Apprentissage automatique (Machine learning)

Syllabus

Visseho Adjiwanou, PhD.

12 June 2024

Description

Ces deux cours (1INF2409 : Machine learning I, 2INF2409 : Machine learning II) vous présente l'utilisation de l'apprentissage automatique pour résoudre les problèmes de sciences sociales et de politiques publiques. Nous abordons le processus d'apprentissage automatique de bout en bout tout en nous concentrant sur quelques méthodes supervisées (régression, classification) et non supervisées (clustering). ce cours fera aussi une présentation de deep learning. Après ce cours, vous devriez avoir une vue d'ensemble des composants d'un pipeline et des méthodes d'apprentissage automatique, et savoir comment les utiliser pour résoudre des problèmes de sciences sociales. Enfin, ce cours vous donnera une explication intuitive des méthodes et vous fournira un cadre et des conseils pratiques sur la façon de les utiliser dans la pratique.

Objectifs du cours

- Comprendre les concepts fondamentaux de l'apprentissage automatique.
- Apprendre à implémenter et évaluer différents algorithmes de machine learning en utilisant R.
- Explorer des applications pratiques et avancées du machine learning.
- Développer des compétences en utilisation des packages R dédiés au machine learning.

Prérequis :

- Connaissances de base en R et en statistiques.
- Connaissances en algèbre linéaire et en calcul différentiel sont recommandées.

Jour 1 (6 heures)

Session 1 : Données digitales: forces et faiblesses (2 heures)

- Définition
- Forces et faiblesses
- Stratégies de recherche
- Qualité des données

Lectures

- https://www.bitbybitbook.com/fr/1st-ed/introduction/
- https://www.bitbybitbook.com/fr/1st-ed/observing-behavior/

Session 2 : Introduction et concepts de base (2 heures)

- Introduction à l'apprentissage automatique
- Types d'apprentissage : supervisé, non supervisé, semi-supervisé, par renforcement
- Cycle de vie d'un projet de machine learning

Lectures

- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2020). Big Data and Social Science: Data Science Methods and Tools for Research and Practice (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429324383
  - Chapitre 7: 7.1 à 7.5; 7.6.5 à 7.6.5.1.2; 7.8 à 7.12

Session 3 : Labo (2 heures)

- Labo
  - Préparation des données
  - Travail pratique : Introduction aux packages tidyverse et data.table pour la manipulation des données.

Jour 2 (6 heures)

Apprentissage automatique supervisé (I)

Session 4 : Régression linéaire et extensions (2 heures)

- Concept de la régression linéaire

- Méthode des moindres carrés
- Évaluation des performances : RMSE, MAE, R^2
- Travail pratique : Implémentation de la régression linéaire avec lm() et visualisation avec ggplot2.

Lectures

- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2020). Big Data and Social Science: Data Science Methods and Tools for Research and Practice (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429324383
  - Chapitre 7: 7.6.2 à 7.6.4
- Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511973000
  - Chapitre 7
- https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf
  - Chapitre 3
  - Chapitre 6

Session 5 : Régression logistique (2 heures)

- Concept de la régression logistique
- Fonction sigmoïde et probabilités
- Coût et optimisation (descente de gradient)
- Travail pratique : Classification binaire avec la régression logistique en utilisant glm().

Lectures

- https://www.bitbybitbook.com/fr/1st-ed/asking-questions/
- https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf
  - Chapitre 4

Session 6: Labo et travail à faire (2 heures)

- Labo et présentation du travail à faire

Jour 3 (6 heures)

Session 7 : Algorithmes de classification (2 heures)

- k-Nearest Neighbors (k-NN)
- Support Vector Machines (SVM)
- Arbres de décision et forêts aléatoires
- Travail pratique : Utilisation des packages class pour k-NN, e1071 pour SVM, et randomForest pour les forêts aléatoires.

Lectures

- Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511973000
  - Chapitre 7
- https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf
  - Chapitre 9
- Michael A. Bailey, Anton Strezhnev, and Erik Voeten. "Estimating Dynamic State Preferences from United Nations Voting Data." Journal of Conflict Resolution, August 2015.

Session 7 (bis): Forêts aléatoires

- Introduction aux forêts aléatoires
- Fonctionnement
- Mise en oeuvre:
  - Introduction aux packages : randomForest et ranger
  - Exemple pratique : Classification d'un jeu de données (par exemple, prédiction d'espèces d'iris)
  - Interprétation des résultats (importance des variables, précision, matrices de confusion)

Session 8 : Évaluation et validation (2 heures)

- Techniques de validation croisée
- Métriques de performance : précision, rappel, F1-score, courbe ROC-AUC
- Problèmes de surapprentissage et sous-apprentissage
- Travail pratique : Utilisation du package caret pour la validation croisée et l'évaluation des modèles.

Lectures

- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2020). Big Data and Social Science: Data Science Methods and Tools for Research and Practice (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429324383
  - Chapitre 7: 7.7
- https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf
  - Chapitre 5

Session 9: Labo

- Labo

Jour 4 (6 heures)

Apprentissage automatique non supervisé (II)

Session 10 : Clustering et méthodes non supervisées (2 heures)

- Introduction au clustering
- Algorithmes : k-means, DBSCAN, agglomératif
- Réduction de dimensionnalité : PCA, t-SNE
- Travail pratique : Application du clustering avec stats et dbscan, et de la réduction de dimensionnalité avec prcomp et Rtsne.

Lectures

- https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf
  - Chapitre 10

Session 11 : Apprentissage profond (deep learning)

- Introduction
  - Concepts de base des réseaux de neurones
  - Différence entre apprentissage automatique traditionnel et apprentissage profond
  - Applications de l'apprentissage profond
- Architecture des Réseaux de Neurones
  - Neurone artificiel et fonction d'activation
  - Réseau multicouche (MLP : Multilayer Perceptron)
  - Entraînement des réseaux de neurones (propagation avant, rétropropagation)
  - Hyperparamètres (nombre de couches, neurones par couche, taux d'apprentissage)
- Mise en oeuvre
  - Introduction aux packages : keras et tensorflow en R
  - Exemple pratique : Classification d'images de chiffres manuscrits (MNIST)
  - Construction, compilation et entraînement d'un modèle simple
  - Évaluation et interprétation des performances du modèle
- Applications et défis

Session 12 : Conclusion et perspectives (1 heure)

- Révisions et discussion sur les avancées récentes du domaine
- Perspectives de carrière dans le machine learning

Session 13: Travail à faire (1 heures)

- Présentation du travail à faire

Ressources et matériel recommandé :

Livres

- Salganik, M. J. (2018). Bit by bit: Social research in the digital age. Princeton University Press.
- Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9780511973000

- Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2020). Big Data and Social Science: Data Science Methods and Tools for Research and Practice (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429324383
- Provost and Fawcett's *Data Science for Business* [@FawcettProvost] is a good practical handbook for using machine learning to solve real-world problems.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (2nd edition). Springer. is a classic and is available online for free.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). An Introduction to Statistical Learning: With Applications in R. Springer New York. from the same authors, includes less mathematics and is more approachable. It is also available online.
- Mitchell, T. M. (1997). Machine Learning (1st edition). McGraw-Hill Education. is a classic introduction to some of the methods and gives a good motivation underlying them.
- Wu, X., & Kumar, V. (Eds.). (2009). The Top Ten Algorithms in Data Mining (1st edition). Chapman and Hall/CRC.

Logiciels:

- R possède de nombreux packages pertinents.[1]
- Python (with libraries like `scikit-learn`, `pandas`, and more).
- Cloud-based: AzureML, Amazon ML, Google
- Free: KNIME, Rapidminer, Weka (mostly for research use).
- Commercial: IBM Modeler, SAS Enterprise Miner, Matlab.

Cours en ligne

De nombreux excellents cours sont disponibles en ligne [@MLcourses].

Conférences

Les principales conférences dans ce domaine comprennent :

- International Conference on Machine Learning,
- Annual Conference on Neural Information Processing Systems (NeurIPS), - ACM International Conference on Knowledge Discovery and Data Mining (KDD).

[1] https://cran.r-project.org/web/views/MachineLearning.html