Statistiques numériques et analyse de données - Projet

Alexandre AHETO, Raphaël GRAFF-MENTZINGER, Anne SPITZ

5 février 2017

Dans le cadre de notre projet, nous avons choisi d'étudier la base de données « Communities and Crime Data Set », fournie librement par le site UCI. Composée de 1994 échantillons comportant chacun 128 attributs réels, elle rassemble des informations diverses, notamment démographiques, sur les *communities* américaines. On y retrouve ainsi pour chaque échantillon des données telles que le revenu moyen, le pourcentage de blancs, de noirs, d'asiatiques et d'hispaniques, le niveau d'éducation ou encore le nombre de résidences inoccupées. Le dernier attribut de chaque échantillon représente, comme le titre de la base de données peut le laisser sous-entendre, son taux de criminalité, ou plus précisément le nombre de crimes violents commis durant l'année 1990, par tranche de 100,000 habitants.

Nous avons choisi d'étudier cette base de données dans le but de répondre à la problématique suivante :

Est-on capable de prédire avec certitude le taux de criminalité d'une ville, et quelles sont alors les principales causes d'une criminalité élevée ?

Le langage que nous utiliserons pour faire cette étude est le langage R, qui possède de puissantes fonctionnalités dédiées aux statistiques et à l'analyse de données.

1. Pré-traitement de la base de données

La base de données que nous utilisons est la synthèse de trois études menées entre les années 1990 et 1995 : les données socio-économiques sont fournies par l'*US Census*, celles traitant de la législation proviennent de l'étude *US LEMAS*, et enfin les données sur la criminalité sont fournies par le FBI, dans une étude publiée en 1995.

Or l'étude *US LEMAS* possède la particularité de n'avoir été menée que sur les commissariats de police possédant au moins 100 officiers, afin d'obtenir des résultats statistiquement significatifs. Cette limitation fait qu'un certain nombre de villes ont dues être exclues de la base de données que nous étudions, et que même parmi celles restantes, moins de la moitié a tous ses attributs renseignés. Notre première étape est donc de « nettoyer » la base afin de pouvoir appliquer dessus nos outils statistiques.

Si l'on applique la fonction `summary` à notre base de données brute, on observe que, parmi ses 128 colonnes, 24 sont constituées presque uniquement de points d'interrogation (de l'ordre de 1600 points d'interrogation sur 2000 échantillons). Ces colonnes correspondent, comme attendu, aux données fournies par l'étude *US LEMAS*. Nous choisissons donc de supprimer tout simplement ces colonnes de notre base de données, puisque bien que présentant des éléments qui nous sembleraient utiles pour répondre à notre problématique, elles contiennent trop peu

d'information pour être exploitables dans le cadre de notre étude.

Après avoir supprimé ces attributs, nous observons qu'une colonne continue de poser problème : l'attribut *OtherPerCap* possède un point d'interrogation, en plein milieu de ses données. Nous envisageons dans un premier temps de conserver cette colonne et de remplacer sa seule valeur manquante par la moyenne de toutes les autres. Cependant, on devine que l'attribut *OtherPerCap* est directement corrélé avec les attributs *whitePerCapita*, *blackPerCapita*, *indianPerCap* et *AsianPerCapita*, et nous pensons donc que cette colonne apportera peu dans le cadre de notre problématique. Nous la supprimons donc simplement, comme nous avons fait avec les attributs précédents.

Finalement, en prévision de l'ACP et de la régression linéaire qui ne peuvent être effectuées que sur des données numériques, nous supprimons la colonne *communityname*, constituée de chaînes de caractères symbolisant le nom des villes répertoriées dans la base.

Notre base de données comporte donc maintenant 102 attributs, toujours pour 1994 échantillons. Les données ont déjà été normalisées et ramenées linéairement à des valeurs entre 0 et 1, colonne par colonne, nous n'avons donc pas besoin d'effectuer cette opération dans notre phase de pré-traitement.

2. Analyse statistique des données