



Ecole d'Été sur
**L'INTELLIGENCE
ARTIFICIELLE**

**SYLLABUS DE LA SEMAINE
DE MACHINE LEARNING**

PAR L'ÉQUIPE PÉDAGOGIQUE



1 Introduction

L'introduction de l'apprentissage automatique (Machine Learning) s'étale sur 6 jours. Durant cette semaine, les apprenants vont construire différents modèles prédictifs, comprendre les types de problèmes dans lesquels ils peuvent être utilisés et interpréter le résultat de ces modèles (éventuellement, ses paramètres). De plus, l'équipe d'enseignants s'assurera que les élèves se familiarisent avec le processus de développement de modèles de la définition d'un problème à la maintenance du modèle. En somme, les objectifs de la semaine, telle que décrite dans les sections suivantes, sont:

1. Les élèves sont capables, étant donné une base de donnée et connaissant la typologie du modèle (régression linéaire, k-means, régression logistique, arbre de décision, réseaux de neurones...) voulu, de calibrer le modèle avec du code Python;
2. Les élèves sont capables, étant donné une base de donnée et un problème défini, de donner une proposition de modèle qui serait le plus adapté pour résoudre le problème, en justifiant leur choix à partir des caractéristiques des données (quantité, nature des variables, etc.) et des objectifs du projet;
3. Les élèves sont capables d'interpréter les paramètres des modèles qui en possèdent et de faire des prédictions avec Python;
4. Les élèves sont familiers avec les différentes étapes d'un projet de développement de modèle, ont un premier contact avec les principaux concepts et sont capables d'expliquer l'intérêt de chacune de ces étapes.

2 Emploi du temps

	Jour 1: Introduction	Jour 2: Classification	Jour 3: Régressions	Jour 4: Apprentissage profond	Jour 5: Optimisation	Jour 6: Synthèse
8h00	Introduction à l'apprentissage statistique (Espéran Padonou) 3h 30mn de pause Processus de développement de Modèles (Part I) 1h	K-means (Johannès) 1h30 30mn de pause TPs K-means - 2h	Régression Linéaire (____) 1h30 30 mn de pause TPs de Reg. linéaire 2h	Introduction aux réseaux de Neurones (Syntyche/Junior) 3h30	Optimisation (Rodolphe Leriche) 3h30	Processus de développement de Modèles (Part II) (____) 2h/3h
14h00	Processus de développement de Modèles Part I 1h 15mn de pause Statistique Graphique (____) 1h30	Arbres de décision (Kevin Kpakpo) - 1h30 15mn de pause TPs arbres de décision 2h	Régression Logistique (Modeste Dayé) - 1h30 15mn de pause TPs de logistique - 2h	Conférence	TP optimisation - 2h	

3 Description des cours

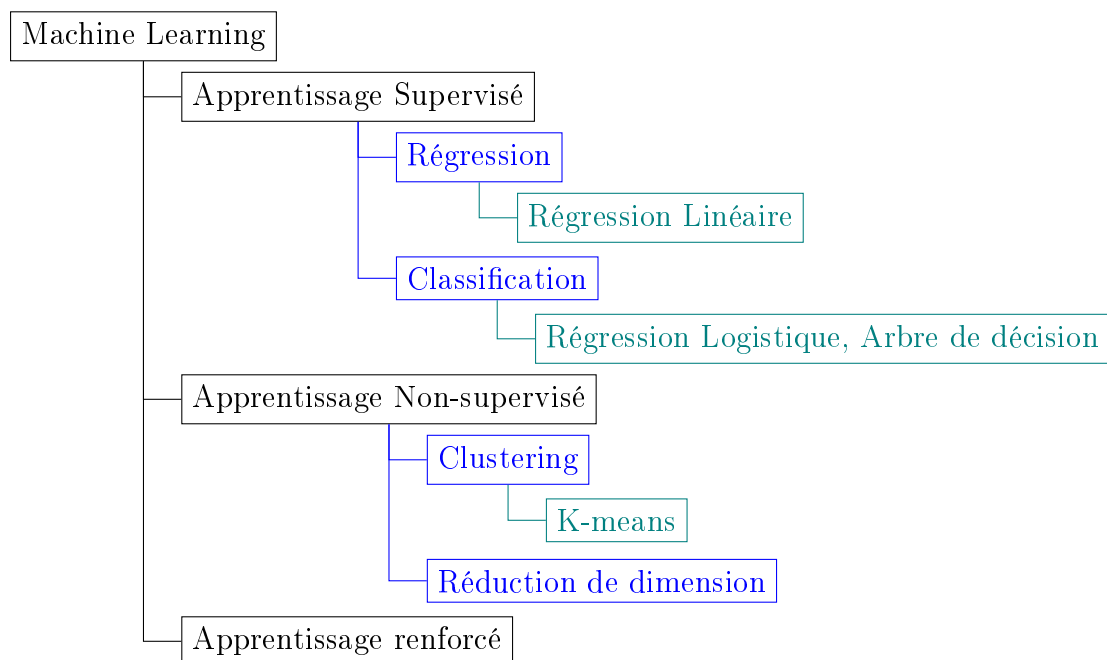
3.1 Introduction à l'apprentissage statistique

Durée: 3h

Résumé: Le cours donne un panorama de différentes applications dans lesquelles l'utilisation de modèles statistiques peut apporter des éléments de réponses. Il doit confronter les élèves à des problèmes concrets et les amener à définir des questions telles qu'un modèle puisse répondre à ceux-ci.

Acquis du cours: A la sortie du cours, les élèves doivent:

1. Connaître les catégories de problème de ML (Régression, Classification, Clustering, Réduction de dimension, Apprentissage renforcé) en définissant les termes “non-supervisé” et “supervisé”;
2. Etre capable de donner des exemples concrets de problèmes pour chacune de ces catégories;
3. Comprendre le concept d'erreur statistique: le modèle prédictif impose un type de relation qui crée une erreur qu'on cherche à minimiser (d'où l'utilité du cours d'optimisation)



3.2 Processus de développement de Modèles - Part I

Durée: 2h

Résumé: Le cours présente les étapes de développement de modèles statistiques. Il se concentre sur l'aspect théorique de chaque étape et justifie leur intérêt. Il doit donner aux élèves le point de vue d'un modélisateur face à un problème: “Comment je collecte les données? Quelle est la qualité de ces données? Comment j'améliore cette qualité? (Prétraitement des données) Je ne connais pas les variables et les relations entre elles. Comment résumer rapidement ces relations? (analyse exploratoire des données) Vu les relations que j'observe, il y a-t-il un intérêt à

créer de nouvelles variables? (ingénierie des caractéristiques) Comment faire pour calibrer et savoir s'il se comportera bien sur de nouvelles données? (entraînement et évaluation de modèle) Pendant que l'utilisation de ce modèle, comment je sais qu'il fait de bonnes prédictions?"

Acquis du cours: A la sortie du cours, les élèves doivent:

1. Connaître les étapes du processus de développement et expliquer leur intérêt;
2. Donner des exemples d'indicateurs de performance simple: RMSE, MAE, Précision, Sensibilité;
3. Comprendre que les cours qui suivent ne décrivent que l'étape d'entraînement de modèle.

3.3 Statistique Graphique

Durée: 1h30

Résumé: Le cours aborde une partie de la problématique de l'Analyse Exploratoire des Données. Il présente quelques graphes utilisés pour extraire des informations d'une base de données. Les élèves sont initiés aux catégories de variables (quantitatives et qualitatives). A partir de ces définitions, le cours leur donne des éléments pour décider du graphe le plus approprié suivant l'objectif voulu.

Acquis du cours: A la sortie du cours, les élèves doivent être:

1. capable d'expliquer l'objectif des graphes;
2. capable de classer une variable dans une catégorie (continue, discrète, ordinale, nominale);
3. capable d'interpréter une moyenne, une médiane, un mode, l'écart type, et l'intervalle interquartile;
4. capable de décider quand utiliser un histogramme, un barplot, un graphique linéaire, nuage de points, une boîte à moustache.

3.4 K-means

Durée: 1h30 - TD 2h

Résumé: Le cours présente l'algorithme des K-means à partir d'un exemple simple. Les élèves s'entraînent à calculer les barycentres et les distances des points à ceux-ci.

Acquis du cours: A la sortie du cours, les élèves doivent être:

1. Être capable d'expliquer l'algorithme de K-means;
2. comprendre dans quels contextes ce type de modèle est conseillé;
3. Être capable de calibrer un modèle et de faire des prédictions.

3.5 Arbre de décision

Durée: 1h30 - TD 2h

Résumé: Le cours présente un jeu qui peut être optimisé à l'aide d'un arbre de décision. A travers ce jeu, l'élève se familiarise à la structure d'un arbre. Il donne à l'élève une compréhension détaillée de l'entropie de Shannon. Elle est utilisée pour quantifier/représenter le gain d'information (réduction d'hétérogénéité).

Acquis du cours: A la sortie du cours, les élèves doivent être:

1. Être capable de définir un arbre de décision et les principales notions;
2. Comprendre les contextes (segmentation des problèmes) dans lesquels ce type de modèle est judicieux;
3. Être capable de décrire l'algorithme de construction de l'arbre;
4. Être capable de calibrer un modèle et de faire la prédiction;
5. Être capable d'évaluer le modèle;
6. (optionnel) citer les inconvénients de l'arbre de décision et présenter la forêt aléatoire.

3.6 Régression Linéaire

Durée: 1h30 - TD 2h

Résumé: le cours implémente une régression linéaire pour modéliser la consommation de poulet avec la population des pays africains.

Acquis du cours: A la sortie du cours, les élèves doivent:

1. Être capable de définir la régression linéaire (et les principaux termes: coefficients, ordonnée à l'origine);
2. Être capable de donner des exemples pour lesquels, calibrer une régression linéaire est possible;
3. Être capable de calibrer le modèle et comprendre l'interprétation des paramètres;
4. Être conscient du risque d'intégrer des valeurs extrêmes dans les données de calibration;
5. Comprendre le lien entre la calibration et l'optimisation.

3.7 Régression Logistique

Durée: 1h30 - TD 2h

Résumé: Comme exemple de méthode de classification, la régression logistique est présentée dans ce cours à travers une étude de cas simple. Il guidera l'élève pour prédire le mode de transport choisi suivant les conditions météorologiques et de circulation.

Acquis du cours: A la sortie du cours, les élèves doivent:

1. Être capable de définir la régression logistique (coefficients, formula);

2. Etre capable de donner des exemples pour lesquels, calibrer une régression logistique est conseillé, en particulier, comprendre pourquoi elle est préférable à une régression linéaire;
3. Etre capable de calibrer le modèle et comprendre l'interprétation des paramètres.

3.8 Introduction de Réseaux de Neurones

Durée: 3h30

Résumé: Afin de prendre la mesure de l'ubiquité de l'apprentissage profond, le cours commence par présenter aux élèves des cas d'application dans divers secteurs. Par ailleurs, une discussion critique des méthodes classiques permet à l'élève de comprendre la valeur ajoutée et les avantages des méthodes d'apprentissage profond. Enfin, dans une partie plus technique, l'élève abordera la structure détaillée des réseaux de neurones.

Acquis du cours: A la sortie du cours, les élèves doivent:

1. Etre capable de citer des cas d'application courant des méthodes d'apprentissage profond;
2. Etre capable de distinguer les ces méthodes aux méthodes plus classiques (citer les situations dans lesquelles elles sont préférables);
3. Etre capable d'entraîner une réseaux de neurones et comprendre comment illustrer le surapprentissage.

3.9 Optimisation

Durée: 1h30

Résumé:

Acquis du cours: A la sortie du cours, les élèves doivent: