

Université de
Sherbrooke

Département d'informatique
IFT 603 / IFT 712 – Techniques d'apprentissage
Plan d'activité pédagogique
Automne 2021

Enseignant

Pierre-Marc Jodoin

Courriel : pierre-marc.jodoin@usherbrooke.ca

Local : D4-1016-1

Téléphone : +1 819 821-8000 x62025

Disponibilités : Du lundi au vendredi de 9 h à 17 h. Il est cependant préférable de prendre rendez-vous avec moi.

Responsable(s) : Direction du département**Site web du cours** : <http://info.usherbrooke.ca/pmjodoin/cours/ift603>

Horaire

Exposé magistral :	Judi	13h30 à 15h20	salle D3-2035
	Vendredi	11h30 à 12h20	salle D3-2040

Description officielle de l'activité pédagogique¹

Cibles de formation : S'initier aux techniques d'apprentissage automatique en intelligence artificielle.

Contenu : Concepts fondamentaux de l'apprentissage automatique, formulation mathématique probabiliste, régression linéaire, classification linéaire, astuce du noyau, machine à vecteurs de support, apprentissage bayésien, modèle de mélange de gaussiennes, algorithme EM, réduction de dimensionnalité et combinaison de modèles.

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/ift603>

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

L'apprentissage automatique ou l'apprentissage par machine (Machine Learning) s'intéresse à la conception, l'analyse, l'implémentation et l'application de programmes d'ordinateur capables de s'améliorer, au fil du temps, soit sur la base de leur propre expérience, soit à partir des données antérieures fournies par d'autres programmes. De nos jours, l'apprentissage automatique joue un rôle essentiel dans de nombreux domaines d'applications, tels que la vision par ordinateur, le traitement automatique du langage, la reconnaissance vocale, les systèmes tutoriels intelligents, la modélisation de l'utilisateur, la robotique, la bio-informatique, les finances, le marketing, les jeux vidéo, la télédétection, etc. En fait, la plupart des programmes de l'intelligence artificielle contiennent un module d'apprentissage. Presque tous les systèmes de reconnaissance de formes sont basés sur des techniques d'apprentissage.

1.2 Cibles de formation spécifiques

Ce cours vise à introduire les concepts fondamentaux de l'apprentissage automatique et à présenter une vaste gamme d'approches différentes utilisées dans des systèmes d'apprentissage automatique. À la fin de cette activité pédagogique, l'étudiante ou l'étudiant sera capable de :

1. Comprendre les notions fondamentales de l'apprentissage automatique ;
2. Comprendre les formulations probabilistes de l'apprentissage automatique ;
3. Maîtriser des algorithmes de régression et classification en apprentissage supervisé ;
4. Comprendre comment utiliser l'astuce du noyau ;
5. Maîtriser les fondements des réseaux de neurones et leur apprentissage par rétro propagation ;
6. Maîtriser des algorithmes de combinaison de modèles.

1.3 Contenu détaillé

Thème	Contenu	Nbr. d'heures	Objectifs	Travaux	Lectures
1	Mathématiques pour l'I.A. : algèbre linéaire de base ; loi gaussienne multivariée, marginale et conditionnelle d'une gaussienne, distance de Mahalanobis, matrice de covariance, valeurs propres et vecteur propres, espérance mathématique ; distribution de vraisemblance, a posteriori, a priori et évidence ; lagrangien.	4	1	✓	[1] sections 1.2, 1.2.1, 1.2.4, 2.3, 2.3.1, 2.3.2 et annexes B, C et E
2	Concepts fondamentaux : apprentissage supervisé vs non-supervisé ; performance d'entraînement vs généralisation, sur- et sous-entraînement ; régularisation ; sélection d'hyper-paramètres et validation croisée ; minimisation d'une fonction de perte, argmin ; exemple de la régression polynomiale ; malédiction de la dimensionalité.	3	1	✓	[1] sections 1.0, 1.1, 1.2.4, 1.3, 2.1, 2.3 et appendice C
3	Formulation probabiliste : théorie des probabilités (cas discret et continu) ; théorie de l'information (entropie, information mutuelle, divergence de Kullback-Leibler) ; bayes et probabilités jointes, conditionnelles, a priori ; introduction au maximum de vraisemblance ; introduction au maximum a posteriori.	4	1 et 2	✓	[1] sections 1.2, 1.2.1, 1.2.2, 1.2.4, 1.2.5, 1.6 et 1.6.1
4	Régression linéaire : formulation probabiliste de la régression polynomiale ; apprentissage par maximum de vraisemblance ; apprentissage par maximum a posteriori.	3	3	✓	[1] sections 3.1, 3.1.1, 3.1.4, 3.1.5 et 3.2
5	Classification linéaire : classification comme un problème de régression ; méthode de Fisher 2 classes ; méthodes génératives ; perceptron binaire vs multiclasse ; régression logistique binaire vs multiclasse ; descente de gradient ; sigmoid vs softmax ; distribution de Bernoulli vs entropie croisée.	6	1, 3 et 5	✓	[1] sections 4.1, 4.1.2, 4.1.3, 4.1.4, 4.2 et 4.3
6	Méthode à noyau : représentation duale et régression à noyau ; construction de noyau ; régression à noyau.	3	4	✓	[1] sections 6.1 et 6.2
7	Méthode à noyau parcimonieuse : classifieur à marge maximale ; machine à vecteurs de support (SVM).	3	3 et 4	✓	[1] sections 7.0, 7.1, 7.1.1 et 7.1.2
8	Réseaux de neurones multi-couches : approche linéaire vs non linéaire ; classification binaire et multiclasse ; fonction d'activation ; couche d'entrée vs couche cachée vs couche de sortie ; dérivée en chaîne, rétro-propagation, problème avec les fonctions d'activation non centrées à zéro, utilisation de mini-batches, disparition du gradient, <i>Dropout</i> ; initialisation d'un réseau ; <i>sanity checks</i> et vectorisation.	6	5	✓	[1] sections 5.1, 5.2, 5.2.1, 5.2.4, 5.3, 5.3.1, 5.3.2, 5.5
9	Combinaison de modèles : <i>bootstrap</i> , <i>boosting</i> .	3	6		[1] sections 14.0, 14.2, 14.3, 14.3.1
10	Théorie de la décision : décomposition biais-variance ; théorie de la décision.	3	1, 2 et 3		[1] sections 1.5.5 et 3.2
11	Mélange de gaussiennes (si le temps le permet) : partitionnement de données (<i>clustering</i>) ; algorithme EM ; fenêtres de Parzen ; EM Spectral Clustering.	5	2		[1] sections 9.2 et 9.4

1. Le cours doit comprendre au moins quatre travaux pratiques couvrant tous les sujets marqués «✓» dans le tableau.
2. 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

Une semaine comprend une à trois heures de cours magistral, ainsi qu'une à trois heures de **vidéo devant être regardés à la maison**. Il est à noter que les vidéos contiennent certaines des parties les plus importantes du cours. Il est donc fondamentalement important que les étudiantes et étudiants suivent les vidéos au rythme mentionné sur le site web du cours. Les présentations magistrales auront pour objectif de clarifier le contenu mathématique des notions introduites dans les vidéos, donner des illustrations concrètes et faire des exercices. Les liens vers les vidéos et les diapositives y étant associées (format PDF) sont disponibles sur le site web du cours. Tout au long de la session, les élèves devront implémenter les notions vues en classe à travers quatre travaux pratiques, ainsi qu'un projet de session pour les étudiantes et étudiants d'IFT712. Des instructions particulières seront données pour chacun des travaux et les périodes de consultation seront déterminées en classe. En ce qui concerne les examens, seules des notes manuscrites non photocopiées sont admises, tout manuel et notes de cours imprimées étant interdits. L'examen final est récapitulatif.

Compte tenu du contexte actuel (pandémie due au COVID-19), il se peut que le cours ait lieu en totalité ou en partie à distance d'une façon différente de ce qui est énoncé ci-dessus. Notez que vous en serez informés rapidement si tel est le cas.

2.2 Calendrier

Semaine	Date	Thème	Matière vue en classe
1	2021-08-30	1 et 2	Présentation, notation vecteurs-colonnes vs vecteurs-lignes, illustration des probabilités, espérance mathématique et gaussienne, loi gaussienne multivariée, distance de Mahalanobis, matrice de covariance
2	2021-09-06	1	argmin, argmax, entraînement, validation, généralisation, conditionnelle, preuve que l'espérance mathématique d'une gaussienne est μ que sa variance est σ^2 et Σ . Jointe, marginale, Bayes, exemple concret (math vs info), présentation du projet (712)
3	2021-09-13	1 et 3	Jointe, marginale, Bayes, exemple concret (math vs info)(suite). Maximum de vraisemblance (MV) et a posteriori (MAP). Lien avec MV - bruit gaussien et régression L2. Malédiction de la dimensionnalité, preuve que $E[ax+by] = aE[x] + bE[y]$ dans le cas de variables indépendantes
4	2021-09-20	1 et 4	Lien entre MAP - Bruit et a priori Gaussien et régression Ridge, calcul de la dérivée d'une fonction par un vecteur. Illustration de la classification linéaire
5	2021-09-27	3 et 4	Théorie de l'information, entropie, divergence KL, Lagrangien.
6	2021-10-04	3 et 5	(suite) Lagrangien et Critère de Fisher. Régression logistique + entropie croisée : lien avec la divergence KL
7	2021-10-11	6 et 7	Révision et autres exemples de MV vs MAP
8	2021-10-18	Examen périodique	
9	2021-10-25	Relâche	

10	2021-11-01	8	Preuve que W est une normale, notion de distance signée, pré-traitement des données, dérivée en chaîne
11	2021-11-08	8	Dérivée en chaîne (suite), rétro-propagation, problèmes avec les fonctions d'activation non centrées à zéro et disparition du gradient
12	2021-11-15	9	Gradient de la fonction Soft-max, initialisation de Xavier, MC Dropout (mesure de confiance) et apprentissage actif.
13	2021-11-22	10	Combinaison de modèles + analyse théorique d'Adaboost et de bagging.
14	2021-11-29	11	Expectation maximisation
15	2021-12-06	Révision	Révision
16	2021-12-13	Examen final	

2.3 Évaluation

Travaux pratiques	40 %
Examen intra	20 %
Examen final	40 %

IMPORTANT : le barème pour les évaluations décrit ci-haut s'applique SEULEMENT pour les étudiantes et étudiants d'IFT603. Pour les étudiantes et étudiants d'IFT712, le barème pour les évaluations est le suivant :

- Devoirs : 30 %
- Examen intra : 15 %
- Examen final : 30 %
- Projet de session : 25 %

Détails à l'intention de tous les étudiantes et étudiants

TP1 : Régression linéaire (Thème principal 4)

TP2 : Classification linéaire (Thème principal 5)

TP3 : Méthodes à noyau – classification non-linéaire (Thèmes principaux 6,7)

TP4 : Réseaux de neurones (Thème principal 8)

- Les devoirs doivent être effectués par équipe de **deux ou de trois** ;
- Le code rédigé dans le cadre de ce cours doit se faire dans le langage de programmation **Python**. Le code soumis doit être compatible avec (c'est-à-dire exécutable sous) la version de Python installée dans les laboratoires, sous Linux - Ubuntu ;
- Il est fortement recommandé de configurer votre **ordinateur personnel** et de vous en servir pour effectuer les travaux pratiques et le projet. Si vous avez **Windows**, il est suggéré d'installer **VMware** (gratuits pour les étudiantes et étudiants) et une **VM Ubuntu** (disponible dans le répertoire public).
- La qualité du français et de la présentation peut être considérée lors de l'évaluation des travaux ;
- Toute soumission en retard implique une **perte de 10 points par jour** (incluant samedi et dimanche), sauf celles motivées par des raisons valables et conformes au règlement des études (par exemple, maladie avec attestation d'un médecin).

Projet de session (IFT712)

Les étudiantes et étudiants inscrits au cours IFT 712 sont tenus.es de faire un projet de session en python par équipe de **deux ou trois personnes**. Le projet a pour objectif de tester au moins six méthodes de classification sur une base de données Kaggle (www.kaggle.com) avec la bibliothèque Sklearn (scikit-learn.org). Les équipes sont libres de choisir la base de données de leur choix, mais une option simple est celle du challenge de classification de feuilles d'arbres (kaggle.com/c/leaf-classification). Pour ce projet, on s'attend à ce que les bonnes pratiques de validation croisée et de recherche d'hyperparamètres soient mises de l'avant pour identifier la meilleure solution possible pour résoudre le problème.

Le barème de correction est le suivant :

- Qualité du code - Commentaires /10
- Choix de design /10
- Gestion de projet (Git) /10

- Rapport
 - Démarche scientifique /50
 - Analyse des résultats /20

Qualité du code et commentaires

On vous demande du code rédigé 100% en anglais ou en français, mais pas de français ! Votre code doit être bien documenté, respecter le **standard pep8** et respecter un standard uniforme pour la nomenclature des variables, des noms de fonctions et des noms de classes. Évitez également les variables « hardcodées » empêchant l'utilisation de votre programme sur un autre ordinateur que le vôtre.

Choix de design

Vous devez organiser votre code de façon professionnelle. Pour ce faire, on s'attend à une hiérarchie de classes cohérente, pas seulement une panoplie de fonctions disparates. Aussi, du code dans un script « qui fait tout » se verra automatiquement attribuer la **note de zéro**. Bien que non requis, on vous encourage à faire un design de classes avant de commencer à coder. Aussi, le code, les données et la documentation doivent être organisés suivant une bonne structure de répertoires. Pour vous aider, vous pouvez utiliser le projet

« cookiecutter » (<https://github.com/audreyr/cookiecutter>). La solution proposée doit aussi être facile à utiliser.

Gestion de projet

Comme tout projet qui se respecte, vous devez utiliser un gestionnaire de version de code. On vous demande d'utiliser « **git** » via la plateforme

« **gitHub** » ou « **bitBucket** ». On s'attend également à ce que vous fassiez une bonne utilisation de **git**. Par exemple : évitez de « pousser » du code dans le master sans « merge », éviter les « méga » commits, etc. L'utilisation d'une plateforme de type « **trello.com** » est recommandée mais non obligatoire.

Démarche scientifique

Pour ce volet, vous devez vous poser les questions suivantes :

- avez-vous bien « cross-validé » vos méthodes ?
- Avez-vous bien fait votre recherche d'hyperparamètres ?
- Avez-vous entraîné et testé vos méthodes sur les mêmes données ?
- Est-ce que cela transparaît dans le rapport ?
- Avez-vous soumis vos résultats de test en ligne afin d'avoir les résultats de test ?
- Avez-vous uniquement utilisé les données brutes ou vous avez essayé de les réorganiser pour améliorer vos résultats ?
- Etc.

À noter qu'une faute aberrante peut entraîner une perte massive de points. Par exemple, ne pas utiliser **git** entraînerait une perte de 30 points.

2.3.1 Qualité de la langue et de la présentation

Conformément à l'article 17 du règlement facultaire d'évaluation des apprentissages² 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 études³. 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'avez 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.

²https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/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/>

2.4 Échéancier des travaux

Les dates de remise des travaux seront indiquées sur les énoncés.

2.4.1 Directives particulières

Pour les étudiantes et étudiants du cours IFT712, un projet de fin de session sera à remettre le **10 décembre 2021**.

2.5 Utilisation d'appareils électroniques et du courriel

Selon le règlement complémentaire des études, section 4.2.3⁴, 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 4.2.3⁵, toute utilisation d'appareils de captation de la voix ou de l'image exige la permission de la personne enseignante.

Note : Je ne réponds à aucun courriel.

Si vous avez des questions, s.v.p. passer au bureau de l'enseignant. Aussi, bien que les appareils électroniques soient acceptés en classe, **les étudiantes et étudiants utilisant un ordinateur portable doivent s'installer à l'arrière afin de ne pas distraire le reste de la classe.**

3 Matériel nécessaire pour l'activité pédagogique

La manuel de Bishop [1] est fortement suggéré, mais non obligatoire pour réussir le cours. Les livres de Duda et al. [2] et de Murphy [3] sont également de bonnes références.

Les liens vers les vidéos et les diapositives (format PDF) sont disponibles sur le site web du cours.

4 Références

- [1] BISHOP, CHRISTOPHER M. : *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.
- [2] DUDA, RICHARD O. AND HART, PETER E. AND STORK, DAVID G. : *Pattern Classification*. Wiley, New York, 2 édition, 2001.
- [3] MURPHY, KEVIN P. : *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

⁴https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Intranet/Informations_academiqes/Sciences_Reglement_complementaire_2017-05-09.pdf

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



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
