BI Data Mining - TME3 - Visualisation et python

Ludovic Denoyer et Laure Soulier

1 Objectifs du TME

L'objectif de ce TME est de réaliser des analyses factorielles avec un langage de programmation, en l'occurrence python. Nous allons procéder en deux étapes : 1) implémenter chacune des étapes permettant de réaliser l'ACP, et 2) utiliser des librairies existantes.

2 L'ACP pas à pas

Nous allons utiliser le jeu de données IRIS, disponible au lien suivant `https://archive.ics.uci.edu/ml/datasets/Iris`.

Réaliser les étapes suivantes (pour chaque question, prenez le temps de regarder les résultats, interprétez) :

1. Charger les données
2. Isoler les données numériques et la classe des individus.
3. Normaliser les données numériques (centrer et réduire)
4. Estimer la matrice de variance-covariance de deux façons :
   — En appliquant la formule $Cov = \frac{1}{n} X^T X$.
   — Avec la commande *np.cov*.
5. Extraire les valeurs propres et vecteurs propres (*np.linalg.eig*). Dessiner l'éboulis des valeurs propres.
6. Estimer l'inertie des valeurs propres. Combien d'axes factoriels retenir ?
7. Concaténer les vecteurs propres associés aux k axes factoriels retenus (*np.hstack((eigvector[1].reshape(4,1), eigvector[1].reshape(4,1)))*).
8. Faire une projection des individus en 2D (axe factoriel 1 et axe factoriel 2). Coordonnées : $X.P$. Vous pouvez aussi colorer les points en fonction des classes des individus.
9. Faire une projection des individus en 3D (axe factoriel 1, axe factoriel 2, axe factoriel 3). Permettre au graphique de faire une rotation pour visualiser les données suivant les différents plans factoriels.

3 L'ACP avec sklearn

— Réaliser l'ACP du dataset iris avec sklearn :
$sklearn_pca = sklearnPCA(n\_components = 2)$
$Y\_sklearn = sklearn\_pca.fit\_transform(X\_std)$
— Projeter les données sur un axe 2D.

4 Encore plus loin...

— Charger les données du dataset *Leaf*
— Centrer et réduire les données
— Appliquer la PCA sans préciser le nombre de composants (*pca = PCA()*
pca.fit(leafN))
— Afficher les valeurs propres et vecteurs propres ($pca.explained\_variance\_ratio\_$ et $pca.components\_$).
— Dessiner l'éboulis des valeurs propres
— Transformer les coordonnées des individus dans le nouvel axe factoriel ($pca.transform$).
— Visualiser ces données en 2D puis 3D

5 A faire à la maison (non relevé)

Vous noterez qu'avec sklearn, il est impossible de réaliser le cercle des corrélations ainsi qu'un biplot individus/variables. A vos heures perdues, vous regarderez la librairie suivante (`https://github.com/mazieres/analysis`) qui propose une implémentation de tout cela et un test sur le dataset iris.