Data mining : classification

Infos pratiques

> **ECTS** : 3.0
> **Nombre d'heures** : 32.0
> **Langue(s) d'enseignement** : Français
> **Niveau d'étude** : BAC +5
> **Période de l'année** : Enseignement neuvième semestre
> **Méthodes d'enseignement** : En présence
> **Forme d'enseignement** : Cours magistral et Travaux dirigés
> **Ouvert aux étudiants en échange** : Oui
> **Campus** : Campus de Nanterre
> **Composante** : Sciences économiques, gestion, mathématiques et informatique
> **Code ELP** : 5E9DADMC

Présentation

Ce cours pose les fondements théoriques et pratiques de la classification supervisée (prédiction de catégories). À partir du concept de classifieur théorique idéal (Bayes), il explore les méthodes concrètes pour concevoir des prédicteurs performants. L'enseignement aborde la notion de modélisation probabiliste et d'optimisation, tout en intégrant les enjeux clés de complexité (sur-apprentissage, dimensionnalité) et de performance du modèle.

Outils : R / RStudio

Applications : Banque, assurance, santé (scoring crédit, fraude, risques, déséquilibre de classes).

PLAN DU COURS

1. **Introduction à la classification supervisée** (cadre statistical learning/data science)
   • Motivation avec des exemples en banque, assurance et santé.
   • La perte 0-1, risque de prédiction, erreur de classification.
   • Classifieur de Bayes.
2. **Quelques méthodes de classification supervisée** (point de vue probabiliste/statistique)
   • Méthode des k plus proches voisins.
   • Modèles génératifs.
   • Régression logistique, estimation des paramètres par maximum de vraisemblance, interprétation et scoring.
3. **Critères de performance** (matrice de confusion, ROC,etc.).
   • Interprétabilité.
   • Déséquilibre des classes avec une attention particulière à la régression logistique.
4. **Sélection de modèles et validation dans le cadre de classification supervisée**
   • Complexité d'un modèle
   • Validation croisée et techniques de rééchantillonnage
5. **Quelques méthodes de classification supervisée** (point de vue optimisation)
   • Problème de minimisation du risque empirique en classification ;
   • Quelques fonctions de perte et majorants de la perte 0-1 (perte hinge, perte exponentielle, perte logistique etc)
   • Introduction à l'optimisation sous contrainte et à la régularisation.
   • Méthodes à noyaux et séparateurs à vaste marge : lien avec la perte hinge, optimisation du risque et la régularisation
   • Introduction aux méthodes d'ensemble : boosting, XGBoost, etc.

Objectifs

- Comprendre les bases statistiques et les étapes d'un pipeline de data science.
- Choisir et justifier la méthode de classification adaptée aux données.
- Maîtriser les enjeux de modélisation, de validation croisée et d'évaluation.
- Interpréter rigoureusement les résultats et métriques de performance.

Évaluation

Modalités : CC
SESSION 1 :
Contrôle Continu
• Type : Écrit, Dossier

• Durée :
Régime Dérogatoire
• Type : Écrit, Dossier
• Durée : 2h00
SESSION 2 :
• Type : Écrit
• Durée : 2h00
Utilisation de l'intelligence artificielle :
Dans le cadre de cet EC, l'usage de l'IA pour aider à la réalisation des travaux de contrôle continu soumis à évaluation ***est interdit***. Vous n'avez pas le droit de faire appel à une IA générative à des fins de documentation, recherche d'idées, construction, rédaction ou édition (hors usages de recherche web augmentée, de correction orthographique et syntaxique).

Compétences visées

- Maîtrise des concepts d'erreur de prédiction, de sur-apprentissage et de régularisation.
- Sélection méthodologique et arbitrage entre différents modèles.
- Évaluation de la performance via les métriques clés (courbes ROC, AUC, etc.)
- Mise en œuvre pratique et interprétation autonome sous R/RStudio.