Projet de traitement de données massives

Pour ce projet, vous aurez accès aux données de l'étude électorale canadienne produite par le Consortium de la démocratie électorale[1]. Il s'agit de réponses à un sondage par des personnes à travers le pays à propos de l'élection fédérale de 2019 (nous n'utiliserons pas celle de 2021 parce qu'elle était plate à mourir).

Votre projet consistera à résoudre une problématique réelle et actuelle tant pour les chercheurs que les partis politiques : comment prédire la position politique d'un individu en utilisant ses réponses au sondage.

Les données

Les données proviennent d'un sondage en ligne. Ce sondage a été répondu par presque 40 000 personnes. Le sondage contenait plus de 600 questions, mais chaque personne recevait un sous-ensemble de ces questions et personne n'a répondu à l'ensemble des questions. Les données sont accompagnées d'une documentation PDF qui inclue la structure et logique du questionnaire ainsi que les énoncées des questions et leurs choix de réponses.

Le défi

Il y a 5 questions qui demandent à la personne, sous une forme ou une autre, pour quel parti elle veut voter. Tous les utilisateurs (sauf environ un millier) ont répondu à une de ces cinq questions. Ces questions sont :

- cps19_votechoice
- cps19_votechoice_pr
- cps19_vote_unlikely
- cps19_vote_unlike_pr
- cps19_v_advance

En plus de celle-ci il y a 9 questions demandant des détails supplémentaires sur le vote, 4 questions sur le deuxième choix et 10 questions sur le parti pour lequel la personne refuse de voter. Il s'agit des questions :

- cps19_votechoice_7_TEXT
- cps19_votechoice_pr_7_TEXT
- cps19_vote_unlikely_7_TEXT
- cps19_vote_unlike_pr_7_TEXT
- cps19_v_advance_7_TEXT
- cps19_vote_lean
- cps19_vote_lean_7_TEXT

[1] Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. *The 2019 Canadian Election Study – Online Collection*. [dataset]

- cps19_vote_lean_pr
- cps19_vote_lean_pr_7_TEXT
- cps19_2nd_choice
- cps19_2nd_choice_7_TEXT
- cps19_2nd_choice_pr
- cps19_2nd_choice_pr_7_TEXT
- cps19_not_vote_for_1
- cps19_not_vote_for_2
- cps19_not_vote_for_3
- cps19_not_vote_for_4
- cps19_not_vote_for_5
- cps19_not_vote_for_6
- cps19_not_vote_for_7
- cps19_not_vote_for_8
- cps19_not_vote_for_9
- cps19_not_vote_for_7_TEX

Les réponses à toutes ces questions ont été masquées (remplacées par numpy.nan) pour 10% des individus. Les individus ciblés sont listés dans un document de test disponible sur le site web. Votre défi pour ce projet est de prédire pour quel parti chacun de ces individus veut voter. La réponse est nécessairement une des réponses possibles du questionnaire :

- Another party (please specify)
- Bloc Québécois
- Conservative Party
- Don't know/ Prefer not to answer
- Green Party
- Liberal Party
- NDP
- People's Party

Dans le cas d'un autre parti, il n'est pas nécessaire de prédire lequel.

Questionnaire (1%)

Un questionnaire est disponible sur le site web du cours. Il pose une série de questions qui vous pousseront à explorer les données, la documentation des données, et Pandas. Ce questionnaire constitue donc un point de départ pour lancer votre travail. La note est participative. Il s'agit de la seule composante individuelle dans votre note de projet.

Question de ne pas vous stresser, ce questionnaire n'a pas de durée limite, et vous pouvez le faire un nombre illimité de tentatives. Vous avez bien entendu le droit de consulter des ressources en ligne

(particulièrement pour vous aider avec Pandas) et de collaborer avec d'autres étudiants (particulièrement ceux de votre équipe).

Vous avez jusqu'à la date de remise du premier rapport pour compléter le questionnaire.

Premier rapport : Analyse et prétraitement des données (10%)

L'objectif de la première partie du projet est de vous familiariser avec les données avec lesquelles vous allez travailler.

Analysez vos données et leurs propriétés statistiques. Portez attention autant aux valeurs normales qu'aux cas problématiques, comme la présence de bruit, le fléau de dimensionnalité, les informations manquantes, le déséquilibre des classes, les valeurs aberrantes, etc. Discutez de vos observations. L'objectif ici n'est pas de faire une grande liste de statistiques sur les données, mais d'en tirer des leçons pour guider la réalisation du projet. (3 points)

Prévoyez les attributs que vous allez utiliser pour votre algorithme de traitement de données. Il y a une immense variété d'attributs qui peuvent être obtenus de ces données, incluant des attributs linguistiques (les mots utilisés dans les réponses permettant du texte libre, etc.), des attributs démographiques sur les répondants, des attributs personnels sur les préférences des répondants, des méta-attributs (la date du questionnaire, le temps passé à répondre, etc.). Vous devez prévoir les premiers attributs sur lesquels vous allez vous concentrer (il vous est bien entendu possible d'en rajouter à n'importe quel moment au cours de la session). Justifiez votre choix d'attributs initiaux. (3 points)

Prévoyez comment vous allez traiter les données d'un point de vue pratique. C'est-à-dire premièrement les algorithmes que vous allez implémenter et comment ils vont manipuler ces données, mais aussi l'optimisation de ceux-ci afin de pouvoir traiter la quantité massive de données disponible pour ce projet de manière efficace. (2 points)

Discutez également de la procédure de tests que vous envisagez. Vous ne pouvez pas tester votre système avec les votes des individus dans la liste de test, car vous n'avez pas le vote réel pour ces individus[2]. Vous devez donc prévoir votre propre procédure de tests afin de savoir si chaque variation que vous implémentez pour votre solution améliore ou non vos prédictions, et ainsi guider votre travail de développement. (2 points)

Ce rapport est dû le 16 février 2022.

Deuxième rapport : traitement des données (10%)

Pour ce rapport, vous devez présenter les algorithmes de traitement de données que vous avez implémentés, leur fonctionnement et les résultats que vous avez obtenu. Je m'attends que vous ayez

[2] Les données étant publiques, vous pouvez les retrouver en ligne et obtenir les votes des individus. Mais ce serait de faire un sur-apprentissage de votre système pour le corpus de test, ce qui est une mauvaise pratique et sera pénalisé.

un processus de développement itératif : implémentez un système, testez-le, découvrez ses points faibles, et raffinez-le en conséquences (en ajoutant des attributs, en modifiant l'entraînement, en corrigeant l'algorithme, etc.).

Décrivez les algorithmes que vous avez choisi d'implanter. Décrivez, d'un point de vue technique, comment ils fonctionnent et les composantes clefs. Le but ici n'est pas de répéter les notions de base des algorithmes que je vous ai enseigné au cours de la session, mais plutôt d'expliquer comment vous avez adaptés et utilisés ces algorithmes pour ce projet. Justifiez vos choix pour les décisions de design et d'implémentation que vous avez pris. Décrivez leur efficacité étant donné le volume de données à traiter. (2 points)

Décrivez également les tests que vous avez faits. Pour chaque test, présentez les résultats attendus et les résultats obtenus. Présentez des statistiques pertinentes (taux de succès, précision, rappel, temps moyen de calcul, complexité algorithmique, etc.). Discutez des leçons que vous avez prises de chaque test et comment elles ont guidé votre travail. (2 points)

Présentez la version finale de votre système. Quels attributs des données sont utilisés par votre algorithme, quels attributs ont une valeur prédictive plus importante, et pourquoi? (3 points)

Présentez des études de cas (i.e. des répondants) comme exemples spécifiques du fonctionnement de votre algorithme. Décrivez autant les cas qui fonctionnent bien que ceux pour lesquels le test échoue, et discutez des raisons pour cette différence. (2 point)

Offrez une rétrospective sur le projet. Comparé à vos réflexions au début de la session, en quoi avez-vous eu raison, et quelles surprises avez-vous eu en chemin? Si le projet était à refaire, que feriez-vous différemment? (1 point)

Ce rapport est dû le 13 avril 2022.

Évaluation des résultats (2%)

En même temps que le deuxième rapport, vous devez soumettre votre prédiction du vote. Pour chaque individu test, vous devez prédire laquelle des 8 réponses possibles l'individu a donné (tous les individus tests ont indiqué une préférence).

Vous trouverez sur le site web un document de test. Ce document est sur deux colonnes séparées par une tabulation (\t). La première colonne contient le ID (le numéro de ligne, premier attribut) de l'individu test, et la deuxième colonne est pour inscrire votre prédiction.

Il y a 4 solutions triviales à ce défi. Vous pouvez assigner à chaque individu une réponse au hasard parmi les 8 possibles avec une distribution uniforme (i.e. 12,5% de chance de chaque réponse). Cette solution correspond au fichier test que je vous ai soumis en exemple. Vous pouvez assigner une réponse au hasard parmi les 8 possibles en utilisant la distribution de partis du fichier d'entraînement. En appliquant un peu de connaissances générales de la politique canadienne et en sachant que la majorité des canadiens votent pour les Libéraux ou les Conservateurs, vous pouvez

assigner à chaque individu un de ces deux partis au hasard. Finalement, vous pouvez assigner à chaque individu la réponse la plus fréquente du jeu de données d'entraînement. Ces solutions triviales donnent entre 13% et 28% de bonnes prédictions, selon la solution adoptée. Il va de soi qu'un algorithme développé intelligemment et bien entraîné devrait performer mieux qu'une solution triviale.

Évaluation des rapports

Les rapports sont limités à 12 pages, incluant les figures, tableaux et références, doivent suivre le format de la conférence Canadian AI, disponible ici :

https://www.caiac.ca/en/conferences/canadianai-2021/call-papers

Les rapports seront remis en-ligne à travers le site web du cours. Une seule soumission par équipe. Chaque rapport doit inclure une page titre indiquant les membres de l'équipe et la date de soumission. Les rapports doivent être écrits en Word ou LaTeX (pas de rapports écrits à la main) et soumis en format PDF (pas de fichiers Word ou texte).

La majorité des points du rapport seront donnés sur l'analyse et la discussion de votre système et de vos résultats. Il est donc important (pour vous) d'écrire une analyse approfondie et scientifique. La question centrale n'est donc pas « qu'est-ce qui se produit », mais « pourquoi est-ce que ça se produit » et « qu'est-ce qu'on peut y faire ». Vous ne devez pas simplement écrire un algorithme et générer des résultats. Vous devez être en mesure de justifier vos décisions qui ont mené à votre algorithme, et expliquer pourquoi il a généré ces résultats.

Un exemple peut clarifier les choses. Supposons que vos tests démontrent que la majorité de la population vote pour le Parti Rhinocéros. Vous pouvez rapporter ce résultat de plusieurs manières :

- « Notre algorithme surévalue la proportion du vote pour le Parti Rhinocéros. » Ceci n'est pas une analyse, mais simplement une observation des faits. Les points donnés seront minimaux.
- « Notre algorithme surévalue la proportion du vote pour le Parti Rhinocéros parce que ce parti obtient un score élevé dans notre algorithme trop souvent. » Ceci est l'inverse d'une analyse utile. Je ne donne pas de points, et je me réserve le droit de rire de vous.
- « Notre algorithme surévalue la proportion du vote pour le Parti Rhinocéros parce que ce parti prend toutes les positions sociales et économiques possibles, et donc il se trouve à obtenir un score élevé parce qu'il correspond à toutes les préférences de tous les électeurs. » Vous avez identifié et analysé le problème et découvert sa source, bien joué! Vous avez des points.
- « Notre algorithme surévalue la proportion du vote pour le Parti Rhinocéros parce que ce parti prend toutes les positions sociales et économiques possibles, et donc il se trouve à obtenir un score élevé parce qu'il correspond à toutes les préférences de tous les électeurs. Nous allons résoudre ce problème en assignant une pénalité à un parti adoptant une position inverse à celle préférée par l'électeur. » Non seulement vous avez découvert la source du problème, mais vous l'avez comprise assez bien pour proposer une solution, c'est fantastique. Vous aurez une bonne note.

- « Notre algorithme surévalue la proportion du vote pour le Parti Rhinocéros parce que ce parti prend toutes les positions sociales et économiques possibles, et donc il se trouve à obtenir un score élevé parce qu'il correspond à toutes les préférences de tous les électeurs. Une solution possible serait de limiter le nombre de positions qu'un parti peut prendre, mais un tel seuil serait arbitraire et pourrait pénaliser un parti avec des positions détaillées et nuancées (donc nombreuses). Une autre option serait de pondérer les positions à la baisse selon le nombre de positions du parti, mais ceci pourrait avoir l'effet de survaloriser un parti n'ayant qu'une position sur un seul dossier (comme le Parti Marijuana) et ainsi causer le problème inverse. Finalement, on pourrait assigner une pénalité à un parti adoptant une position inverse à celle préférée par l'électeur, ce qui neutralisera le bénéfice d'adopter deux positions contradictoires pour plaire aux électeurs opposés. C'est la solution que nous avons choisie d'appliquer. » Vous avez identifié le problème, vous l'avez analysé pour trouver sa source, puis vous avez exploré plusieurs pistes de solutions et justifié votre choix d'une en particulier. C'est parfait. 100%.

Notez finalement que jusqu'à 10% des points d'un rapport peuvent être enlevés en pénalité pour un rapport de mauvaise qualité. Ceci inclut particulièrement une abondance de fautes d'orthographe et de grammaire, des figures mal préparées (ou dessinées à la main), les rapports écrits à la main, le non-respect du format et de la longueur maximale, et les textes incohérents.

Équipes

Le projet doit être réalisé en équipes de 3 étudiants. La note sera donnée pour l'équipe, et non par individu, sauf dans des situations jugées extrêmes par moi-même. Choisissez bien vos coéquipiers.

Plagiat

Le plagiat est une offense académique sérieuse. Tout étudiant qui tente de soumettre un travail qui n'est pas le sien sera pénalisé. Ceci inclut de copier le travail ou rapport d'un autre étudiant du cours ou un système trouvé ailleurs. Un étudiant coupable de plagiat recevra automatiquement la note de zéro pour le projet entier (c'est-à-dire toutes les parties) et s'exposera à d'autres sanctions telles que décidées par l'Université.

Conseils

- Il y a plusieurs solutions possibles pour ce projet. Quoiqu'un algorithme de classification qui associe chaque individu à un parti (classe) semble naturel, on pourrait aussi faire un partitionnement de l'espace d'individu et associer chaque partition à un parti, ou encore créer un algorithme de recommandation de parti pour les individus. Il y a aussi une temporalité aux données, tant pour les répondants que pour la popularité des partis politiques qui varie avec le temps, alors vous pourriez traiter les données comme un flux de données (ou encore un flux par parti). Tous ces sujets sont couverts dans des cours à venir, et le matériel du cours est disponible pour vous permettre de prendre de l'avance et d'explorer ces pistes de solution.
- Essayez plusieurs de vos idées et parlez-en dans vos rapports. Décrivez quelle est l'idée, pourquoi vous pensez que c'est intéressant à essayer (qu'est-ce que vous voulez découvrir ou que pensez-

vous va arriver), et quel est le résultat obtenu (est-ce celui que vous attendiez, et sinon pourquoi). À force de réfléchir et d'expérimenter, vous trouverez une bonne solution. Et ce n'est pas mauvais que plusieurs de vos idées ne fonctionnent pas; c'est la nature même de la recherche! De plus, ça justifie expérimentalement que la version finale de votre système est la meilleure, et non simplement la première que vous avez essayé. Pour l'évaluation, je donne des points pour les explorations intéressantes (à condition qu'elles soient bien présentées, justifiées, et analysées, bien entendu). Je ne donnerai pas de points pour des idées farfelues ou mal présentées. Mais par contre je n'enlèverai jamais de points pour avoir essayé quelque chose. Et en contrepartie, si vous ne décrivez pas vos idées et expériences dans votre rapport, je ne peux pas vous donner de points du tout.

- Considérez les extrêmes logiques de vos idées. Par exemple, si augmenter le poids d'un attribut améliore les résultats, pourquoi ne pas l'augmenter encore plus, ou ne conserver que cette variable? Ce sera rarement le bon choix, mais d'explorer le comportement de votre système dans les cas extrêmes peut souvent aider à mieux comprendre le problème et à développer une nouvelle intuition pour sa solution.
- Justifiez votre analyse avec des démonstrations mathématiques lorsque possible.
- Ne soumettez pas un copier-coller de votre code au complet dans votre rapport. Expliquez comment votre algorithme fonctionne en utilisant des descriptions du processus et des étapes, la logique du système, des formules mathématiques, et du pseudo-code.
- Je suis disponible durant mes heures de bureau pour vous aider en discutant de votre projet, des difficultés que vous rencontrez, et en suggérant des idées et des pistes. Je n'ai pas la solution du projet. Et je ne vais pas déboguer votre code pour vous.

Prix Pierre Ardouin

Depuis l'automne 2013, le Département d'informatique et de génie logiciel a mis en place un concours récompensant l'équipe qui aura produit le meilleur TP/projet dans le cadre d'un cours. Ces travaux de session ont l'envergure d'un mini-projet qui est admissible par rapport aux normes fixées par le Département. À la suite des évaluations des travaux, l'enseignant du cours détermine l'équipe gagnante; chaque membre de l'équipe gagnante reçoit alors une bourse de 50$ ainsi qu'une attestation remises par le Département.

De plus, le Département d'informatique et de génie logiciel a mis en place une bourse Élite, appelée bourse « Pierre Ardouin », qui vise à récompenser le meilleur projet de session, tous cours confondus. Deux principaux critères guident le choix des évaluateurs dans l'identification du lauréat : l'excellence du travail (par rapport à ce qui est demandé dans l'énoncé) et l'aspect créativité/innovation. Il est actuellement prévu une bourse de 200$ pour récompenser chaque membre de l'équipe « élite » gagnante (pour un maximum de 1000$ pour toute l'équipe). Aussi, le Département veille à publier l'information sur un site Web dédié : http://www.ift.ulaval.ca/vie-etudiante/prix-pierre-ardouin.

À la deuxième moitié du mois de mai de chaque année universitaire, le Département organise une cérémonie pour honorer les finalistes et le lauréat du prix Pierre Ardouin des sessions d'automne et d'hiver, et leur remettre une attestation.