Analyse des principaux pays du monde en matière de développement. Aide à la décision pour les organisations de solidarité internationale.

Projet - UV ODATA

24 octobre 2021

Objectifs du projet

L'objectif du projet est d'effectuer une classification non supervisée des principaux pays du monde en matière de développement à partir de données socio-économiques et de santé sur chaque pays. Cette classification permettra de guider les choix d'affectation de l'aide humanitaire internationale en identifiant les pays dans lesquels il est nécessaire d'intervenir en priorité et de préciser les domaines d'action à privilégier.

Les méthodes de clustering à considérer / tester sont les suivantes :
— K-means
— Classification ascendante hiérarchique (CAH)
— Modèle de mélange de Gaussiennes
— DBSCAN
— Partitionnement spectral (Spectral clustering)

D'abord, une étude préalable vous permettra de comparer ces 5 méthodes sur des données simulées. Vous étudierez le principe des méthodes DBSCAN et Spectral clustering par vous-même, vous analyserez les performances des différentes techniques sur plusieurs jeux de données simulées et vous préciserez quelles sont les techniques les mieux adaptées à chaque type de données.

Ensuite, vous vous intéresserez au problème de clustering des principaux pays du monde en matière de développement à partir d'un fichier de données regroupant des indicateurs socio-économiques et de santé sur chacun des pays.

Vous analyserez les partitions obtenues par les différentes méthodes à l'aide des métriques de votre choix et vous justifierez les résultats obtenus. Enfin vous proposerez une liste d'environ 10 pays qui ont le plus besoin de soutien de la part des organisations de solidarité internationale. Si possible, vous préciserez également les domaines d'action à développer.

Livrables attendus

A la fin du projet, vous devrez fournir 2 fichiers :

1. Un rapport (format pdf) comprenant les éléments suivants :
   — un rappel des objectifs du projet,
   — les réponses aux différentes questions posées,
   — une description du protocole expérimental mis en place : objectifs, métriques utilisées, ... Pour chaque méthode vous préciserez les paramètres choisis (initialisation, distance, ...) et vous justifierez vos choix,
   — une analyse et une interprétation des résultats obtenus.
2. Le code, clair et commenté (archive au format zip).

1 Etude préalable : Comparaison des méthodes de clustering sur des données simulées

1.1 Etude des méthodes DBSCAN et Spectral Clustering

Parmi les méthodes considérées, seules les méthodes DBSCAN et Spectral clustering n'ont pas été étudiées en cours, il faut donc faire une recherche sur ces méthodes et comprendre leur principe.

Expliquez de façon succincte le principe de chaque méthode et indiquez ses avantages.

1.2 Etude des classes et modules relatifs aux différentes méthodes

Dans le projet, vous allez considérer les classes ou modules suivants des librairies `scikit-learn` et `scipy` :

— K-means : `sklearn.cluster.KMeans`
— CAH : `scipy.cluster.hierarchy`
— Modèle de mélange de Gaussiennes : `sklearn.mixture.GaussianMixture`
— DBSCAN : `sklearn.cluster.DBSCAN`
— Spectral clustering : `sklearn.cluster.SpectralClustering`

Etudiez ces différentes classes et modules : paramètres d'appel, attributs, méthodes, fonctions.

1.3 Expérimentations

Il s'agit d'étudier les performances des 5 méthodes sur des données simulées correspondant à des clusters de formes différentes. Ces données sont stockées dans les fichiers suivants : `jain.txt`, `aggregation.txt` et `pathbased.txt`.

Chaque fichier contient 3 colonnes. Pour chaque individu (ou point), les 2 premières colonnes correspondent aux valeurs de 2 caractéristiques, la 3ème colonne indique sa classe d'appartenance. Cette information pourra servir de référence pour évaluer les partitions obtenues par clustering.

Importez les données et visualisez les nuages de points correspondants.

Pour chaque méthode :
- Précisez les valeurs choisies pour les paramètres, par exemple pour l'initialisation, le type de distance, la méthode de linkage en classification hiérarchique....
- Proposez un partitionnement de chaque jeu de données. Pour le nombre de clusters K, vous choisirez le nombre réel de classes.

Evaluez les performances des 5 méthodes :
- de façon qualitative : représentez les points des clusters obtenus par des couleurs différentes et comparez visuellement avec les "vraies" classes.
- de façon quantitative : comparez le partitionnement obtenu avec la vraie classification en calculant l'indice de Rand ajusté (ARI).

Comparez les performances obtenues avec les 5 méthodes. Précisez quelles sont les méthodes les mieux adaptées aux différentes formes de clusters.

2 Partitionnement des principaux pays du monde en fonction de leur développement

L'objectif est maintenant d'utiliser ces 5 méthodes pour classer automatiquement les principaux pays du monde en fonction de leur développement à partir de données socio-économiques et de santé sur chaque pays.

L'intérêt de cette classification est de guider les choix d'affectation de l'aide humanitaire internationale en identifiant les pays dans lesquels il est nécessaire d'intervenir en priorité et éventuellement de préciser les domaines d'action.

Le fichier `data.csv` stocke pour 167 pays les données suivantes :
- Taux de mortalité infantile (nombre de décès d'enfants de moins de 5 ans pour mille naissances)
- Exportations de biens et de services par habitant (exprimé en pourcentage du PIB par habitant)

— Dépenses totales de santé par habitant (exprimé en pourcentage du PIB par habitant)
— Importations de biens et de services par habitant (exprimé en pourcentage du PIB par habitant)
— Revenu net par personne
— Inflation (taux de croissance annuel du PIB)
— Espérance de vie
— Taux de fécondité (nombre moyen d'enfants par femme)
— PIB (produit intérieur brut) par habitant

Visualisez le tableau de données, puis importez les données du fichier.

2.1 Examen des données

Après l'importation, il est important d'examiner plus en détail ces données, en particulier :
— la taille du jeu de données,
— le type des données (numérique : int, float ou qualitatif/catégoriel : object),
— la qualité des données (est-ce qu'il y a des données manquantes ?),
— la distribution des données (est-ce qu'il y a des données aberrantes ?).

En utilisant les méthodes de la classe `DataFrame`, procédez à l'examen des données et notez les informations qui vous paraissent pertinentes.

En particulier, il est important d'identifier les données manquantes (représentées par le symbole 'NA' : Not Available) qui devront être pré-traitées avant le clustering (voir section suivante). En utilisant la méthode `isna()` de la classe `DataFrame`, vous pouvez identifier les valeurs manquantes et déduire le nombre de valeurs manquantes pour chacune des variables.

Il est aussi intéressant de connaître les statistiques des données à traiter. Pour cela, vous pouvez utiliser la méthode `describe()` de la classe `DataFrame` et construire une visualisation de type histogramme pour chaque variable numérique avec la méthode `hist()` de la classe `DataFrame`. Vous pouvez ainsi identifier les éventuelles données aberrantes, c'est-à-dire en dehors de l'échelle de valeurs prises habituellement par une variable.

2.2 Préparation des données

Pour faire fonctionner correctement les algorithmes de clustering, il est nécessaire d'avoir des données numériques de bonne qualité.

Pour résoudre le problème des valeurs manquantes et des valeurs aberrantes, plusieurs solutions sont possibles :
— rechercher la vraie valeur via d'autres sources d'information,

— attribuer une valeur conforme à la distribution de la variable (moyenne, médiane, valeur la plus probable...) en utilisant la méthode `fillna()` de la classe `DataFrame`,
— supprimer la variable correspondante, si le nombre de valeurs manquantes ou aberrantes est très important (plus d'un tiers des données environ).

Enfin, il est nécessaire de centrer les données et de les réduire si les plages de valeurs sont différentes.

Après avoir réalisé toutes ces transformations, il est intéressant d'examiner à nouveau toutes les variables qui seront utilisées pour le clustering.

2.3 Recherche de corrélations

Pour mieux comprendre les données, il faut s'intéresser aux relations qui existent entre les variables. Pour cela, il faut calculer le coefficient de corrélation entre chaque couple de variables numériques.

Commentez les résultats obtenus.

2.4 Analyse exploratoire des données

Avant d'appliquer les méthodes de clustering, il est intéressant d'effectuer une ACP pour aller plus loin dans l'analyse des données. L'ACP va permettre de mieux comprendre les relations (corrélations) entre les variables, de faire des regroupements de pays similaires (donc d'effectuer un premier clustering) et éventuellement de réduire la dimension des données.

Effectuez une ACP sur les données. Combien d'axes proposez-vous de conserver ?
Représentez la projection des pays dans le(s) premier(s) plan(s) principal(aux) ainsi que la projection des variables dans le(s) cercle(s) des corrélations. Donnez une interprétation des axes conservés.

2.5 Clustering des données

Effectuez le clustering proprement dit sur les données des différents pays avec les 5 méthodes proposées.

Pour chaque méthode :
— Précisez les valeurs choisies pour les paramètres.
— Evaluez la qualité des partitions obtenues pour différentes valeurs de K, le nombre de clusters, en utilisant les métriques de votre choix disponibles dans le module `sklearn.metrics`.
— Proposez une valeur de K, justifiez votre choix,

— Pour la valeur de K choisie, analysez les clusters obtenus au regard des données du fichier.
— Visualisez graphiquement les clusters obtenus sur le premier plan principal défini par les 2 premiers axes de l'ACP.

Vous pouvez appliquer les algorithmes sur les données complètes (sans ACP) et/ou les données réduites obtenues après l'ACP.

Comparez les résultats obtenus avec les 5 méthodes, interprétez et commentez les résultats.

Etudiez plus précisément le cluster correspondant aux pays les moins avancés en terme de développement. Donnez vos préconisations pour l'affectation de l'aide humanitaire internationale en identifiant une dizaine de pays dans lesquels il vous paraît nécessaire d'intervenir en priorité. Si c'est possible, précisez également les domaines d'action à privilégier.

Le projet est ouvert, donc soyez curieux, n'hésitez pas à proposer, tester, expérimenter... Votre démarche, vos idées, votre analyse comptent plus que le résultat de clustering lui-même.