CS-330

Printemps 2020

Projet remplaçant le test intermédiaire annulé du 30 Mars 2020

(suivant les nouvelles directives de la direction de l'EPFL direction).

Dans ce projet, vous développerez un système d'IA qui aide à traiter les personnes susceptibles de développer une maladie cardiaque, avec les objectifs suivants :

1) Utiliser un ensemble de données accessible au public avec un algorithme d'apprentissage automatique (ID3) pour obtenir un modèle permettant de prédire si une personne développera une maladie cardiaque.
2) Tester la précision du modèle sur les données de test.
3) Générer des règles qui peuvent être utilisées pour expliquer les prédictions et les recommandations.
4) Mettre en œuvre un outil qui indique les traitements possibles pour éviter les maladies cardiaques.
5) Améliorer les performances de l'algorithme d'apprentissage et proposer de nouvelles améliorations au processus.

Pour l'implémentation, nous vous encourageons à utiliser les modèles développés dans certains des exercices du cours. Les détails du rendu sont donnés à la fin du document.

Le projet sera noté sur la base de votre rapport et d'une brève présentation via Zoom, au cours de laquelle il vous sera demandé d'expliquer votre solution. L'évaluation compte 100 points et jusqu'à 20 points bonus pourront être attribués pour des solutions particulièrement bonnes. La répartition des points est indiquée dans la description des tâches.

Vous devez faire ce projet par équipe de 2, les deux membres de l'équipe recevront la même note. Un squelette de code de base est disponible ici:

http://lia.epfl.ch/CS330_project_2020/project.py

Taches:

1) (20 points) :

Utilisez l'algorithme ID3 des exercices du 27 Avril pour traiter les données d'entraînement que vous trouvez ici :

http://lia.epfl.ch/CS330_project_2020/train_bin.csv

L'ensemble de données d'apprentissage (train_bin.csv) a déjà été traité de sorte que les attributs numériques soient discrétisés. Analysez l'ensemble de données, implémentez l'algorithme et documentez l'arbre de décision résultant (par exemple, quelle est la taille maximale / moyenne, le nombre d'enfants ?)

2) (10 points):

En utilisant les données de test :

http://lia.epfl.ch/CS330_project_2020/test_public_bin.csv

Implémentez un environnement de test qui évalue la précision des prédictions obtenues grâce à l'arbre de décision. (Précision = pourcentage de classifications correctes).

3) (20 points):

Implémentez l'algorithme de génération de règles présenté dans la classe du 27 Avril. Vous pouvez utiliser une technique similaire à l'algorithme DFS des exercices du 16 Mars pour analyser l'arbre de décision. Ecrivez une fonction qui prend comme entrée un exemple et affiche une justification de la prédiction en utilisant les conditions de la règle qui en est responsable!

4) (20 points):

Mettez en œuvre un outil qui oriente les patients des données de tests vers les traitements possibles pour prévenir une maladie cardiaque. Pour ce faire, traitez cette tâche comme une tâche de diagnostique où chacun des attributs (à l'exception du sexe et de l'âge) pourrait être considéré comme «défectueux». En d'autres termes, le traitement pourrait changer la valeur d'un ou plusieurs attribut(s) : avec la/les valeur(s) «correcte(s)», le modèle ne prédit plus une maladie cardiaque. Effectuez une abduction et utilisez un algorithme de recherche pour trouver une combinaison de changements qui entraînerait la prédiction souhaitée. Modifiez l'explication de 3) pour inclure également la suggestion de changement. Combien de patients pouvez-vous aider avec au plus un ou deux changements d'attributs ?

5) (30 points):

La discrétisation des valeurs d'attribut à l'avance peut entraîner une perte de performances de votre arbre de décision. Au lieu de cela, vous pouvez essayer une division plus sophistiquée des valeurs d'attribut, basée sur les notions de minimisation d'entropie (ou de maximisation du gain d'informations) présentées dans la classe du 27 Avril.

Pour ce faire, vous devez modifier l'algorithme ID3 à partir des exercices du 27 Avril (id3.py et noeud_de_decision.py). L'algorithme doit prendre en entrée un ensemble de données avec des valeurs d'attribut continues:

http://lia.epfl.ch/CS330_project_2020/train_continuous.csv

À chaque itération de l'algorithme, un nœud est divisé en nœuds enfants en fonction non seulement de l'attribut, mais de la combinaison de (attribut, valeur d'attribut) qui minimise l'entropie (ou, de manière équivalente, maximise le gain d'informations). Par exemple, l'algorithme peut diviser un nœud en fonction de l'attribut "âge" et de la

valeur "37" de telle sorte que tous les participants de moins de 37 ans iraient sur l'enfant de gauche, tandis que les plus âgés sur l'enfant de droite. Ne considérez que les arbres de classification binaires, c'est-à-dire que chaque nœud non feuille doit avoir deux enfants. L'algorithme continue sa récursion sur chaque sous-ensemble, mais en considérant TOUS les attributs (c.-à-d., dans une itération ultérieure, l'algorithme peut choisir de se diviser à nouveau en fonction de l'attribut "âge", mais cette fois sur la valeur 20 de sorte que tous les participants qui sont plus jeunes de 20 ans iraient sur l'enfant de gauche, tandis que ceux qui ont entre 20 et 36 ans iraient sur l'enfant de droite). Utiliser le jeu de données de test suivant :

http://lia.epfl.ch/CD330_project_2020/test_public_continuous.csv

Évaluez la précision des prédictions obtenues grâce à l'arbre de décision. Votre précision s'est-elle améliorée par rapport à l'arbre de décision de la question (1)?

DELIVRABLES:

Un fichier .zip file nommé lastname1-lastname2.zip qui contient :

1) Le fichier project.py complet, avec votre code pour les questions (1) - (5).
2) Les fichiers modifiés id3.py et noeud_de_decision.py pour la question (5).
3) Un court rapport (maximum 2 pages) décrivant vos résultats.

Dates importantes:

Lundi, 25 Mai : date-limite pour la soumission du projet.

Mercredi 27 Mai - Vendredi 29 Mai: brève interview de chaque groupe pour expliquer sa solution via Zoom (nous allons mettre en place un doodle); la période disponible pour les interviews pourrait être prolongé jusqu'à la semaine suivante si aucun temps convenable n'était trouvé.

Nous vous demandons de respecter le code d'honneur de l'EPFL et de ne pas partager les résultats entre les groupes. S'il existe des copies évidentes et que vous n'êtes pas en mesure d'expliquer comment vous avez obtenu votre solution, vous pourriez ne pas obtenir les points!