DEPARTEMENT DE MATHEMATIQUES ET DE GENIE INDUSTRIEL

MTH8302 - Analyse de Régression et Analyse de Variance

Hiver 2025

Informations sur l'enseignant

Enseignant : Chiheb Trabelsi, PhD
E-mail : chiheb.trabelsi@polymtl.ca
Horaire du cours : Chaque Mercredi de 16h00 à 19h00
- 16h00-17h20 (cours)
- 17h20-17h30 (pause)
- 17h30-18h45 (cours)
- 18h45-19h00 (Q&R)

Cours : Présentiel
Consultation : Après chaque cours et par courriel

Dates importantes

Premier jour de classe : Mercredi 08 janvier 2025
Dernier jour de cours : Mercredi 16 Avril 2025
Date limite d'inscription (ajout) : 22 Janvier 2025
Date limite de retrait (Drop) : 22 Janvier 2025 (sans frais); 5 Février 2025 (avec frais)

Structure du cours

La structure du cours comprend :
- Des cours magistraux, devoirs théoriques et pratiques (4 ou 5 devoirs et pas d'examens).
- Des outils et bibliothèques modernes (en Python tels que les Scikit-learn, PyTorch)

Connaissances préalables

Une connaissance préalable est souhaitable **mais n'est pas obligatoire** en matière de programmation en Python, d'algèbre Matriciel, de calcul différentiel et des techniques d'optimisation, de probabilités et de statistique inférentielle. Une revue assez exhaustive est l'objet des premières séances.

Objet du cours

Le Cours couvre un éventail de méthodes **d'analyse supervisée** (Supervised Learning), allant des approches fondamentales aux techniques avancées adaptées à des problématiques complexes.

L'analyse supervisée est une approche d'apprentissage (d'estimation statistique) où un modèle est appris (estimé) à partir de données étiquetées. Cela signifie qu'un ensemble de données comprend des **entrées (**« features », attributs, inputs, prédicteurs, variables explicatives, « régresseurs », variables explicatives, ...) et leurs sorties associées (étiquettes ou cibles).

$$y = f(x) + \varepsilon$$

Le but est de trouver une fonction (transformation) f qui relie les entrées aux sorties pour faire des analyses d'influence et des prédictions sur de nouvelles données. **Elle comprend deux grandes familles de tâches :**

- **Régression** : Lorsque la variable **cible** est **continue** (par exemple, prédire un prix ou une température).
- **Classification** : Lorsque la variable **cible** est **discrète ou catégorique** (par exemple, reconnaître si un e-mail est spam ou non).

Objectif Général du cours

Ce cours offre une couverture des techniques **d'Analyse Supervisée (*Supervised Learning*),** en se concentrant sur les modèles de **régression, d'analyse de variance (ANOVA)**, **et leur extension vers les méthodes modernes de fouille des données (*Data Mining*), d'apprentissage machine et d'apprentissage profond**. Ces outils permettent d'analyser des données observationnelles (big data) et données expérimentales (small data), pour résoudre des problématiques complexes en modélisation tels que la prédiction, la classification des données, et les études d'impact (influence).

Objectifs Spécifiques du Cours

1. **Modélisation prédictive par régression** :
    - Développer et interpréter des modèles de régression pour établir des relations entre des variables explicatives X (continues ou catégoriques) et une variable réponse Y, qu'elle soit continue ou catégorique soit : $y = f(x) + \varepsilon$
    - Appliquer ces modèles à des données observationnelles et historiques, en explorant leur utilisation pour des prédictions précises dans des contextes variés.

2. **Classification par régression** :
    - Appliquer les techniques de régression (logistique, polynomiale, etc.) pour des tâches de classification, où la variable cible appartient à des catégories distinctes.

- Identifier les cas d'utilisation pertinents et évaluer la performance des modèles de classification par régression.

3. **Évaluation de l'influence des facteurs avec l'ANOVA** :
   - Comprendre et appliquer les modèles d'analyse de variance (ANOVA) pour tester l'influence de variables catégoriques X (facteurs) ou continues sur une variable réponse Y.
   - Explorer l'utilisation de l'ANOVA dans des contextes expérimentaux pour valider des hypothèses et évaluer des effets significatifs.
4. **Apprentissage machine pour l'analyse et la classification** :
   - Intégrer des techniques modernes d'apprentissage machine (Machine Learning), telles que les **forêts aléatoires (Random Forests)**, les **arbres de décision** pour l'analyse et la classification.
   - Comparer les performances des approches traditionnelles (régression, ANOVA) et modernes (Machine Learning) selon les types de données et objectifs.
5. **Apprentissage profond pour des applications avancées** :
   - Introduire les modèles d'apprentissage profond, y compris les **réseaux de neurones (ANN)** et les **réseaux de neurones convolutionnels (CNN)**, pour des tâches complexes d'analyse et de classification.
   - Explorer leurs applications spécifiques dans des contextes analytiques et prédictifs, notamment dans la classification d'images, le traitement de données volumineuses et non structurées, et l'analyse de relations complexes.

Compétences acquises à la suite du cours

À l'issue de ce cours les étudiants seront capables de :

- Identifier et appliquer les techniques adaptées (régression, ANOVA, Machine Learning) aux problématiques d'analyse supervisée.
- Construire, évaluer et interpréter des modèles analytiques avancés pour la prédiction et la classification.
- Exploiter les outils et bibliothèques modernes (tels que Scikit-learn ou PyTorch) pour résoudre des problématiques pratiques dans divers domaines d'application.

Ressources du cours

Site(s) web ayant trait au cours :

Stanford CS229 Machine Learning (Bonne resource pour le rappel en algèbre linéaire, optimization, probabilité, statistiques et Python)**:** https://cs229.stanford.edu

Du matériel pédagogique personnalisé, des documents académiques s et des ressources en ligne seront communiqués au fur et à mesure que le cours avance (Notes de cours, devoirs, travaux de programmation, slides, …)

Manuels et matériel de cours :

Voici Quelques manuels appropriés pour un cours sur la régression, l'analyse de la variance (ANOVA) et les extensions à l'apprentissage automatique et à l'apprentissage profond :

1. **An Introduction to Statistical Learning with Applications in Python (ISLP)** (Référence Principale pour la Régressioion)
   https://www.statlearning.com/?ref=dataschool.io

- **Authors :** Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
- **Publisher :** Springer (1st Edition, 2023)
- **ISBN :** 978-3032424751

- **Focus :**

  - Se concentre sur les applications pratiques avec des implémentations claires en Python.
  - Le contenu couvre les sujets suivants :
    - Qu'est-ce que l'apprentissage statistique ?
    - Régression
    - Classification
    - Méthodes de rééchantillonnage
    - Sélection et régularisation des modèles linéaires
    - Aller au-delà de la linéarité
    - Méthodes basées sur les arbres
    - Machines à vecteurs de support
    - Apprentissage profond
    - Analyse de survie
    - Apprentissage non supervisé
    - Tests multiples

2. **Applied Linear Statistical Models** (Référence Principale pour l'Analyse de la Variance)

- **Authors:** Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li
- **Publisher:** McGraw-Hill Education (5th Edition, 2004)
- **ISBN :** 978-0073108742

- **Focus :** Analyse de la variance (ANOVA) et de la covariance (ANCOVA) pour intégrer des prédicteurs continus et catégoriques.

3. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**

- **Authors:** Trevor Hastie, Robert Tibshirani, Jerome Friedman
- **Publisher :** Springer (2nd Edition, 2009)
- **ISBN :** 978-0387848570

- **Focus :** Couverture complète des techniques de régression, de classification et des techniques avancées d'apprentissage automatique. Les sujets incluent:
  - Les méthodes de régularisation (Lasso, Ridge)
  - Les modèles basés sur les arbres de décision.
  - Fait le lien entre les méthodes statistiques classiques et l'apprentissage automatique moderne.

Evaluation des travaux

- 4 ou 5 Devoirs à domicile

Planning des cours / Calendrier provisoire 2025

1. **Introduction**
    - Rappel en Algèbre Linéaire, Optimisation, Probabilité, Statistiques et Python
2. **Régression Simple**
    - Concepts de base et application de la régression linéaire simple
3. **Régression Multiple 1**
    - Introduction à la régression multiple
    - Interprétation des coefficients, et diagnostics de régression
4. **Régression Multiple 2**
    - Approfondissement des techniques de régression multiple
    - Gestion de la multicollinéarité
    - Sélection de modèles
5. **Régression Multiple 3**
    - Techniques avancées en régression multiple
    - Modèles linéaires généralisés
    - Régression logistique
6. **Régression MARS**
    - Présentation et application de la régression multivariée par spline adaptative (MARS)
7. **Introduction au Data Mining**
    - Concepts de base du data mining et introduction aux techniques et outils statistiques
8. **Réseaux de Neurones**
    - Réseaux de neurones artificiels pour la prédiction et la classification
9. **Arbres de Classification**
    - Techniques d'arbres de décision
    - Forêts aléatoires pour la classification et la régression
10. **Techniques de data mining avancées en Python**
    - Utilisation de Python pour appliquer des techniques de data mining avancées
11. **Introduction à l'ANOVA**
    - Principes de base de l'analyse de variance et conception des expériences
12. **ANOVA à Plusieurs Facteurs**
    - Analyse des effets de multiples facteurs et de leurs interactions via ANOVA
13. **ANOVA avec Mesures Répétées**
    - Gestion des données avec mesures répétées et analyse des effets longitudinaux
14. **Modèles Mixtes**
    - Combiner les facteurs fixes et aléatoires dans les modèles statistiques pour des analyses plus robustes

Ressources Supplémentaires pour le Cours

Accès aux notes de cours. Des lectures supplémentaires peuvent être également recommandées pour chaque chapitre.