Data mining : classification

Infos pratiques

> **ECTS** : 3.0

> **Nombre d'heures** : 36.0

> **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

> **Composante** : Sciences économiques, gestion, mathématiques et informatique

Présentation

Plan du cours

- Introduction à l'apprentissage supervisée.
  - Motivation : Le questionnement statistique à travers quelques exemples de la vie réelle (en en économie, finance, santé par exemple avec des vrais données). Définition de Machine Learning et apprentissage supervisé.

- Introduction au sélection de modèles
  - Décomposition et compromis biais-variance. Sur-apprentissage et sous-apprentissage
  - Techniques de re-échantillonnage et Validation Croisée.

- Classification supervisée - point de vue probabiliste/ statistique (1)
  - Classifieur de Bayes (CM)
  - La méthode de k-plus proches voisins. Choix d'un bon k.

- Classification supervisée - point de vue probabiliste/ statistique (2)
  - Modèles génératives : Analyse discriminante (linéaire, quadratique), nayve bayes, etc

- Classification supervisée - point de vue probabiliste/ statistique (3)
  - Régression Logistique.
  - Motivation à la courbe ROC. AUC

- Méthodes pénalisés, régularisation (pénalisation)
  - Ridge, Lasso, Elastic-net
  - Optimisation sous contrainte et Formule de Lagrange.

- Classification supervisée - point de vue optimisation (1)
  - Séparateur à Vaste marge (SVM)

- Classification supervisée - point de vue optimisation (2)
  - Méthodes d'arbres, Boosting,...

Objectifs

- Comprendre le vocabulaire et les concepts fondamentaux de l'apprentissage statistique.
- Etre capable d'identifier des questions concrètes.
- Analyser les données d'un point de vue de l'apprentissage statistique, modéliser, prédire, interpréter et répondre aux questions posées, expliquer les résultats fournis par le logiciel R, Rstudio.

Évaluation

Examen écrit ou projet : 100%

Pré-requis nécessaires

Ce cours peut être suivi par des étudiants ayant une connaissance basique des statistiques et probabilités.

Compétences visées

- Comprendre le vocabulaire et les concepts fondamentaux de l'apprentissage statistique.
- Etre capable d'identifier des questions concrètes.
- Analyser les données d'un point de vue de l'apprentissage statistique, modéliser, prédire,

interpréter et répondre aux questions posées,
expliquer les résultats fournis par le logiciel R, Rstudio.

Bibliographie

- Hastie, R. Tibshirani, and J. Friedman (2009) The Elements of Statistical LearningSpringer Series in Statistics.
- James, D. Witten, T. Hastie and R. Tibshirani (2013) An Introduction to Statistical Learning with Applications in RSpringer Series in Statistics.
- Philippe Besse. Statistique et Big Data Mining https://www.math.univ-toulouse.fr/~besse/enseignement.html