
PLAN DE COURS

*Apprentissage automatique pour données massives
(8INF919)*

*Session : A2022
Département d'informatique et de mathématique*

Formule pédagogique

Les cours magistraux sont dispensés **en mode hybride (présentiel et virtuel en synchrone via zoom)**, lors de la période réservée au cours et spécifiée à l'horaire de cours officiel. Pour la connexion à distance au cours, veuillez consulter l'annexe en page 7.

Insertion du cours dans le programme

Le cours est optionnel dans les divers programmes d'études de cycles supérieurs du département d'informatique et de mathématique (DIM) de l'Université.

Objectifs généraux du cours

Le monde de l'informatique décisionnelle d'aujourd'hui est confronté à de nouvelles difficultés liées à une augmentation exponentielle du volume de données, qu'il est irréaliste de fouiller (analyser) avec une approche centralisée. Mais le volume n'est pas la seule contrainte pour la fouille de données. En effet, les données ne sont pas structurées et arrivent souvent sous forme de flux de données et doivent être traitées en temps réel. Ces contraintes ont donné naissance à ce qu'on appelle le BigData. Le passage à l'échelle de l'apprentissage automatique pour faire face à la règle du 3V qui caractérise le BigData, à savoir la Volumétrie, la Variété et la Vitesse des données, consiste à paralléliser les algorithmes d'apprentissage et/ou à construire d'une manière incrémentale les modèles d'apprentissage, il s'agit alors de l'apprentissage en ligne.

Dans ce cours, nous allons aborder les approches actuelles pour faire passer à l'échelle l'apprentissage automatique à partir des données et étudierons plus longuement des opérations d'apprentissage automatique en environnement distribué et en ligne. Nous aborderons dans ce contexte certains problèmes fréquents dans l'analyse des données massives. À la fin du cours, l'étudiant sera capable de mener des projets de d'apprentissage automatique pour le bigdata, de bout en bout.

Objectifs spécifiques du cours

Le cours s'inscrit dans une problématique générale liée à l'extraction de connaissances et présente les concepts fondamentaux, ainsi que les techniques principales d'apprentissage automatique en ligne et distribué. Au terme de ce cours, les étudiants devront être en mesure:

- (i) de maîtriser le processus général de la fouille de données massives;
- (ii) de distinguer les avantages et les inconvénients des principales approches;
- (iii) de se familiariser avec les algorithmes d'apprentissage automatique en ligne et distribué

Contenu du cours

Cours magistraux

Le cours permettra aux étudiants l'apprentissage des éléments suivants (voir l'annexe 1 pour plus de détails) :

1. Concept d'apprentissage automatique pour les données massives;
2. Algorithmes de classification et de clustering pour les données massives;
3. Apprentissage distribué et en ligne,
4. Évaluation d'un modèle d'apprentissage- pour les données massives
5. Ingénierie de la connaissance et processus de fouille de données massives.

Laboratoire

Travaux sous forme de synthèse d'articles de recherche et d'expérimentation sont prévus dans l'environnement d'apprentissage Spark, PySpark, Weka-MOA, WekaSpark,

Modalité d'évaluation

Calendrier

Synthèse d'articles et travaux pratiques	40 %
Présentation de mini-projets	60%

Qualité du français écrit

Tout travail remis doit être conforme aux exigences de la politique institutionnelle en matière de la maîtrise du français écrit du Manuel de Gestion (www.uqac.ca > Employés >. Le manuel de gestion PDF > lien de l'index, section 3.1.1-012).

Pénalité pour retard

En cas de retard dans la remise des travaux, une pénalité de 10% par jour sera appliquée. Un retard de plus d'une semaine ne sera pas accepté. Les règlements de l'UQAC concernant le plagiat seront strictement appliqués.

Note de passage

La note de passage est fixée C+.

Évaluation du cours

Ce cours sera évalué, conformément à la résolution du programme de cycle supérieur, à une date à déterminer entre le milieu et la fin du trimestre.

Soutien pédagogique

N'hésitez pas à m'envoyer des courriels. Veuillez noter qu'il est également possible que le professeur soit également disponible en dehors de ces périodes en mode virtuel.

Références

J. Leskovec, A. Rajaraman, J. Ullman. Mining of Massive Datasets. Cambridge University Press. (<http://www.mmids.org>)
Karau, H., A. Konwinski, P. Wendell et M. Zaharia. Learning Spark - Lightning-Fast Big Data Analysis. O'Reilly,
D. Cielien, A. Meysman, M. Ali, Introducing data science, Manning Publications Co., <http://www.manning.com>
I. White, Hadoop the Definitive Guide, (4th Edition), O'Reilly,
B. Chambers, M. Zaharia, Spark The Definitive Guide, O'Reilly
P.Zececiv, M. Bonaci, Spark In Action, Manning Publications Co., <http://www.manning.com>

Quelques sites intéressants :

L'environnement de la machine d'apprentissage pour flux de données MOA (Massive Online Analysis)
<https://moa.cms.waikato.ac.nz/>

L'environnement de la machine d'apprentissage Weka-Spark
<http://www.cs.waikato.ac.nz/ml/weka/>

Le site de téléchargement de spark
<http://spark.apache.org/>

Le site du cours. Les notes de cours seront disponibles sur moodle_
<https://moodle.uqac.ca/>

Annexe 1 –

Sem. / Séance	Contenu : calendrier approximatif des activités (L'ordre et le contenu peuvent être modifiés)
30/08 Présentiel	Vue d'ensemble de l'apprentissage automatique pour le BigData • Le BigData et la science des données ? • Problématique du passage à l'échelle ? • Qu'est-ce que l'apprentissage automatique à large échelle • Démonstration avec PySpark, MOA
06/09 Virtuel	Environnement d'apprentissage automatique : BigData • Initiation à l'aide d'un tutoriel dans un laboratoire • Étude comparative des outils d'apprentissage en ligne PySpark • Une étude de cas d'apprentissage automatique de BigData de bout en bout
13/09 Présentiel	Apprentissage supervisé distribué • Distribuer les données sur plusieurs processeurs, • MapReduce et l'apprentissage parallèle • Apprentissage parallèle d'un arbre de décision • Expérimentation avec PySpark,
20/09 Virtuel	Apprentissage par la recherche de similarité • Recherche k plus proches voisins • Hachage sensible à la similarité
27/09 Présentiel	Application : systèmes de recommandation • Réduction de dimensionnalité • Décomposition-UV
04/10 Virtuel	Apprentissage supervisée sur flux de données stationnaires • Architecture d'un système de flux de données • Échantillonnage et filtrage dans un flux • Construction d'un arbre de décision en ligne • Algorithme Hoeffding Tree (VFDT) et ses applications • Test et mesure de performance en ligne, MOA
11/10	Semaine de relâche
	Apprentissage supervisée sur flux de données non-stationnaires

18/10 Présentiel	- Fenêtres sur un Flux - Dérive de concepts - Algorithme cVFDT et ses applications, MOA
25/10 Virtuel	Consultations pour les mini-projets
01/11 Présentiel	Apprentissage non-supervisé distribué - Retour sur K-means, K-means++ - Kmeans-MapReduce - Application avec PySpark
09/11 Virtuel	Apprentissage non-supervisé pour flux de données - Algorithmes BIRCH, CluStream, DenStream - Application avec MOA
16/11 Présentiel	Motifs fréquents - L'algorithme FP-Growth - Faire face au volume de données en mémoire - Algorithmes à passes limitées - Comptage de motifs fréquents dans un flux de données
23/11 Virtuel	Machine à vecteurs de support (SVM) - Séparateurs à large marge : SVM - Distribution du SVM - Exemple d'application avec PySpark
30/11 Présentiel	Ensemble de classeurs : - Combinaison de classeurs - Ensemble de classeurs distribués - Expérimentation avec PySpark
06/12 Virtuel	Présentation de projets
13/12 Virtuel	Présentation de projets

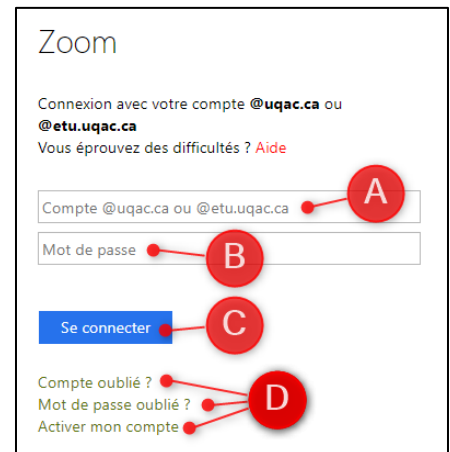
Note importante : Une licence Zoom Pro est attribuée automatiquement à toute la communauté UQAC. Vous n'avez donc pas besoin de faire de demande de création de compte.

1. Accéder au site Web suivant <https://uqac.zoom.us/> ;
2. La fenêtre suivante s'ouvrira, cliquer sur le bouton « **Connexion** » :



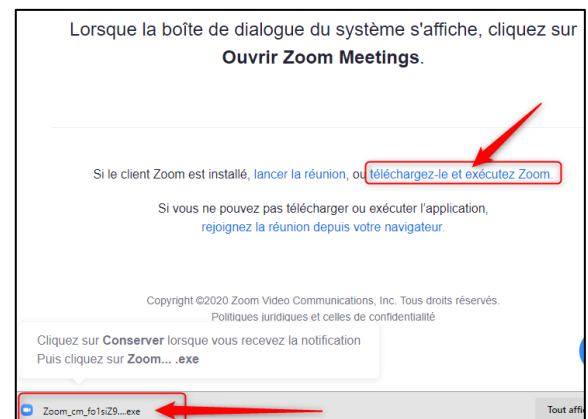
3. Par la suite, vous allez être redirigé vers le portail **UQAC**.
Utiliser vos informations de connexion avec votre compte UQAC.

- A) Dans le premier champ, vous devez entrer **votre compte UQAC employé** (identifiant court suivi de @uqac.ca). Exemple : j1smith@uqac.ca;
- B) Dans le deuxième champ, entrer votre **mot de passe** (le même que vous utilisez pour accéder à votre **Intranet** et à votre **portail employé** de l'UQAC);
- C) Cliquer sur le bouton « **Connexion** »;
- D) Si vous ne vous souvenez plus de votre compte UQAC et/ou de votre mot de passe, cliquer sur le lien « **Compte oublié ?** » ou sur « **Mot de passe oublié ?** » et suivre les indications qui s'afficheront à l'écran. Si vous êtes un employé nouvellement embauché à l'UQAC et que vous n'avez jamais accédé à votre compte Intranet et/ou votre portail employé, vous devez, au préalable, cliquer sur le lien « **Activer mon compte** ».



Si c'est la **PREMIÈRE** fois que vous tentez de vous brancher à Zoom, effectuer l'étape #4 sinon, aller directement à l'étape #5.

4. La **PREMIÈRE** fois que vous tenterez de vous brancher à Zoom, la fenêtre suivante s'affichera et vous devrez installer Zoom sur votre ordinateur. Pour ce faire, cliquer sur le lien « **téléchargez-le et exécutez Zoom** », le fichier d'installation s'affichera dans le coin inférieur gauche de votre écran, cliquer dessus et suivre les indications à l'écran :

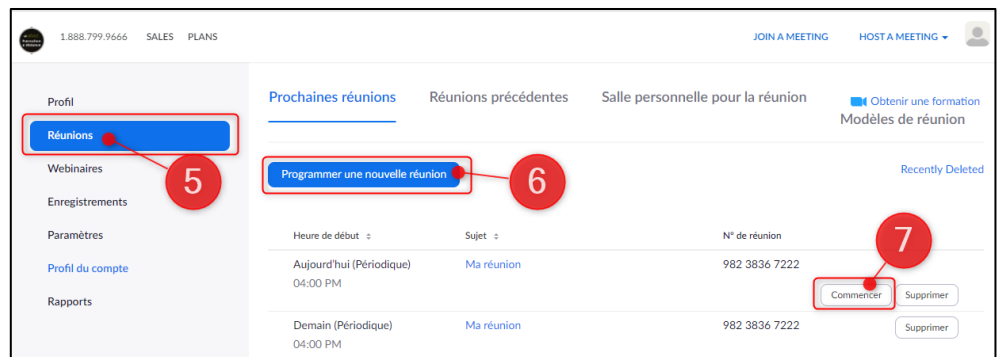


5. La fenêtre suivante s'affichera, cliquer sur « **Réunions** »;

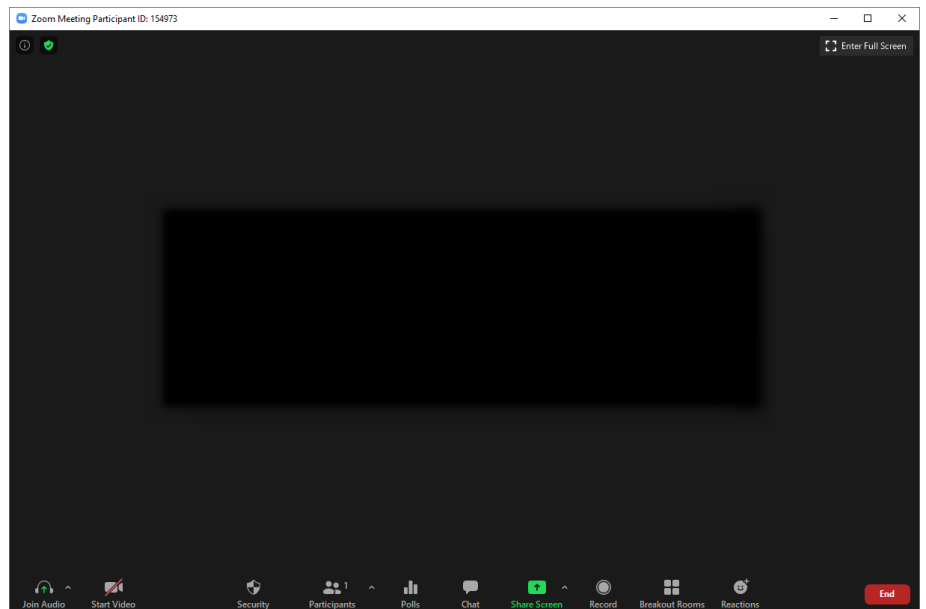
6. Vous pourrez, si vous le souhaitez, programmer une nouvelle réunion en cliquant sur le bouton « **Programmer une nouvelle réunion** »;

7. Vous serez également en mesure de vous brancher à une réunion en cliquant sur le bouton « **Commencer** » vis-à-vis la réunion de votre choix;

8. Si vous cliquez sur le bouton « **Commencer** » une réunion, la boîte suivante s'affichera, cliquer sur « **Ouvrir Zoom Meetings** » :



9. Vous serez alors branché à la réunion :



Si vous avez des questions, commentaires ou suggestions, n'hésitez pas à communiquer PAR COURRIEL avec le Centre de soutien à la prestation virtuelle de cours : ✉ prestation_virtuelle@uqac.ca