

3MD1530 - Modèles statistiques pour l'apprentissage automatique

Responsables : **Frederic PENNERATH**

Langues d'enseignement : **FRANCAIS**

Campus où le cours est proposé : **CAMPUS DE METZ**

Nombre d'heures d'études élèves (HEE) : **50**

Nombre d'heures présentielles d'enseignement (HPE) : **26**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

Les cours "Modélisation statistique" ModStat1 et ModStat2 traitent de la modélisation de systèmes pour lesquels les sorties sont suffisamment incertaines pour qu'elles nécessitent d'être modélisées par des variables aléatoires.

Le cours débute avec des rappels en statistique et l'introduction de modèles élémentaires (e.g. naive Bayes, régression linéaire, etc) pour aller progressivement vers des modèles plus complexes.

Si les cours présentent les modèles et méthodes élémentaires les plus utiles dans ce contexte de modélisation, ils ne se veulent pas être un catalogue exhaustif.

L'objectif est davantage de présenter au sein d'une théorie cohérente les concepts et outils théoriques communs à l'ensemble de ces modèles et méthodes et de montrer comment à partir d'hypothèses de modélisation propres à chaque problème concret, ces concepts sont assemblés logiquement avant d'aboutir à une méthode opérationnelle.

D'un point de vue pratique l'enjeu de ce cours est non seulement de donner aux étudiants les moyens de comprendre et d'utiliser à bon escient les implémentations existantes de modèles mais aussi de concevoir ses propres implémentations pour prendre en compte les particularités d'un problème donné.

Le cours s'attache à rattacher la théorie à la pratique : sont d'abord identifiées en cours les hypothèses associées à une classe de problèmes donnée, s'ensuit un travail théorique de modélisation, qui conduit à définir un modèle et ses algorithmes d'estimation. Ces résultats sont alors appliqués à un cas d'étude en TD avant de faire l'objet en TP d'un travail d'implémentation (en Python) et d'évaluation sur des données.

Le cours ModStat1 introduira les outils de bases de la modélisation statistique quand le cours ModStat2 se concentrera sur les modèles à variables cachées.

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

SD9

Prérequis :

- Connaissances de base en théorie des probabilités, statistique et apprentissage machine (machine learning)
- Niveau débutant en programmation Python / Numpy

Plan détaillé du cours (contenu) :

Volume total : 13,5 h. de cours, 4,5 h. de TD, 6 h. de TP et 2 h. d'examen écrit

Cours (13.5h)

- Rappels sur l'estimation statistique et l'inférence bayésienne (3h)
- Introduction aux réseaux bayésiens (1.5h)
- Modélisation, évaluation, calibration, etc : exemple de Naive Bayes (1,5h)
- Modèles gaussiens pour la classification : QDA, LDA, etc (1,5h)
- Modèles de régression linéaire : ridge, LASSO, etc (1,5h)
- Introduction à la théorie de l'information (1.5h)
- Famille exponentielle et modèles linéaires généralisés (3h)

TD (4.5h)

- Inférence bayésienne (1,5h)
- Modélisation d'un problème (1,5h)
- Régression linéaire et GLM (1,5h)

TP (6h)

- Modélisation d'un problème (3h)
- Régression linéaire et GLM (3h)

Examen écrit (2h)

Déroulement, organisation du cours :

- Les cours font appel dans la mesure du possible à des démos ou des exemples pour illustrer les concepts abstraits.
- Des TD (1,5h) et TP sur machine (3h) permettent de mettre en application d'un point de vue théorique et pratique les chapitres du cours.

Organisation de l'évaluation :

Examen écrit de 2 heures sans document portant sur des questions de cours et des exercices.

Moyens :

- Enseignant : Frédéric Pennerath (Cours, TD, TP), Joël Legrand (TP)
- Programmation en Python et ses bibliothèques standards.
- 1 groupe de TD
- 2 groupes de TP
- 2 étudiants max par PC

Acquis d'apprentissage visés dans le cours :

- Être capable de choisir un modèle/ une méthode statistique adaptée au problème considéré et de la mettre en œuvre de façon appropriée.
- Être capable de comprendre les concepts théoriques sous-jacents à une méthode d'inférence statistique présentée dans un article scientifique.
- Être capable d'implémenter un modèle / une méthode statistique dans un langage tel que Python.
- Être capable d'adapter un modèle / une méthode pour tenir compte des spécificités du problème traité.

Description des compétences acquises à l'issue du cours :

C1.2 - Use and develop appropriate models, choose the right modeling scale and relevant simplifying assumptions to address the problem

C2.1 - Have studied in depth a field or discipline related to the basic or engineering sciences

C6.4 - Solving problems in a computational thinking process

Bibliographie :

- Support de cours (Présentation du cours, polycopié)
- "Machine Learning - A Probabilistic Perspective" de Kevin Murphy (MIT Press, 2012)
- "Bayesian Reasoning and Machine Learning" de David Barber (Cambridge University Press, 2012)
- "All of Statistics: A Concise Course in Statistical Inference" de Larry Wasserman (Springer-Verlag, 2004)