CSI 4506 Introduction à l'intelligence artificielle

Devoir 1 : Préparation des données

Marcel Turcotte

Version: sept. 15, 2024 16h34

🎯 Objectifs d'apprentissage

- **Écrire** et **exécuter** un document en format Jupyter Notebook.
- **Télécharger** et **analyser** des données dans un nuage (*cloud*) informatique.
- **Préparer** des données pour un projet d'apprentissage automatique.
- **Réaliser** une analyse exploratoire des données.

La préparation des données est l'étape fondamentale de tout projet d'apprentissage automatique. Elle consiste à transformer des données brutes en un format propre et structuré, adapté à l'analyse et à l'entraînement du modèle. L'adage "garbage in, garbage out" ("entrées mauvaises, sorties mauvaises") souligne de manière appropriée le rôle crucial de la qualité des données dans le succès d'un projet d'apprentissage automatique.

En outre, une compréhension approfondie des données est essentielle pour sélectionner des algorithmes d'apprentissage automatique appropriés et concevoir un protocole expérimental efficace. Cette compréhension aide à identifier les caractéristiques pertinentes et informe les décisions de prétraitement, améliorant ainsi la robustesse et la précision du modèle.

Dans ce devoir, vous travaillerez avec plusieurs jeux de données parmi les options fournies. Dans le devoir suivant, vous réaliserez une étude empirique en appliquant des algorithmes de classification d'apprentissage automatique à un jeu de données choisi.

📤 Soumission

- **Date limite :**
  - Soumettez votre document (Jupyter Notebook) avant le 29 septembre, 23h.

- **Devoir individuel :**
  - Ce devoir doit être réalisé individuellement.
- **Plateforme de soumission :**
  - Téléchargez votre soumission sur Brightspace dans la section Devoir (Devoir 1).
- **Format de soumission :**
  - Soumettez une copie de votre cahier (notebook) sur Brightspace. Cette copie servira de référence temporelle officielle de votre soumission.
  - Optionnellement, votre cahier **peut** inclure un lien vers un Jupyter Notebook hébergé sur Google Colab, permettant au correcteur d'accéder et d'exécuter les cellules de code. Si vous préférez une autre plateforme que Colab, assurez-vous que le correcteur puisse accéder à votre cahier sans avoir à installer de logiciels supplémentaires ou à copier des données.

Avis important : Si le correcteur ne peut pas exécuter votre code, votre soumission recevra la note de zéro. Il est de votre responsabilité de vous assurer que votre soumission fonctionne depuis un autre ordinateur que le vôtre et que toutes les cellules de votre cahier sont exécutables.

Exigences

1. Sélection du jeu de données

Dans ce devoir, vous travaillerez avec plusieurs jeux de données parmi les options fournies. Tous les jeux de données sont destinés à des tâches de classification multi-classes. Plus précisément, vous devez implémenter du code dans votre Jupyter Notebook qui récupère les jeux de données depuis le web.

1. Jeu de données pour l'identification du verre :
   - Nombre d'échantillons : 214, Nombre d'attributs : 9, Nombre de classes : 7 (types de verre comme la verrerie, les phares, les véhicules)
   - www.kaggle.com/datasets/danushkumarv/glass-identification-data-set
2. Jeu de données de dermatologie :
   - Nombre d'échantillons : 366, Nombre d'attributs : 34, Nombre de classes : 6 (troubles – psoriasis, dermatite séborrhéique, lichen planus, pityriasis rosea, dermatite chronique et pityriasis rubra pilaris)
   - www.kaggle.com/datasets/olcaybolat1/dermatology-dataset-classification

3. Risque de santé maternelle :

    - Nombre d'échantillons : 1013, Nombre d'attributs : 6, Nombre de classes : 3 (niveau de risque – élevé, moyen, faible)
    - archive.ics.uci.edu/dataset/863/maternal+health+risk

4. Jeu de données de voitures :

    - Nombre d'échantillons : 1728, Nombre d'attributs : 6, Nombre de classes : 4 (inacceptable, acceptable, bon, très bon)
    - archive.ics.uci.edu/dataset/19/car+evaluation

5. Jeu de données sur la qualité du vin :

    - Nombre d'échantillons : 4898, Nombre d'attributs : 11, Nombre de classes : 11 (0 à 10)
    - https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

6. Jeu de données des 16 personnalités :

    - Nombre d'échantillons : 60K, Nombre d'attributs : 60, Nombre de classes : 16 (type de personnalité)
    - www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt

7. Jeu de données sur les cotes de crédit :

    - Nombre d'échantillons : 100K, Nombre d'attributs : 27, Nombre de classes : 3 (Bon, Standard, Faible)
    - www.kaggle.com/datasets/parisrohan/credit-score-classification

Les jeux de données de Kaggle nécessitent un compte (et un mot de passe) pour le téléchargement, ce qui ajoute une complexité inutile à ce devoir. Pour simplifier le processus, j'ai téléchargé les données sur un dépôt public sur GitHub.

- github.com/turcotte/csi4106-f24/tree/main/assignments-data/a1

Dans votre cahier, vous pouvez accéder et lire les données directement depuis ce dépôt GitHub.

2. Analyse exploratoire

1. **Analyse des valeurs manquantes** : Examinez les jeux de données pour identifier et évaluer les valeurs manquantes dans divers attributs. Les valeurs manquantes peuvent être représentées par des symboles tels que '?', des chaînes vides ou d'autres substituts.

   1.1 Parmi la liste d'options, quels sont les jeux de données qui contiennent des valeurs manquantes ? Plus précisément, quel attribut ou quels attributs ont des valeurs manquantes ?

   1.2 Décrivez la méthodologie utilisée pour cette investigation, et fournissez le code correspondant.

   1.3 L'imputation des données consiste à remplacer les données manquantes ou incomplètes par des valeurs substituées pour préserver l'intégrité du jeu de données pour l'analyse ultérieure. Proposez des stratégies d'imputation pour chaque attribut avec des valeurs manquantes.

2. **Sélectionnez et familiarisez-vous avec une tâche de classification :** Choisissez l'un des jeux de données fournis pour une investigation plus approfondie. Il est conseillé de sélectionner un jeu de données contenant un nombre suffisamment important d'exemples, idéalement autour de 1 000, pour garantir des résultats robustes lors de l'application des algorithmes d'apprentissage automatique dans le devoir suivant.

   1.1 Quel est l'objectif de la tâche ? Est-elle destinée à une application spécifique ? Possédez-vous une expertise dans ce domaine d'application particulier ?

3. **Analyse des attributs** :

   3.1 Déterminez quels attributs manquent d'informativité et doivent être exclus pour améliorer l'efficacité de l'analyse d'apprentissage automatique. Si toutes les attributs sont jugés pertinents, indiquez explicitement cette conclusion.

   3.2 Examinez la distribution de chaque attribut (colonne) dans le jeu de données. Utilisez des histogrammes ou des boxplots pour visualiser les distributions, en identifiant les tendances ou les valeurs aberrantes.

4. **Analyse de la distribution des classes** : Examinez la distribution des étiquettes de classe dans le jeu de données. Utilisez des diagrammes en barres pour visualiser la fréquence des instances pour chaque classe, et évaluez si le jeu de données est équilibré ou déséquilibré.

5. **Prétraitement**

   5.1 Pour les attributs numériques, déterminez la meilleure transformation à utiliser. Indiquez la transformation qui semble appropriée et pourquoi. Incluez le code illustrant comment appliquer la transformation. Pour au moins un attribut, montrez la distribution avant et après la transformation. Voir Prétraitement des données.

5.2 Pour les attributs catégoriels, montrez comment appliquer l'encodage one-hot. Si votre jeu de données ne contient pas de données catégorielles, montrez comment appliquer l'encodeur one-hot à l'étiquette (variable cible).

6. **Données d'entraînement et cible** : Définissez la variable Python `X` pour désigner les données et `y` pour désigner la classe cible. Assurez-vous de sélectionner uniquement les caractéristiques informatives.
7. **Ensembles d'entraînement et de test** : Divisez le jeu de données en ensembles d'entraînement et de test. Réservez 20 % des données pour les tests.

3. Documentation de l'analyse exploratoire

Votre rapport doit documenter de manière exhaustive tout le processus suivi pendant ce devoir. Le document (Jupyter Notebook) doit inclure les éléments suivants :

- Votre nom, numéro d'étudiant et un titre de rapport.
- Une section pour chaque étape de l'analyse exploratoire. Chaque section doit contenir le code Python pertinent ainsi que les explications ou résultats.
  - Pour les sections nécessitant du code Python, incluez le code dans une cellule.
  - Pour les sections nécessitant des explications ou des résultats, incluez-les dans une cellule séparée ou en combinaison avec des cellules de code.
- Assurez une séparation logique du code dans différentes cellules. Par exemple, la définition d'une fonction doit être dans une cellule et son exécution dans une autre. Évitez de placer trop de code dans une seule cellule pour maintenir la clarté et la lisibilité.
- Le document que vous soumettez doit inclure les résultats de l'exécution, complets avec des graphiques. En d'autres termes, votre assistant d'enseignement doit pouvoir noter votre document sans avoir besoin d'exécuter le code.

Évaluation

- **Effort global dans le rapport (20%)**
  - Assurez-vous que l'écriture est claire et descriptive, permettant au correcteur de comprendre facilement ce qui a été fait, comment cela a été accompli et les raisons sous-jacentes.
  - Maintenez une bonne séparation entre le texte, le code et les résultats pour une meilleure lisibilité.
  - Fournissez des tests sur divers exemples que le correcteur peut facilement exécuter.

  - Facilitez la comparaison entre différentes approches grâce à des visualisations utilisant des tableaux et/ou des graphiques.
  - Assurez-vous que le rapport est suffisamment détaillé pour permettre la reproductibilité des résultats.

- **Description du jeu de données (10%)**
  - Justifiez le choix du jeu de données.
  - Fournissez une description détaillée du jeu de données.
  - Définissez clairement ce que représentent les attributs et la cible.

- **Analyse exploratoire (60%)**
  - Analyse des valeurs manquantes (10%)
  - Analyse des attributs (10%)
  - Analyse de la distribution des classes (10%)
  - Prétraitement (20%)
  - Données d'entraînement et cible (5%)
  - Ensembles d'entraînement et de test (5%)

- **Ressources et références (10%)**
  - Citez toute partie de votre code qui provient de sites web, y compris les sites de tutoriels ou Stack Overflow.
  - Référencez toute théorie ou algorithme utilisé qui se trouve dans des livres, des présentationa ou des tutoriels.

Ressources

Comme mentionné précédemment, assurez-vous de citer toute partie de votre code provenant de sites web, de manuels ou d'autres ressources externes.

De nos jours, de nombreux programmeurs utilisent l'intelligence artificielle pour améliorer leur productivité, une tendance qui est susceptible de continuer à croître. Pour mieux vous préparer au marché du travail, il est plausible d'utiliser ces technologies. Cependant, il est impératif que vous compreniez pleinement les concepts sur lesquels vous êtes évalué, car ces outils ne seront pas disponibles pendant les évaluations en personne.

Si vous utilisez une assistance IA, documentez soigneusement vos interactions. Incluez les outils et leurs versions dans votre rapport, ainsi qu'une transcription de toutes les interactions. La plupart des assistants IA gardent une trace de vos conversations. La pratique recommandée est de créer une nouvelle conversation spécifiquement pour le devoir et de réutiliser systématiquement cette conversation tout au long de votre travail sur le devoir. Assurez-vous que cette

conversation est exclusivement dédiée au devoir. Soumettez cette transcription de conversation dans la section des références de votre Jupyter Notebook.

Questions

- Vous pouvez poser vos questions dans le sujet du devoir sur le forum de discussion de Brightspace.
- Alternativement, vous pouvez envoyer un courriel à l'un des deux assistants d'enseignement. Cependant, l'utilisation du forum est fortement préférée, car elle permet à vos camarades de classe de bénéficier des questions et des réponses correspondantes fournies par les assistants d'enseignement.

Remerciements

Ce devoir utilise une liste de jeux de données de classification compilée par Caroline Barrière, et suit le format général de ses directives pédagogiques. Merci Caroline!