1CC5000 - Statistique et Apprentissage

Responsables : **Arthur TENENHAUS**

Département de rattachement : **DÉPARTEMENT MATHÉMATIQUES**

Langues d'enseignement : **ANGLAIS , FRANCAIS**

Type de cours : **Cours commun**

Campus où le cours est proposé : **CAMPUS DE PARIS - SACLAY**

Nombre d'heures d'études élèves (HEE) : **60**

Nombre d'heures présentielles d'enseignement (HPE) : **31**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

Dans ce cours, les étudiants devront acquérir les bases mathématiques, méthodologiques et numériques permettant de réaliser à partir d'observations d'un phénomène aléatoire (les données) une inférence sur la distribution de probabilité sous-jacente. Ainsi, ils seront en mesure d'analyser un phénomène passé ou de réaliser des prévisions pour un phénomène futur de nature similaire.
Pour cela, les étudiants devront dans un premier temps acquérir les formalismes, concepts et résultats élémentaires de la statistique mathématique. Cela inclut en particulier la définition de modèles statistiques, les principes de la théorie de l'estimation et de la théorie des tests d'hypothèses.
Dans un deuxième temps, les élèves se familiariseront avec les méthodes et algorithmes d'apprentissage statistique à partir des données, dans le cadre de l'apprentissage supervisé pour la statistique prédictive ou de l'apprentissage non-supervisé pour la statistique descriptive. Dans ce cadre, ils seront en particulier sensibilisés à la problématique de la grande dimension.
Finalement, les étudiants découvriront et utiliseront par des travaux pratiques en Python des bibliothèques et algorithmes d'apprentissage statistique.

Période(s) du cours (n° de séquence ou hors séquence) :

ST4

Prérequis :

Convergence-Intégration-Probabilités

Plan détaillé du cours (contenu) :

1. Variables aléatoires et échantillons, statistique descriptive, mesure empirique.

2. Estimation paramétrique
a. Familles de distributions et modèles paramétriques
b. Quelques estimateurs ponctuels : méthode de substitution, méthode des moments, maximum de vraisemblance
c. Propriétés des estimateurs ponctuels (biais, consistance, risque, Borne de Cramer- Rao, vitesse de

convergence, propriétés asymptotiques, normalité asymptotique, consistance et normalité asymptotique de l'EMV)
d. Théorème central limite, méthode delta, th. De continuité, et th. De Slutsky
e. Régions de confiance (fonctions pivotales, cas gaussien), et régions de confiance asymptotiques.

3. Estimation Bayésienne : théorème de Bayes, distributions a priori et a posteriori, exemples de distributions conjuguées, intervalle de crédibilité, fonctions de perte et estimateurs bayésiens ponctuels

4. Tests d'hypothèses statistiques
a. Principes et démarches générale d'un test : hypothèses alternatives, risques et puissance, statistique de test, région de rejet, p-valeur
b. Tests paramétriques : Lemme de Neyman-Pearson, tests asymptotiques
c. Tests non paramétriques d'ajustement (chi2, Kolmogorov)
5. Modèle Linéaire de régression et modèles additifs généralisés, arbres.

6. Sélection de modèles, Pénalisation L1 (lasso) et L2 (ridge), validation croisée.

7. Modèle logistique pour la classification

8. Introduction aux Réseaux de neurones.

9. Analyse en composantes principales + Méthodes non supervisées : Clustering (K-means)

Déroulement, organisation du cours :

9 x 1H30 de CM + 10 x 1H30 de TD + 3H de Contrôle.
Avec alternance de CM et TD.

Organisation de l'évaluation :

Un examen écrit obligatoire de 3h couvrant la totalité du programme

Moyens :

- Equipe enseignante (noms des enseignants des cours magistraux) :
  - 9 CM : Arthur Tenenhaus + Laurent Le Brusquet + Julien Bect (Anglais)
- 10 TDs/TPs (par défaut 50 élèves) :
- Outils logiciels : Les TPs seront réalisés en Python, avec notamment les librairies ScikitLearn, StatsModels, Scipy, Keras.
- Salles de TP (département et capacité d'accueil) : A priori TPs en classes normales avec portables des élèves.

Acquis d'apprentissage visés dans le cours :

A la fin de cet enseignement, l'élève sera capable de :
- modéliser un problème d'inférence statistique
- estimer les paramètres de ce modèle
- valider ou remettre en cause des hypothèses statistiques
- résoudre des problèmes de régression et de classification à partir de données
- identifier des sous-familles homogènes à partir de données

Bibliographie :

- Slides de cours + sujet de TDs
- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, pp. 241-249). New York: Springer series in statistics.