Université de Sherbrooke

Département d'informatique

IFT 599 / IFT 799 – Science des données

Plan d'activité pédagogique

Automne 2024

Enseignant	Shengrui Wang
Courriel :	Shengrui.Wang@USherbrooke.ca
Local :	D4-1018-1
Téléphone :	+1 819 821-8000 x62022
Disponibilités :	Officiellement, les mardis entre 14h et 16h. Les personnes étudiantes peuvent venir poser des questions en tout temps en personne au bureau.

Site web du cours : MS Teams

Horaire

Groupe 1 :	Exposé magistral :	Mardi	10h30 à 12h20	salle D4-2025
		Vendredi	13h30 à 14h20	salle D3-2030
Groupe 18 :	Exposé magistral :	Lundi	9h00 à 11h50	salle L1-11615/L1-4660/L1-11620/L1-4600/L1-4

Description officielle de l'activité pédagogique[1]

Cibles de formation :	Comprendre et maîtriser des théories et méthodes de base pour la science des données.
Contenu :	Inférence statistique : procédures statistiques fondamentales, estimation des paramètres d'un modèle, tests des hypothèses liées aux caractéristiques structurelles d'un modèle, intervalle de confiance pour les paramètres de modèle. Techniques de forage de données : analyse exploratoire des données, prétraitement, visualisation, recherche et extraction des règles d'association, classification et prédiction, analyse de regroupement. Recherche d'information : principe, concepts de base, indexation, engins de recherche. Applications dans divers domaines tels que la santé, l'intelligence d'affaires, les réseaux sociaux et la finance.
Crédits	3
Organisation	3 heures d'exposé magistral par semaine 6 heures de travail personnel par semaine
Préalable	IFT 436, STT 418
Particularités	Aucune

[1] https://www.usherbrooke.ca/admission/fiches-cours/ift599

1 Présentation

Cette section présente les cibles de formation spécifiques et le contenu détaillé de l'activité pédagogique. Cette section, non modifiable sans l'approbation du comité de programme du Département d'informatique, constitue la version officielle.

1.1 Mise en contexte

De nos jours, le volume de plus en plus important de données disponibles (que l'on appelle données massives, mégadonnées ou bien « big data ») et leur démocratisation combinée avec les puissances de calcul et de stockage, nous amènent vers une nouvelle ère d'analyse des données qui comporte beaucoup de défis. En effet, les mégadonnées qui ne cessent d'augmenter jour après jour sont souvent peu exploitées, alors qu'elles cachent des connaissances décisives face au marché et à la concurrence. La plupart des techniques d'analyse de données traditionnelles échouent à extraire de l'information pertinente sur ce type de données. Pour combler ce besoin, la science des données s'est rapidement développée ces dernières années. Elle s'intéresse à tous les aspects liés aux données notamment la collecte, le nettoyage, le filtrage, l'anonymisation, le cryptage, la visualisation, le stockage, le transfert, l'analyse et la modélisation (classification et prédiction), etc. L'appellation « science des données » est issue de celle du forage de données (« Data Mining ») qui a été le porteur de la discipline avec l'apprentissage automatique (« machine learning ») depuis une vingtaine d'années. La science des données est donc l'ensemble des algorithmes et méthodes destinés à l'exploration et l'analyse de grandes bases de données informatiques en vue de détecter dans ces données des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières restituant de façon concise l'essentiel de l'information utile, et de prédire les résultats des processus sous jacents afin de générer des valeurs pour la société.

1.2 Cibles de formation spécifiques

Ce cours vise à initier l'étudiante et l'étudiant aux concepts fondamentaux de sciences de données et d'appliquer ces notions à des problèmes concrets. À la fin de cette activité pédagogique, l'étudiante ou l'étudiant sera capable de :

1. comprendre les principaux concepts de la science des données et le processus d'extraction des connaissances dans les bases de données ;
2. maîtriser des concepts et techniques d'inférence statistique ;
3. maîtriser les techniques d'analyse des données transactionnelles, plus particulièrement la recherche des règles d'association ;
4. maîtriser des techniques de « clustering » et de classification les plus utilisées dans la pratique ;
5. maîtriser des techniques de prétraitement des données et de réduction de la dimension (essentiellement la gestion de la redondance et la sélection des attributs) ;
6. comprendre la problématique des données de grande dimension et maîtriser des techniques d'agrégation utilisées sur ce type de données ;
7. se familiariser avec des algorithmes de classement des pages Web tels que HITS et PageRank ;
8. se familiariser avec des algorithmes pour les systèmes de recommandation.
9. comprendre le lien qu'il y a entre la sciences des données et divers applications en lien avec l'informatique de gestion.

1.3 Contenu détaillé

Thème	Contenu	Nbr. d'heures	Objectifs	Travaux	Lectures
1	Concepts de base : Processus d'extraction des connaissances dans les bases de données ; Prétraitement des données ; Mesures de similarités ; Techniques de réduction de la dimension (ACP : Analyse en composantes principales) ; Techniques de sélection d'attributs ; Concepts de base pour la détection des anomalies ; Visualisation ; Introduction de Weka (progiciel d'analyse des données).	8	1 et 5	✔	[2] [3] chapitres 1 et 2
2	Analyse d'association et analyse des séquences : Introduction aux données transactionnelles ; Principe de l'algorithme *Apriori* et, si le temps permet, l'algorithme *FP-growth* ; Recherche et extraction des règles d'associations ; Extraction et utilisation des patrons significatifs pour l'analyse des séquences.	8	3	✔	[2] [3] chapitres 5 et 6
3	Inférence statistique : Procédures statistiques fondamentales ; Estimation des paramètres d'un modèle ; Tests des hypothèses liées aux caractéristiques structurelles d'un modèle ; Intervalle de confiance pour les paramètres de modèle.	4	1 et 2		[2]
4	Agrégation (méthodes de base) : Clustering par partition et hiérarchique ; Agrégation basé sur la densité ; Agrégation des données catégoriques et transactionnelles ; Validation des résultats d'agrégation.	6	4 et 5		[2] [3] chapitre 7
5	Agrégation (méthodes avancées) : Détection des anomalies ; Agrégation des données de grandes dimensions et des données complexes.	6	4, 5 et 6	✔	[2] [3] chapitre 8 [4]
6	Systèmes de recommandation : Principes et éléments de base ; Principe des algorithmes basés sur le contenu ; Approches basées sur le filtrage collaboratif ; Méthodes basées sur la ressemblance directe ; Méthodes basées sur la sémantique latente.	6	5 et 8	✔	[2] [1]
7	Exploration du Web : Principes et éléments de bases ; Algorithmes de classement des pages Web (HITS et PageRank) ; Extraction des communautés dans le Web.	2	7		[2]

1. Les lectures indiquées ne sont là qu'à titre indicatif. L'enseignant est libre de choisir un autre document de référence.

2 Organisation

Cette section propre à l'approche pédagogique de chaque enseignante ou enseignant présente la méthode pédagogique, le calendrier, le barème et la procédure d'évaluation ainsi que l'échéancier des travaux. Cette section doit être cohérente avec le contenu de la section précédente.

2.1 Méthode pédagogique

Ce cours est constitué de cours magistraux en raison de 3 heures par semaine. Les principaux documents du cours sont des diapositives de l'enseignant qui sont assez détaillées pour présenter les concepts et les méthodes enseignés. Les diapositives de chaque thème seront mises à la disposition des personnes étudiantes à l'avance. Les matières théoriques et les exercices seront étroitement intégrés, tandis que les travaux pratiques comporteront non seulement l'application des méthodes enseignées pour résoudre un problème, mais aussi la recherche ou la conception de nouvelles méthodes. Ces travaux pratiques jouent un rôle très important dans le développement de la capacité des personnes étudiantes pour résoudre des problèmes réels en sciences données.

2.2 Calendrier

Semaine	Date	Thème	Travaux pratiques
1	2024-08-26	1	
2	2024-09-02	1	
3	2024-09-09	1 et 3	
4	2024-09-16	3	
5	2024-09-23	3 et 4	
6	2024-09-30	4	Remise TP1
7	2024-10-07	4	
8	2024-10-14	Examen périodique	
9	2024-10-21	Relâche	
10	2024-10-28	5	
11	2024-11-04	5	
12	2024-11-11	6	
13	2024-11-18	6	Remise TP2
14	2024-11-25	2	
15	2024-12-02	Révision et 2	Remise TP3
16	2024-12-09	Examen final	
17	2024-12-16	Examen	

2.3 Évaluation

Travaux pratiques (3)	25 %
Examen intra	30 %
Examen final	45 %

Pour les examens, tout document papier est permis. L'examen final est cumulatif, toutefois, l'accent sera mis sur les matières enseignées après la semaine de relâche. Tel que précisé ici, il y aura trois travaux pratiques (TPs) pour lesquels des efforts importants seront nécessaires pour trouver des solutions (surtout pour les TP1 et TP2). Les travaux peuvent se réaliser en solo ou en équipe de 2 ou 3 personnes. Un rapport de projet sera exigé pour chaque TP. Les critères d'évaluation pour les travaux et examens sont principalement basés sur la structure et clarté du code source, l'exactitude et la précision des réponses aux questions théoriques ainsi que la pertinence des solutions proposées aux problèmes énoncés.

2.3.1 Qualité de la langue et de la présentation

Conformément à l'article 17 du règlement facultaire d'évaluation des apprentissages[2] l'enseignante ou l'enseignant peut retourner à l'étudiante ou à l'étudiant tout travail non conforme aux exigences quant à la qualité de la langue et aux normes de présentation.

2.3.2 Plagiat

Le plagiat consiste à utiliser des résultats obtenus par d'autres personnes afin de les faire passer pour sien et dans le dessein de tromper l'enseignante ou l'enseignant. Vous trouverez en annexe un document d'information relatif à l'intégrité intellectuelle qui fait état de l'article 9.4.1 du Règlement des études[3]. Lors de la correction de tout travail individuel ou de groupe une attention spéciale sera portée au plagiat. Si une preuve de plagiat est attestée, elle sera traitée en conformité, entre autres, avec l'article 9.4.1 du Règlement des études de l'Université de Sherbrooke. L'étudiante ou l'étudiant peut s'exposer à de graves sanctions qui peuvent être soit l'attribution de la note E ou de la note zéro (0) pour un travail, un examen ou une activité évaluée, soit de reprendre un travail, un examen ou une activité pédagogique. Tout travail suspecté de plagiat sera transmis au Secrétaire de la Faculté des sciences. Ceci n'indique pas que vous n'ayez pas le droit de coopérer entre deux équipes, tant que la rédaction finale des documents et la création du programme restent le fait de votre équipe. En cas de doute de plagiat, l'enseignante ou l'enseignant peut demander à l'équipe d'expliquer les notions ou le fonctionnement du code qu'elle ou qu'il considère comme étant plagié. En cas d'incertitude, ne pas hésiter à demander conseil et assistance à l'enseignante ou l'enseignant afin d'éviter toute situation délicate par la suite.

2.4 Échéancier des travaux

Travaux pratiques	Sujet	Réception	Remise	Points
TP1	Projection des données	2024-09-13	2024-10-01	10
TP2	Analyse des données par regroupement	2024-10-29	2024-11-18	10
TP3	Système recommendation	2024-11-19	2024-12-03	5

2.4.1 Directives particulières

Pour chaque TP, il y aura au moins deux semaines entre la distribution de la description du TP et la remise du TP. Aucun délai ne sera autorisé.

2.5 Utilisation d'appareils électroniques et du courriel

Selon le règlement complémentaire des études, section 4.2.3 [4], l'utilisation d'ordinateurs, de cellulaires ou de tablettes pendant une prestation est interdite à condition que leur usage soit explicitement permise dans le plan de cours.

Dans ce cours, l'usage de téléphones cellulaires, de tablettes ou d'ordinateurs est autorisées. Cette permission peut être retirée en tout temps si leur usage entraîne des abus.

Tel qu'indiqué dans le règlement universitaire des études, section 4.2.3[5], toute utilisation d'appareils de captation de la voix ou de l'image exige la permission de la personne enseignante.

Note : Je réponds aux questions posées par courriel à l'extérieur des périodes de cours.

Bien que l'enseignant réponde à des questions posées par courriel ou par Teams, il sera plus disposé de répondre aux questions en personne.

[2] https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Etudiants_actuels/Informations_academiques_et_reglements/2017-10-27_Reglement_facultaire_-_evaluation_des_apprentissages.pdf

[3] https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

[4] https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Etudiants_actuels/Informations_academiques_et_reglements/Sciences_Reglement_complementaire.pdf

[5] https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

3 Matériel nécessaire pour l'activité pédagogique

4 Références

[1] DESROSIERS, CHRISTIANAND KARYPIS, GEORGE : *A Comprehensive Survey of Neighborhood-based Recommendation Methods*, pages 107–144. Springer US, Boston, MA, 2011.

[2] SHENGRUI WANG : Acétates du cours et des documents supplémentaires .

[3] TAN, PANG-NING AND STEINBACH, MICHAEL AND KUMAR, VIPIN : *Introduction to Data Mining, (Second Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2018. Ce livre contient plus que 65% de la matière du cours (thèmes : 1, 2, 3 et 4). Voir le site web du livre`http://www-users.cs.umn.edu/~kumar/dmbook/index.php`.

[4] WU, SHU AND WANG, SHENGRUI : Information-theoretic outlier detection for large-scale categorical data. *IEEE transactions on knowledge and data engineering*, 25(3):589–602, 2013.

UNIVERSITÉ DE SHERBROOKE

L'intégrité intellectuelle passe, notamment, par la reconnaissance des sources utilisées.

À l'Université de Sherbrooke, on y veille!

Extrait du *Règlement des études* (Règlement 2575-009)

9.4.1 DÉLITS RELATIFS AUX ÉTUDES

Un délit relatif aux études désigne tout acte trompeur ou toute tentative de commettre un tel acte, quant au rendement scolaire ou une exigence relative à une activité pédagogique, à un programme ou à un parcours libre.

Sont notamment considérés comme un délit relatif aux études les faits suivants :

a) commettre un plagiat, soit faire passer ou tenter de faire passer pour sien, dans une production évaluée, le travail d'une autre personne ou des passages ou des idées tirés de l'œuvre d'autrui (ce qui inclut notamment le fait de ne pas indiquer la source d'une production, d'un passage ou d'une idée tirée de l'œuvre d'autrui);

b) commettre un autoplagiat, soit soumettre, sans autorisation préalable, une même production, en tout ou en partie, à plus d'une activité pédagogique ou dans une même activité pédagogique (notamment en cas de reprise);

c) usurper l'identité d'une autre personne ou procéder à une substitution de personne lors d'une production évaluée ou de toute autre prestation obligatoire;

d) fournir ou obtenir toute aide non autorisée, qu'elle soit collective ou individuelle, pour une production faisant l'objet d'une évaluation;

e) obtenir par vol ou toute autre manœuvre frauduleuse, posséder ou utiliser du matériel de toute forme (incluant le numérique) non autorisé avant ou pendant une production faisant l'objet d'une évaluation;

f) copier, contrefaire ou falsifier un document pour l'évaluation d'une activité pédagogique;

[…]

Par plagiat, on entend notamment :

- Copier intégralement une phrase ou un passage d'un livre, d'un article de journal ou de revue, d'une page Web ou de tout autre document en omettant d'en mentionner la source ou de le mettre entre guillemets;
- reproduire des présentations, des dessins, des photographies, des graphiques, des données… sans en préciser la provenance et, dans certains cas, sans en avoir obtenu la permission de reproduire;
- utiliser, en tout ou en partie, du matériel sonore, graphique ou visuel, des pages Internet, du code de programme informatique ou des éléments de logiciel, des données ou résultats d'expérimentation ou toute autre information en provenance d'autrui en le faisant passer pour sien ou sans en citer les sources;
- résumer ou paraphraser l'idée d'un auteur sans en indiquer la source;
- traduire en partie ou en totalité un texte en omettant d'en mentionner la source ou de le mettre entre guillemets ;
- utiliser le travail d'un autre et le présenter comme sien (et ce, même si cette personne a donné son accord);
- acheter un travail sur le Web ou ailleurs et le faire passer pour sien;
- utiliser sans autorisation le même travail pour deux activités différentes (autoplagiat).

Autrement dit : mentionnez vos sources

Document informatif V.3 (août 2017)