étudiant est de :

- 6h pour les challenges,
- 30h pour le projet.

3. Modalités pratiques

Calendrier indicatif du module

Semaine	Séances cours	Tâches et échéances cours	Tâches et échéances challenges	Tâches et échéances projet
07/02 - 13/02	10/02 (Présentiel) Présentation générale			
14/02 - 20/02	17/02 (Distanciel)			16/02 Constitution des groupes (et choix du langage)
21/02 - 27/02		Cours & QCM Séquences 1, 2, 3 et 4		
28/02 - 06/03				
07/03 - 13/03	10/03 (Présentiel)			13/03 Choix définitif du sujet et du jeu de données
14/03 - 20/03	17/03 (Distanciel)			
21/03 - 27/03	24/03 (Présentiel)		27/03 : 1er rapport challenge	
28/03 - 03/04	31/03 (Distanciel)	Cours & QCM Séquences 5, 6 et 7		
04/04 - 10/04	07/04 (Présentiel)			
11/04 - 17/04	14/04 (Distanciel)			
18/04 - 24/04	21/04 (Présentiel)	Cours & QCM Séquences 8 et 9	24/04 : 2e rapport challenge	
25/04 - 01/05	28/04 (Distanciel)			
02/05 - 08/05				
09/05 - 15/05	12/05 (Présentiel)			
16/05 - 22/05	19/05 (Distanciel)			
23/05 - 29/05				
30/05 - 05/06	02/06 (Présentiel) Soutenance projets			05/06 Rendu du rapport

Équipe enseignante

Les responsables de PC sont :

- Valentin CADORET : valentin.cadoret@enpc.fr
- Vincent LEFIEUX : vincent.lefieux@enpc.fr

Évaluation

	Prorata de la note finale	Objectif
QCM	10%	Vérifier votre compréhension des concepts
Challenges	35%	Mettre directement en application les concepts vus dans le cours
Projet	55%	Mettre en application les concepts vus dans le cours sur un cas d'étude circonstancié

Attention, pour valider le module, tous les QCM devront être réalisés, et les rapports rendus (challenges & projet). Dans le cas contraire, une pénalité de 0.5 point par QCM manquant et 2 points par challenge manquant sera appliquée sur la moyenne globale.