

Big Data Analytics

Clustering par ACP

Réduction de la dimensionalité : PCA

(Souvenez-vous)

Matrice de covariance (resp. corrélation) :

$$\frac{1}{N} * \bar{X}^T * \bar{X}, \left(\frac{1}{N} * \tilde{X}^T * \tilde{X} \right)$$

ACP :

Retrouver les valeurs et vecteurs propres de la matrice de covariance (resp. corrélation), donc diagonaliser la matrice carrée obtenue.

Vecteur propre : vecteur permettant de projeter les données

Valeur propre : “proportion d’information” conservée par la projection suivant le vecteur propre correspondant

Réduction de dimension : On ne projette que suivant le nombre de vecteurs propres voulus

$$\frac{1}{N} * \bar{X} * \bar{X}^T, \left(\frac{1}{N} * \tilde{X} * \tilde{X}^T \right)$$

En considérant les individus comme des features et les features comme des individus, les vecteurs propres ayant une grande valeur propre peuvent être considérés comme des centre de cluster d'individus.