Description des problèmes soumis au 13e atelier de résolution de problèmes de Montréal

1. Prévision des besoins en dispositifs de charge unitaire (Air Canada)

Air Canada utilise des dispositifs de charge unitaire (dont l'acronyme est ULD en anglais) pour transporter du fret et des bagages sur ses vols. Lorsque l'horaire change, la compagnie a besoin de déterminer le nombre d'ULD dont elle a besoin pour chaque type d'avion et chaque vol ; ceci permet de prévoir le nombre de déplacements d'ULD vides. En effet ces déplacements sont nécessaires lorsque apparaît un déséquilibre entre stations. Un déficit d'ULD empêche le transfert du fret et des bagages mais un surplus d'ULD fait augmenter les frais de location.

Le but de ce problème est donc de prévoir, pour les mois suivants, les nombres d'ULD pour chaque type d'avion et chaque vol. Air Canada désire aussi prévoir les proportions de fret et de bagages. Pendant l'atelier, l'objectif de l'équipe sera de concevoir un algorithme pour déterminer le nombre d'ULD par type d'avion et par vol, étant donné l'horaire de la compagnie. Les données disponibles incluront (a) l'horaire des vols, (b) les types d'avion pour les vols, (c) le nombre d'ULD par type d'avion, (d) l'historique du nombre d'ULD par vol et (e) les proportions de fret et de bagages observées dans le passé.

2. Interactions entre les évènements de la vie d'un client assuré (Beneva)

Beneva veut analyser le temps pendant lequel un client est en invalidité et reçoit des prestations d'un assureur jusqu'à un éventuel retour au travail (évènement d'intérêt). Toutefois il existe d'autres évènements (appelés évènements concurrents) tels que (a) la référence du client à des organismes gouvernementaux comme le RRQ et le RPC, (b) la proposition d'un règlement forfaitaire au client ou (c) un décès. Chacun de ces évènements peut être modélisé par une analyse de survie en considérant les autres évènements « concurrents » comme des censures. Dans ce cas, on considère que ces évènements sont indépendants, ce qui n'est pas le cas.

Plutôt que d'utiliser des courbes de Kaplan-Meier (KM), on pourrait utiliser des courbes d'incidence cumulative, qui estiment la probabilité marginale d'un évènement en présence d'évènements concurrents. Les probabilités de ces évènements, qui sont aussi appelées « risques compétitifs », peuvent être calculées en utilisant des méthodes de survie classiques adaptées à la problématique comme des modèles de

Cox ou des modèles d'incidence cumulatifs, ainsi qu'en utilisant des forêts aléatoires en survie.

Le but de l'équipe sera d'explorer les différentes méthodes en tenant compte de plusieurs critères tels que (a) les données manquantes, (b) les censures à droite et à gauche et (c) les variables qui évoluent dans le temps. Un échantillon des données sera mis à la disposition de l'équipe afin que les méthodes proposées soient appliquées à des données concrètes.

3. Évaluation d'une solution de création de données tabulaires synthétiques (Desjardins)

L'utilisation des données est au cœur de deux orientations stratégiques du Mouvement Desjardins, en apparence difficiles à concilier : la création de valeur pour nos membres et clients grâce à l'analytique avancée, exploitant au maximum toutes les informations pertinentes, et l'impératif de sécurité des données, qui limite l'accès aux données confidentielles.

De nouvelles approches innovantes en intelligence artificielle (IA), développées à l'interne, permettent de fabriquer des données synthétiques qui ne sont liées à aucun membre ou client de Desjardins et ont les mêmes propriétés statistiques que les données confidentielles. Elles pourraient donc servir à développer, sans risque de bris de confidentialité, des modèles analytiques pour la création de valeur.

Le défi proposé pour l'atelier est de mettre à l'épreuve l'approche innovante de fabrication de données tabulaires et d'évaluer sa performance. En particulier, Desjardins aimerait obtenir des réponses aux questions suivantes. (a) Est-il possible de reconstruire une observation réelle à partir d'observations synthétiques ? (b) Est-il possible de distinguer les observations réelles des observations synthétiques ?

Un jeu de données avec des observations réelles et synthétiques identifiées sera fourni à l'équipe pour ses analyses. Un jeu de données de test sera aussi fourni afin d'évaluer les modèles qui seront proposés pour répondre aux deux questions ci-dessus.

4. Méthodes en ligne et taux d'apprentissage (Ericsson)

Les méthodes d'apprentissage supervisé en ligne sont des techniques permettant à un modèle d'apprentissage automatique (« machine learning » ou ML en anglais) d'apprendre à partir de données qui arrivent à la volée et ne peuvent être stockées nulle part [1]. Par conséquent l'entraînement du modèle ne peut se faire qu'une observation à la fois : dans le meilleur des cas quelques observations peuvent être utilisées. Ceci rend impossible l'emploi de méthodes comme celle de la descente de gradient stochastique.

Les méthodes variées d'apprentissage en ligne ont toutes un point commun : l'utilisateur doit spécifier un hyperparamètre, appelé taux d'apprentissage. Si la valeur de cet hyperparamètre est trop grande, le modèle aura un biais en faveur des observations futures et oubliera rapidement les données passées. Si sa valeur est trop petite, le modèle aura un biais en faveur des données passées et donc aura du mal à tenir compte des observations futures. Pour le moment le choix de l'hyperparamètre est un art. Il serait très utile de proposer une technique automatique ou un ensemble d'orientations permettant d'affecter une valeur à cet hyperparamètre. L'idéal serait de proposer une méthode s'inspirant des mêmes principes que la méthode ADAM [2].

Des séries temporelles seront fournies au début de l'atelier. Ericsson recueille et traite des séries temporelles régulièrement et s'intéresse donc beaucoup aux méthodes d'apprentissage en ligne. L'approche que nous désirons mettre au point pourrait avoir des applications intéressantes et utiles dans le domaine de l'intelligence artificielle pour les télécommunications.

Références

[1] HOI, Steven C. H.; SAHOO, Doyen; LU, Jing; and ZHAO, Peilin. Online learning: A comprehensive survey. (2021). *Neurocomputing*. 459, 249-289. Research Collection School Of Computing and Information Systems.
[2] D.P. Kingma, L.J. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference for Learning Representations*, San Diego, (2015).

5. Équilibrage de séries temporelles pour l'apprentissage automatique (Hitachi)

Les modèles d'apprentissage automatique (dont l'acronyme anglais est ML) dépendent largement des données utilisées pour les entraîner. Ces modèles peuvent atteindre une grande précision lorsque leur jeu de données d'entraînement est équilibré, c'est-à-dire lorsqu'il contient des nombres comparables d'observations pour tous les scénarios. Dans beaucoup d'applications, la construction d'un jeu de données équilibré est difficile ou impossible et le jeu de données est déséquilibré. Lorsqu'un modèle de ML est entraîné à partir d'un jeu déséquilibré, le modèle a un biais en faveur des classes majoritaires (ou de la classe majoritaire) et néglige les classes minoritaires (ou la classe minoritaire). L'équilibrage d'un jeu déséquilibré est un domaine de recherche important en apprentissage automatique.

Les techniques les plus répandues pour l'équilibrage utilisent le rééchantillonnage. Ces techniques, par exemple SMOTE [1], requièrent un paramètre d'ajustement (représentant une borne inférieure pour le degré d'équilibre) et un critère d'arrêt pour le processus itératif de rééchantillonnage. En général déterminer le paramètre d'ajustement requiert une bonne connaissance du domaine d'application et l'expertise de l'utilisateur. Il serait hautement désirable d'automatiser le choix du paramètre d'ajustement et du critère d'arrêt et le but de l'équipe sera de proposer une approche automatique de sélection de paramètres ne consommant pas trop de temps ou de mémoire.

Un jeu de données relié à la protection d'une ligne de transmission électrique sera fourni à l'équipe. Le jeu consiste de 50 000 séries temporelles de signaux électriques et les étiquettes des signaux (binaires) indiquent si une défaillance s'est produite dans une zone spécifique ou non. Chaque série a 576 estampilles temporelles et vingt-sept caractéristiques. Les données d'entraînement ont été fabriquées grâce à un simulateur de réseau énergétique. Le but proposé à l'équipe est de trouver une méthode automatique pour équilibrer le jeu de données afin de prédire de manière satisfaisante et rapide l'occurrence d'une défaillance dans la ligne de transmission. Notez qu'un faux positif (c'est-à-dire la prédiction qu'une défaillance s'est produite dans une zone spécifique alors qu'elle s'est produite ailleurs) est plus dommageable qu'un faux négatif.

Référence

[1] Chawla, Nitesh V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (2002): 321-357.

6. Détection de motifs dans les séries d'alarmes (Hitachi)

Les systèmes d'alarmes sont un outil vital pour les opérateurs de réseaux électriques et constituent un élément essentiel de l'interface entre le contrôle automatique et le contrôle manuel. Un système d'alarmes bien conçu peut faciliter grandement les opérations tandis qu'un système mal conçu peut détourner l'attention d'informations importantes et alourdir le travail des opérateurs. Le nombre d'alarmes dans les systèmes électriques augmente rapidement et cette situation est d'autant plus gênante que les directives pour créer des alarmes cohérentes et utiles manquent.

Lors d'une perturbation, le nombre et la fréquence des alarmes sont trop élevés pour que les alarmes soient traitées par les opérateurs et il se peut que le système d'alarmes gêne le travail des opérateurs. Un des plus grands défis est dû au fait qu'une alarme est souvent accompagnée d'une suite d'alarmes reliées entre elles. Dans ce cas, il est difficile de déterminer quelle est l'alarme « originelle » et le processus reliant les alarmes entre elles. Le problème proposé pour l'atelier consiste à analyser des séries temporelles d'alarmes afin de déterminer quelles sont les alarmes « suiveuses » qui peuvent être prédites à partir d'autres alarmes. Ce problème est difficile parce que (a) les relations causales à l'intérieur d'un réseau ne sont pas stables, (b) les estampilles temporelles des alarmes peuvent être imprécises et (c) il y a du « bruit » dans les données, c'est-à-dire que des alarmes indépendantes peuvent être mélangées à des suites d'alarmes reliées entre elles.

Le jeu de données contient 1,5 million d'alarmes cueillies sur une période de huit mois. Pour chaque alarme nous avons un texte la décrivant, un identificateur de l'objet ayant généré l'alarme, une région géographique pour l'objet, un type de données, une classe d'objet et une valeur de transition d'état. En plus des données brutes, des périodes de *vagues d'alarmes* ont été identifiées : ces vagues correspondent à un taux d'alarme très élevé (de plusieurs centaines d'alarmes en quelques secondes) et résultent de perturbations dans le réseau ou de mauvaises configurations d'alarmes. Un très petit nombre de ces vagues ont des étiquettes manuelles indiquant le type de perturbation les ayant causées.

La tâche de l'équipe consistera à identifier, dans le jeu de données, des relations de causalité ou des relations basées sur des processus. Ce travail pourra mener à une application aidant les opérateurs à régler une perturbation en temps réel.

7. Valeur à long terme d'un client en assurance (Intact)

Pour déterminer la valeur à long terme d'un client en assurance, il est critique de comprendre comment les différents clients évoluent dans le temps. Dans le cadre de l'assurance de biens, ces transitions peuvent être de nouvelles réclamations, des infractions au code de la route, mais également des changements plus complexes comme des déménagements ou des substitutions/ajouts de véhicules. Cela résulte en une multitude de transitions et d'états possibles et rend très complexe, voire impossible, l'évaluation de toutes les combinaisons. Modéliser toutes ces transitions séparément de façon stochastique peut aussi être coûteux. Ces difficultés nous empêchent de faire des estimations justes et représentatives de la valeur à long terme d'un client. Pour résoudre ce problème, nous avons déjà tenté avec un succès mitigé de créer une représentation dense des clients afin de simplifier le processus de transition, mais malheureusement cette méthode nous fait perdre en général l'interprétation par les informations du client.

Dans le cadre de cet atelier, nous aimerions tenter de modéliser l'évolution d'un client dans le temps par une chaîne de Markov, comme cela se fait dans d'autres secteurs, par exemple le secteur bancaire. Cette méthode mène cependant à une quantité importante d'états possibles, si on considère toutes les combinaisons de véhicules, d'adresses, de choix de produits, etc. Il faut donc réduire la dimensionnalité du jeu de données pour limiter le nombre d'états, notamment en ce qui a trait au véhicule et au code postal d'un assuré.

Pour l'atelier, Intact fournira un jeu de données anonymisé de clients d'assurance et de certaines variables d'assurance. Dans le jeu de données, il y aura des clichés des situations des clients à des moments spécifiques. Les transitions pourront être observées grâce aux changements de caractéristiques d'un cliché à l'autre. Afin d'assurer la confidentialité, le jeu de données sera extrait de la période 2006-2010 ; de plus l'adresse, le code postal et toute autre caractéristique confidentielle ne seront pas inclus dans le jeu de données. Le travail de l'équipe ne portera donc pas sur les changements d'adresse, mais la méthodologie proposée pourra être appliquée aux changements d'adresse également. Cette méthodologie permettra à Intact de modéliser efficacement plusieurs types de transitions afin de calculer une valeur à long terme représentative.

8. Détermination du moment opportun pour solliciter la création d'un compte (Radio-Canada)

À CBC/Radio-Canada, servir le public canadien est notre raison d'être. L'un de nos objectifs stratégiques pour 2019-2024 est la personnalisation de nos services numériques. Le but est de proposer une offre numérique dans laquelle chaque Canadien pourra se reconnaître et de mettre en valeur la diversité des voix, des communautés et des points de vue qui font la richesse de notre pays.

Dans cette optique, une meilleure connaissance de nos utilisateurs est capitale : c'est pourquoi nous souhaitons encourager nos publics à s'identifier sur nos plateformes, tout en le faisant de façon responsable.

Nous pensons que des méthodes statistiques et d'apprentissage automatique pourraient nous aider à prédire les moments adéquats pour solliciter un utilisateur durant sa visite sur nos plateformes.

Ce modèle optimiserait ainsi les chances que l'utilisateur se crée un compte Radio-Canada, tout en diminuant l'irritation d'une interruption non désirée.

Le premier défi auquel nous faisons face est qu'il existe sur nos plateformes des centaines de milliers de pages, vidéos et articles. Ceci complique l'analyse des parcours puisque nous sommes contraints de regrouper/segmenter les pages afin de pouvoir analyser et interpréter les parcours.

Le second défi est que nous travaillons avec un déséquilibre des classes (« class imbalance ») au niveau de la création de comptes.

Nous travaillons actuellement avec un échantillon d'un mois de données. Cet échantillon contient des données clics par clics anonymisées, pour l'ensemble des visites sur nos plateformes Web exclusivement.

Sources

Notre plan stratégique 2019-2024
Notre politique de données

9. Prévision des pièces de rechange (Société de transport de Montréal)

L'objectif de la gestion des inventaires de pièces de rechange est d'éviter les ruptures d'inventaire tout en limitant les coûts d'entreposage. Le défi est d'avoir en permanence les bonnes pièces de rechange au bon endroit, au bon moment, dans tous les centres d'entretien de l'organisation, tout en minimisant les coûts d'entreposage. Plus précise est la prévision de la demande, meilleure sera la performance de la STM en gestion des inventaires et dans la maintenance de ses actifs. À la différence d'une grande surface qui peut retirer de son catalogue un produit qui ne se vend pas bien ou ne procure pas de bénéfice, une entreprise de maintenance ne peut pas décider de rebuter de son catalogue une pièce qui ne bouge pas ou peu. Pour maintenir ses actifs dans un bon état de fonctionnement, la STM doit s'assurer que toutes les pièces de rechange de son catalogue sont disponibles.

Les modèles de prévision classiques fonctionnent très bien pour les pièces dont la consommation est relativement stable. Pour les pièces aux profils de demande plus sporadiques et grumeleux, ces modèles sont moins efficaces et les analyses statistiques de l'historique de consommation peuvent aider à obtenir des prévisions. Des chercheurs comme Croston, Boylan et Syntetos ont beaucoup étudié ces modèles statistiques et la STM désire évaluer leur performance sur ses catalogues de pièces.

D'autre part, les pièces du catalogue de la STM peuvent être regroupées en catégories. Certaines pièces sont critiques pour les opérations. Certaines ont des profils de consommation très stables et d'autres, au contraire, sont immobiles ou inutilisées. La STM classifie son catalogue de pièces par profils de consommation mais se demande comment améliorer sa classification. Combien de catégories devons-nous former ? Comment déterminer la « distance » entre les profils de consommation de deux pièces de rechange ? Que recommandent les chercheurs qui ont étudié ce problème ? Chaque profil de consommation de pièce de rechange a un impact sur la gestion des pièces et les stratégies de stockage. La STM doit donc tenter de prédire la demande pour chaque catégorie de pièces.

Le travail de l'équipe consistera donc à classer les pièces de rechange par profils de consommation, à déterminer un modèle de prévision de la demande pour chaque catégorie et à évaluer la qualité de la prévision fournie par ce modèle. L'étude concernera les pièces du catalogue des autobus de l'organisation.