Préparation des données

Objectifs:

— Découvrir l'importance de la préparation des données dans la modélisation prédictive avec apprentissage automatique

Introduction

Les projets de modélisation prédictive impliquent la préparation de données.
Les données font référence à des exemples ou des cas du domaine qui caractérise le problème que vous souhaitez résoudre.
Sur un projet de modélisation prédictive, tel que la classification ou la régression, les données brutes ne peuvent généralement pas être utilisées directement.

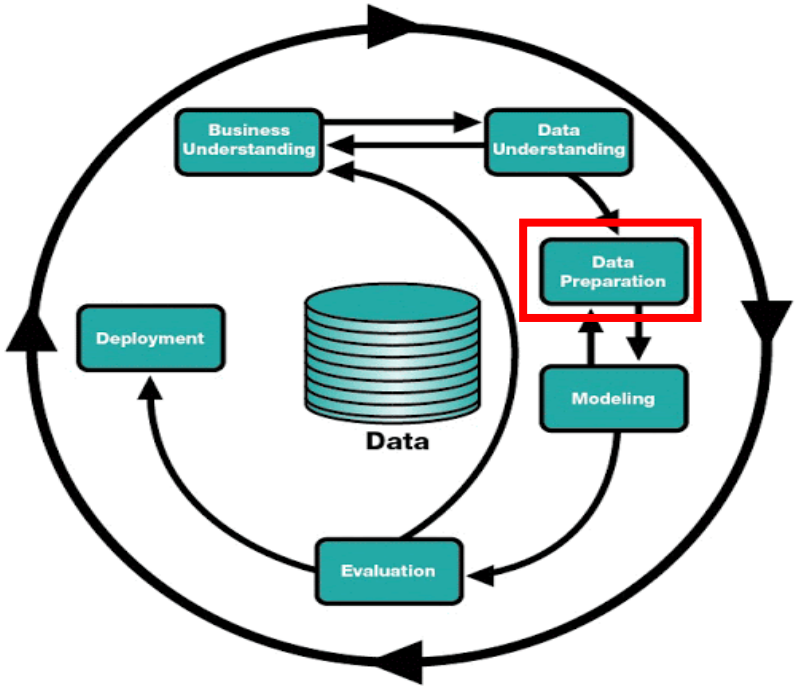
Il faut alors :

— Nettoyage de données/ Types de données : certains algorithmes d'apprentissage automatique exigent que les données soient des nombres.
— Exigences en matière de données : certains algorithmes d'apprentissage automatique imposent des exigences aux données.
— Erreurs de données : le bruit statistique et les erreurs dans les données doivent être corrigés.
— Complexité des données : des relations non linéaires complexes peuvent être extraites des données.

Préparation des données: Intérêt

Définition:

Les données brutes doivent être prétraitées avant d'être utilisées pour ajuster et évaluer un modèle d'apprentissage automatique. Cette étape d'un projet de modélisation prédictive est appelée préparation des données.



Les activités de la préparation

Il existe des activités courantes ou standard que vous pouvez utiliser ou explorer lors de l'étape de préparation des données dans un projet d'apprentissage automatique.

Ces activités comprennent :

Nettoyage des données : identifier et corriger les erreurs dans les données.
Sélection des caractéristiques (Feature Selection) : identification des variables d'entrée les plus pertinentes pour la tâche.
Transformations de données : modification de l'échelle ou de la distribution des variables.
Ingénierie des caractéristiques (Feature engineering) : dérivation de nouvelles variables à partir des données disponibles.
Réduction de dimensionnalité : création de projections compactes des données.

Chacune de ces tâches est un domaine d'étude complet avec des algorithmes spécialisés.

Nettoyage des données : Imputation des données

— Supprimer les individus qui comportent des données manquantes
— Remplacer les données manquantes par une valeur fixe
— Remplacer les données manquantes avec un arbre de décision
— Remplacer les données manquantes par les valeurs les plus proches
— Remplacer les données manquantes avec des algorithmes dédiés…





La meilleure valeur à utiliser?

Nettoyage des données : Imputation des données

Comment identifier toutes les valeurs manquantes dans les données ?

Les données du monde réel ont souvent des valeurs manquantes généralement par ce que les observations n'ont pas été bien enregistrées ou à cause de la corruption des données.

La gestion des données manquantes est importante car de nombreux algorithmes d'apprentissage automatique ne prennent pas en charge les données avec des valeurs manquantes.

Le remplissage des valeurs manquantes avec des données est appelé imputation des données.

Consiste à remplir les valeurs manquantes avec l'imputation

Nettoyage des données : Imputation des données

Une approche courante pour l'imputation des données consiste à calculer une valeur statistique pour chaque colonne et à remplacer toutes les valeurs manquantes pour cette colonne par la statistique.

Stratégies :

— Si «moyenne», remplacez les valeurs manquantes en utilisant la moyenne de chaque colonne. Ne peut être utilisé qu'avec des données numériques.

— Si «médiane», remplacez les valeurs manquantes en utilisant la médiane de chaque colonne. Ne peut être utilisé qu'avec des données numériques.

— Si «most_frequent», remplacez la valeur manquante en utilisant la valeur la plus fréquente de chaque colonne. Peut être utilisé avec des chaînes ou des données numériques.

— Si «constant», remplacez les valeurs manquantes par *fill_value*. Peut être utilisé avec des chaînes ou des données numériques.

— Lorsque stratégie == «constant», *fill_value* est utilisé pour remplacer toutes les occurrences de missing_values.

Si la valeur par défaut est laissée à la valeur par défaut, fill_value sera 0 lors de l'imputation de données numériques et «missing_value» pour les chaînes ou les types de données d'objets.

Nettoyage des données : Imputation des données

1.Le choix de la méthode pour remplir les valeurs NaN dépend de la situation.

2.La méthode habituelle pour traiter les valeurs manquantes est de les remplacer par -999, la moyenne ou la médiane.

3.Les valeurs manquantes peuvent déjà être remplacées par quelque chose par les organisateurs.

4.La caractéristique binaire "isnull" peut être bénéfique.

5.En général, évitez de remplir les valeurs manquantes avant la génération des caractéristiques.

6.Xgboost peut gérer les valeurs NaN.

Atelier: imputation des données (Voir exemple Data preparation Lab I)

Feature Selection

Feature Selection est un processus dans lequel vous sélectionnez automatiquement les features (caractéristiques) de vos données qui contribuent le plus à la variable de prédiction ou à la sortie qui vous intéresse.

Le fait d'avoir des caractéristiques non pertinentes dans vos données peut diminuer la précision de nombreux modèles, en particulier les algorithmes linéaires comme la régression linéaire et logistique.

Les trois avantages de la sélection des caractéristiques avant de modéliser vos données sont :
— Réduit le surajustement : moins de données redondantes signifie moins de possibilités de prendre des décisions basées sur le bruit.
— Améliore la précision : moins de données trompeuses signifie que la précision de la modélisation s'améliore.
— Réduit le temps d'entraînement (Training): moins de données signifie que les algorithmes s'entraînent plus rapidement.

Transformation de données

Cette méthode est utilisée pour supprimer le bruit d'un jeu de données. On appelle bruit toute donnée déformée et dénuée de sens dans un ensemble de données.

Exemple de transformation:

- Normalisation
- Agrégation de données
- Discrétisation
- Généralisation…

Transformation de données: Normalisation

Aussi appelé prétraitement des données, c'est l'une des techniques cruciales pour la transformation des données dans le Machine Learning. Ici, les données sont transformées pour qu'elles tombent sous une plage donnée. Lorsque les attributs sont sur des plages ou échelles différentes, la modélisation des données peut être difficile. La normalisation aide à appliquer des algorithmes du Machine Learning et à extraire les données plus rapidement.

Vers [0:1]

`sklearn.preprocessing.MinMaxScaler`

$$X = (X - X.min())/(X.max() - X.min())$$

Vers Mean = 0 et STD = 1

`sklearn.preprocessing.StandardScaler`

$$X = (X - X.mean())/X.std()$$

Transformation de données: Autres méthodes

Agrégation de données: L'agrégation est le processus de collecte de données à partir de diverses sources et de stockage dans un format unique. Ici, les données sont collectées, stockées, analysées et présentées sous forme de rapport ou de résumé. Il aide à collecter plus d'informations sur un cluster de données particulier. La méthode aide à collecter de grandes quantités de données.

Discrétisation de données: Il s'agit d'un processus de conversion de données continues en un ensemble d'intervalles de données. Les valeurs d'attribut continues sont remplacées par des étiquettes d'intervalle. Cela rend les données plus faciles à étudier et à analyser. Si un attribut continu est géré par une tâche de Machine Learning, ses valeurs discrètes peuvent être remplacées par des attributs de qualité constante. Cela améliore l'efficacité de la tâche.

Généralisation des données: Dans ce processus, les attributs de données de bas niveau sont transformés en attributs de données de haut niveau à l'aide d'hiérarchies de concepts. Cette conversion d'un niveau inférieur à un niveau conceptuel supérieur est utile pour obtenir une image plus claire des données. Par exemple, les données d'âge peuvent être sous la forme (20, 30,…) dans un ensemble de données. Il se transforme en un niveau conceptuel supérieur en une valeur catégorielle (jeune, vieux).

Transformation de données: Feature engineering

Nous savons que les algorithmes d'apprentissage automatique utilisent certaines données d'entrée pour produire des résultats.

Mais bien souvent, les données qui vous ont été fournies peuvent ne pas être suffisantes pour concevoir un bon modèle d'apprentissage automatique. C'est là que la puissance du feature engineering entre en jeu.

Exemple de techniques de Feature Engineering :

Création des Features:

Somme, soustraction, moyenne, min, max, produit, quotient du groupe de features.

Extraction de Features à partir d'un texte:

Bag of words: est une représentation de texte qui décrit l'occurrence de mots dans un document.

Word2vec: Il repose sur des réseaux de neurones à deux couches et cherche à apprendre les représentations vectorielles des mots composant un texte, de telle sorte que les mots qui partagent des contextes similaires soient représentés par des vecteurs numériques proches.

topic extraction: extraire les sujets principaux à partir d'un texte…

Extraction des features à partir d'une image: Variance, Ecart type, Skewness, Kurtosis…