Université du Québec en Outaouais **Département d'informatique et d'ingénierie**

Sigle : INF5103 Gr. 01
Titre : Concepts statistiques pour la science des données
Session : Hiver 2023 Horaire et local
Professeur : Allili, Mohand Said

1. Description du cours paraissant à l'annuaire :

Objectifs

Maîtriser les concepts statistiques avancés utilisés dans la science des données. Développer les connaissances pour développer des algorithmes d'analyse et de prédiction à partir des données en utilisant ces concepts statistiques.

Contenu

Rappels sur les concepts de probabilités et statistiques. Statistique et science des données. Modèles paramétriques vs. non paramétriques. Techniques d'échantillonnage des données et estimation de paramètres. Modèles statistiques pour la classification et la régression. Tests d'hypothèses. Propriétés d'un paramètre statistique: biais, consistance, efficacité. Maximum de vraisemblance. Statistique bayésienne. Analyse factorielle et analyse de variance. Réduction de dimensions. Modèles graphiques probabilistes. Méthodes de Monte-Carlo. Réduction du biais statistique dans l'analyse de données. Études d'applications (ex. régression, classification, ordonnancement, etc.).

2. Objectifs spécifiques du cours :

Au terme de cette activité, l'étudiant(e) aura acquis des connaissances suivantes:

- Les principes de la science des données;
- Les concepts statistiques importants pour la science des données
- Les statistiques descriptives;
- Les statistiques inférentielles;
- L'apprentissage automatique à partir de données;

3. Stratégies pédagogiques :

Séances de cours en présentiel, de 3h/semaine comprenant une ou plusieurs stratégies :

- Cours magistral

Évaluations :

- Devoirs
- Examen intra (en présentiel, à livre ouvert)
- Examen final (en présentiel, à livre ouvert)

4. Heures de disponibilité ou modalités pour rendez-vous :

Sur rendez-vous.

5. Plan détaillé du cours sur 15 semaines :

Semaine	Thèmes	Dates
1	**Généralités** • Phénomène des mégadonnées (big data). • Données structurées et non-structurées. • Principes de la science des données. • Statistique pour la science des données.	12 Janv. 2023
2	**Éléments de cucul des probabilités** • Éxpérience aléatoire • Variables aléatoires et propriétés • Loi de Bayes et loi marginale • Lois de probabilités discrètes & continues.	19 Janv. 2023
3	**Statistique descriptive** • Prétraitement des données. • Population versus échantillon. • Techniques d'échantillonnage. • Mesures de tendance des données. • Visualisation des données. **Travail pratique I**	26 Janv. 2023
4	**Statistique inférentielle** • Modélisation à partir de données • Modèles paramétriques et non-paramétriques • Propriétés des paramètres : biais, consistance, efficacité.	02 Fév. 2023
	Estimation de paramètres • Estimation par intervalle de confiance. • Tests d'hypothèses statistiques • Méthode du maximum de vraisemblance • Estimation Bayésienne **Travail pratique II**	09 Fév. 2023
6	**Modèles statistiques supervisés I (régression)** • Principes de la régression • Régression linéaire et non-linéaire • Concept de sur-apprentissage • Concept de validation croisée	16 Fév. 2023
7	**Modèles statistiques supervisés II (classification)** • Méthode du Bayes naïf pour la classification. • Méthode de la régression logistique	23 Fév. 2023

	Travail pratique III	
8	**Examen intra**	02 Mars 2023
9	**Semaine d'étude**	09 Mars 2023
10	**Modèles non-paramétriques supervisés** • Distributions non-paramétriques de données. • Distributions de données temporelles (chaines de Markov). • Modèles supervisées basés sur les réseaux de neurones.	16 Mars 2023
11	**Modèles statistiques non-supervisés I (regroupement)** • Méthode des K-moyennes • Méthode des mélanges Gaussiens • Validation de groupes **Travail pratique IV**	23 Mars 2023
12	**Modèles statistiques non-supervisés I (Reduction de dimensions)** • Analyse à composantes principales • Analyse factorielle • Auto-encoders	30 Mars 2023
13	**Étude de quelques applications** • Classification de documents multimédias • Détection d'intrusions en cybersécurité • Analyse de risques • Détection d'anomalies	06 Avr. 2023
14	**Présentation des projets**	13 Avr. 2023
	Examen final	20 Avr. 2023

6. Évaluation du cours :

L'étudiant(e) dans ce cours sera évalué(e) par les examens de mi-session et final, ainsi que par des projets de session. La pondération de la note finale se fera comme suit :

- **Devoirs : 15 %**
- **Projets : 25 %**
- **Examen de mi-session : 30 %**
- **Examen final : 30 %**

Pour les projets, l'évaluation sera répartie comme suit :

- **Devoirs : 15 points :** Il s'agit d'implanter quelques méthodes statistiques pour la science des données en utilisant le langage Python.
- **Projets : 25 points :** Il s'agit de faire des projets sur les techniques statistiques pour l'analyse de données. Ces derniers peuvent être dans la description, l'analyse ou la modélisation statistique des données. Ils peuvent aussi être reliés l'évaluation statistique des techniques d'apprentissage automatique. Quelques projets seront fournis par le professeur. Néanmoins, les étudiants auront la liberté de choisir des projets dans des domaines d'applications qu'ils souhaitent. L'évaluation se fera sur deux volets :

    - Un rapport de 15 pages. Les normes de présentation de travaux (p. ex. page de garde, marge d'un pouce, interligne à 1,5, taille des caractères de 12 points) doivent être **absolument** respectées.
    - Une présentation de 15 à 20 minutes sera faite en classe. Les projets se feront individuellement ou en équipes de deux si le nombre d'étudiants inscrits dépasse 10.

Une moyenne générale inférieure à **64 %** est éliminatoire et conduit automatiquement à l'échec de l'étudiant(e).

Tout retard dans la remise d'un travail entraîne une pénalité de **15 %/**jour sur la note attribuée à ce travail, **jusqu'au maximum d'une semaine.** La qualité du français sera considérée lors de la correction des travaux.

uiue

- Politique du département d'informatique et d'ingénierie relative à la tenue des examens
- Note sur le plagiat et sur la fraude
- Politique relative à la qualité de l'expression française écrite chez les étudiants et les étudiantes de premier cycle à l'UQO
- Absence aux examens : cadre de gestion, demande de reprise d'examen (formulaire)

La communauté universitaire s'engage à lutter contre les inconduites, le harcèlement et les violences à caractère sexuel. Dénonçons toute forme de violence.

Ensemble, accomplissons un pas de plus en complétant la formation obligatoire en ligne : "La banalisation des violences à caractère sexuel".

uqo.ca/bimi/formation-obligatoire

Pour de plus amples renseignements consultez :

bimi@uqo.ca

UQO BUREAU D'INTERVENTION EN MATIÈRE D'INCONDUITE

8. Principales références :

1. Stanley H. Chan. Introduction to Probability for Data Science. Michigan Publishing Services, 2021. (https://probability4datascience.com/index.html)
2. Peter Bruce, Andrew Bruce, and Peter Gedeck. Practical Statistics for Data Scientists. O'Reilly books, 2020.
3. P-N. Tan, M. Steinbach, A. Karpatne, V. Kumar. Introduction to Data Mining (2nd Edition), Pearson 2018.
4. Witten et al. Data mining : Practical machine learning Tools and Techniques. Morgan Kaufmann, 2017.
5. K-P. Murphy. Probabilistic machine Learning. MIT press, 2022.

9. Page Web du cours :

https://moodle.uqo.ca