

# Université Lyon 2, Master 2 Informatique, parcours MALIA MAchine Learning pour l'Intelligence Artificielle

Syllabus des cours

2022-2027

## Table des matières

<b>1</b>	<b>Optimisation</b>	<b>2</b>
<b>2</b>	<b>Ensemble methods in Machine Learning</b>	<b>3</b>
<b>3</b>	<b>Manifold learning</b>	<b>4</b>
<b>4</b>	<b>Model-based Machine Learning</b>	<b>5</b>
<b>5</b>	<b>Deep Learning</b>	<b>6</b>
<b>6</b>	<b>Temporal data analysis</b>	<b>7</b>
<b>7</b>	<b>Representation learning for Natural Language Processing</b>	<b>8</b>
<b>8</b>	<b>Big Data Management</b>	<b>9</b>
<b>9</b>	<b>Interpretability and explainability</b>	<b>10</b>
<b>10</b>	<b>Network data analytics for Information retrieval</b>	<b>11</b>
<b>11</b>	<b>Recent advances in ML</b>	<b>12</b>
<b>12</b>	<b>Travail d'Etude et de Recherche</b>	<b>13</b>

# 1 Optimisation

## Pré-requis

Connaissance de calcul différentiel et intégral en contexte multivarié, notions de gradient, de jacobien, de hessiennes, etc.

## Mots clefs

Minimisation sans contrainte, minimisation avec contraintes, analyse convexe, dualité lagrangienne, algorithmes du gradient, accélération de la méthode du gradient, adagrad, méthode du gradient stochastique, méthodes proximal et de descente miroir, applications au machine learning.

## Intervenants

Stéphane Chrétien

## Volume horaire

10h CM

## Objectifs

Connaître les algorithmes appropriés pour les problèmes d'optimisation qui surviennent dans l'implémentation de modèles en machine learning.

## Connaissances à acquérir

Méthodes classiques et leurs raffinements récents (méthodes de gradient, méthodes de descente miroir, variantes stochastiques, accélération de Nesterov et autres, adagrad, etc), méthodes lagrangiennes pour les problèmes sous contraintes, problèmes classique (programmation linéaire, quadratique, positive semi-définie, conique).

## Compétences à acquérir

Savoir programmer les méthodes vue en cours sous python pour les problèmes classiques du machine learning.

## Modalités pédagogiques

Essentiellement des CM avec quelques applications sous la forme de TD.

## 2 Ensemble methods in Machine Learning

### Pré-requis

Algèbre linéaire, Analyse et Probabilités de Licence, Optimisation, Modélisation statistique, Méthodes de base en apprentissage automatique.

### Mots clefs

Décider en comité, Fusion/Agrégation d'information, Bagging, Random Forest, Boosting, Gradient Boosting

### Intervenants

Julien Ah-Pine (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Maîtriser les fondements, concepts, et la mise en pratique avec des outils numériques, de plusieurs méthodes d'ensemble en apprentissage automatique.

### Connaissances à acquérir

- Les motivations et les concepts importants de l'apprentissage ensembliste,
- Les différentes grandes stratégies pour l'apprentissage ensembliste en catégorisation, en régression (et en classification automatique si le temps le permet),
- Les concepts, fondements, et méthodes classiques issues du bagging,
- Les concepts, fondements et méthodes classiques issues du boosting,

### Compétences à acquérir

Savoir mettre en œuvre et mettre en concurrence des méthodes d'ensemble sur des jeux de données avec un logiciel de data science (R ou Python).

### Modalités pédagogiques

L'enseignement en présentiel représente un volume horaire de 21h et comporte des CM ainsi que des TD/TP. Les supports de cours sont mis à disposition et projetés lors des séances. Des exercices et travaux supplémentaires devront également être effectués en dehors des séances. L'évaluation comportera un examen individuel et un projet à réaliser à plusieurs.

## 3 Manifold learning

### Pré-requis

Notions d'Algèbre matricielle et optimisation

### Mots clefs

réduction non linéaire de la dimension, données en grandes dimensions, visualisation des données,

### Intervenants

Jairo Cugliari (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Maîtriser les approches de réduction (linéaire et non linéaire) de la dimension, pour l'extraction d'attributs et la représentation visuelle des données en grande dimension.

### Connaissances à acquérir

1. Introduction. Différences entre apprentissage supervisé et non supervisé. 2. Estimation de la densité. 3. Méthodes linéaires de réduction de la dimension. 4. Méthodes non linéaires de réduction de la dimension.

### Compétences à acquérir

Bonne maîtrise des méthodes d'apprentissage non supervisée notamment celles adaptées aux grandes bases des données.

### Modalités pédagogiques

Étude théorique des principales approches. Mise en pratique par des TP et un projet.

## 4 Model-based Machine Learning

### Pré-requis

Notions de probabilités et statistiques : variable aléatoire, loi de probabilité, indépendance, estimation, vraisemblance, espérance conditionnelle.

### Mots-clefs

classification, clustering, modèles de mélanges, algorithme EM et variantes, sélection de modèles, co-clustering, données complexes

### Intervenants

Julien Jacques (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Après un rappel général des principales méthodes de classification et de clustering, l'objectif de ce cours est de découvrir la classification et le clustering par le biais des modèles de mélanges, leur estimation, les avantages et les inconvénients de ces approches. Ces techniques seront mises en pratique sur données réelles à l'aide du logiciel R.

### Connaissances à acquérir

Le modèle de mélange, ses hypothèses et ses conséquences en termes de règle de classification. L'estimation du modèle de mélange par maximum de vraisemblance dans le cadre supervisé (classification) et non supervisé (clustering). L'algorithme EM et ses variantes pour la maximisation de la vraisemblance dans le cadre supervisé. Les modèles de mélanges spécifiques aux données modernes : grande dimension, natures diverses. Les techniques de choix de modèles. Le modèle de mélange dans le cadre du co-clustering.

### Compétences à acquérir

Savoir résoudre un problème de classification supervisée ou non supervisée avec des approches à base de modèles probabilistes, savoir identifier les bons modèles, choisir entre différents modèles. Face à un nouveau modèle, être capable de mettre en place une stratégie d'estimation, de choix de modèles, et les implémenter informatiquement sous le logiciel R.

### Modalités pédagogiques

Les séances alterneront entre cours et mise en pratique sous R. Les cours seront à la fois sous la forme de présentation électronique vidéo-projetée mais également au tableau pour les développements mathématiques les plus pointus. Des travaux pratiques seront à réaliser entre les séances. Le cours sera évalué à la fois par un examen écrit et par un projet d'implémentation sous le logiciel R.

## 5 Deep Learning

### Pré-requis

algèbre linéaire (espaces vectoriels), bases en apprentissage automatique et en optimisation

### Mots clefs

perceptron multicouches, apprentissage par descente de gradient, réseaux convolutifs, auto-encodeurs, réseaux récurrents

### Intervenants

Jairo Cugliari (Université Lyon 2) et Julien Velcin (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Présenter le concept des réseaux de neurones profonds (deep learning) et ses applications notamment dans le traitement des données images, textuelles et séries chronologiques

### Connaissances à acquérir

- Historique des réseaux de neurones
- Principales définitions et applications
- Architecture du perceptron simple
- Réseaux de neurones multicouches
- Apprentissage des paramètres par descente du gradient
- Architectures spécialisées : Auto-encodeurs, réseaux récurrents, réseaux de convolution

### Compétences à acquérir

- Mettre en place une architecture profonde
- Mise en pratique sur logiciel spécialisé (ex. Python ou R)

### Modalités pédagogiques

30% pour la présentation des concepts et des algorithmes, 70% pour les applications avec un logiciel spécialisé

## 6 Temporal data analysis

### Pré-requis

Méthodes de bases en modélisation (régression linéaire) et apprentissage automatique (clustering, classification).

### Mots clefs

Times series, forecasting, clustering, classification.

### Intervenants

Julien Jacques et Jairo Cugliari (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Savoir résoudre les principales tâches de machine learning lorsque les objets étudiés sont des séries temporelles : prévision du futur d'une série temporelle, clustering de séries temporelles et classification de séries temporelles.

### Connaissances à acquérir

- Prévision de la suite d'une série temporelle (lissage exponentiel, SARIMA, modèles neuronnux)
- Clustering de séries temporelles
- Classification de séries temporelles

### Compétences à acquérir

- Prédire la suite d'une série temporelle
- Résumer de grand ensembles de séries temporelles
- Classer des séries temporelles dans des groupes prédéfinis

### Modalités pédagogiques

- Étude théorique des principales approches
- Mise en pratique par des TP sous R et un projet

## 7 Representation learning for Natural Language Processing

### Pré-requis

Notions en algèbre linéaire, notions en optimisation

### Mots clefs

Traitement automatique du langage naturel, fouille de textes, représentation des mots et des documents, réseaux de neurones profonds, apprentissage supervisé

### Intervenants

Adrien Guille (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

- Comprendre les fondements théoriques des techniques les plus récentes pour résoudre efficacement des problèmes de classification et de régression définis sur du texte.
- Savoir programmer des solutions intégrant ces techniques.

### Connaissances à acquérir

- Principales techniques pour l'apprentissage de représentations des mots : Word2Vec, GloVe
- Apprentissage supervisé à partir du texte avec les réseaux de neurones convolutifs et les réseaux de neurones récurrents
- Apprentissage supervisé à partir du texte par spécialisation de modèles de langage pré-entraînés

### Compétences à acquérir

- Modéliser un problème de classification/régression défini sur du texte
- Programmer une solution efficace avec TensorFlow

### Modalités pédagogiques

- Étude théorique des principales approches
- Mise en pratique par des TP et un projet



## 8 Big Data Management

### Pré-requis

- Maîtrise des bases de données relationnelles.
- Connaissances SQL, connaissances de base du concept de Datawarehouse.

### Mots clefs

Entrepôts de données, OLAP, ETL, NoSQL, Hadoop, Apache Spark

### Intervenants

Jérôme Darmont, Mohamed-Lamine Messai (Université Lyon 2), Nicolas Dubois (BIAL-X)

### Volume horaire

12h CM, 9h TD

### Objectifs

- Aborder les concepts et la pratique de l'entreposage de données, ainsi que les bases de données NoSQL et le traitement de données en environnements distribués à grande échelle.
- Découvrir l'environnement PDI
- Découvrir les composants d'extraction de transformation et d'alimentation
- Alimenter des tables de dimension et des tables de fait à travers un cas réel (Alimentation d'un Datamart Ressources Humaines)

### Connaissances à acquérir

- Modélisation multidimensionnelle, intégration de données (ETL), analyse en ligne (OLAP)
- Modèles NoSQL, architectures distribuées
- Modélisation dimensionnelle avec les formalismes de Ralph KIMBALL
- Principes d'alimentation des tables de dimension et des tables de fait
- Rôle et bénéfices d'un référentiel de développement

### Compétences à acquérir

- Modélisation multidimensionnelle avec les formalismes UML et Golfarelli-Rizzi
- Parallélisation de code informatique
- Pratique des logiciels Indyco Builder, MicroStrategy Desktop, Tableau Public, MongoDB, Apache Spark
- Développer des flux ETL
- Connaître un Système de Gestion de Base de données relationnelles
- Maîtriser le SQL pour interroger une base de données

### Modalités pédagogiques

Cours/TD et projets en groupes

## 9 Interpretability and explainability

### Pré-requis

Les étudiants sont encouragés à avoir une aisance en algèbre linéaire, probabilité, algorithmes et une bonne connaissance générale du machine learning. Il est attendu une bonne maîtrise de Python avec numpy et scikitlearn.

### Mots clefs

Interprétabilité, explicabilité, deep learning, human in the loop.

### Intervenants

Stéphane Chrétien (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Les modèles d'apprentissage automatique sont de plus en plus utilisés pour aider les décideurs dans des contextes à enjeux élevés tels que les soins de santé et la justice pénale. Il est important de s'assurer que les décideurs et utilisateurs de ces outils comprennent correctement. Ce cours vise à familiariser avec les avancées récentes dans le domaine émergent du Machine Learning interprétable et explicable.

### Connaissances à acquérir

Les différentes classes de modèles interprétables (par exemple, les approches basées sur des prototypes, les modèles linéaires parcimonieux, les techniques basées sur des règles de décision, les modèles additifs généralisés), les explications post-hoc (explications de boîte noire, y compris les explications contrefactuelles), et une bonne compréhension des liens entre interprétabilité et causalité.

### Compétences à acquérir

Savoir mettre en œuvre diverses approches récentes permettant de rendre interprétables et explicables des résultats obtenus par machine learning.

### Modalités pédagogiques

Étude théorique des principales approches Mise en pratique par des TP et un projet

## 10 Network data analytics for Information retrieval

### Pré-requis

Notions en algèbre linéaire, théorie des graphes, optimisation

### Mots clefs

modélisation des réseaux d'information, graphes, données textuelles, recherche d'information, apprentissage de représentation pour les graphes

### Intervenants

Julien Velcin (Université Lyon 2)

### Volume horaire

12h CM, 9h TD

### Objectifs

Comprendre les fondements théorique de l'analyse de données, en particulier textuelles, organisées en réseau (modélisation à l'aide de graphes) Savoir implémenter des solutions efficaces pour analyser des données textuelles organisées en réseau dans une optique de recherche d'information

### Connaissances et compétences à acquérir

Modéliser un problème d'analyse d'un réseau d'information dans une optique de recherche d'information Déployer des algorithmes de machine learning (supervisé et non supervisé) adapté au problème posé

### Modalités pédagogiques

Présentation de la théorie qui fonde les principales approches Mise en pratique de ce qui a été vu par des TP et un projet sur un cas d'étude

## 11 Recent advances in ML

Le contenu de cet UE évoluera chaque année, notamment en fonction des chercheurs internationaux invités au laboratoire. Ce sera l'occasion de découvrir un domaine complémentaire aux autres UE.

## **12 Travail d'Etude et de Recherche**

### **Pré-requis**

Ensemble des enseignements du master MALIA

### **Mots clefs**

Initiation à la recherche

### **Intervenants**

Tous les intervenants du master MALIA

### **Volume horaire**

80h de projets.

### **Objectifs**

S'initier à la recherche par la pratique.

### **Compétences à acquérir**

Répondre à une problématique de recherche scientifique.

### **Modalités pédagogiques**

Projet à réaliser en binôme, avec une partie bibliographique et planification du projet sur l'ensemble de l'année, puis une partie mise en pratique avec un mois à temps plein.