Le projet sera rendu sous la forme d'un fichier de commande `Rmd` ou python notebook et de sa version compilée en pdf envoyés à l'adresse `christophe.ambroise@univ-evry.fr`.

Le projet est à rendre au plus tard le 18 décembre 2023.

Projection orthogonale

Considérons un droite $\mathcal{D}$ et un point $\boldsymbol{x}$ dans $\mathbb{R}^p$. Ecrivez les formules explicites et le code informatique associé qui donne

- les coordonnées de la projection orthogonale de $\boldsymbol{x}$ sur $\mathcal{D}$
- la distance entre $\boldsymbol{x}$ et $\mathcal{D}$.

Soit K droites $\mathcal{D}_1, \cdots, \mathcal{D}_K$ et un point $\boldsymbol{x}$ dans $\mathbb{R}^p$. Ecrire le code qui détermine la droite la plus proche de $\boldsymbol{x}$. En cas d'égalité de distance un tirage aléatoire sera utilisé.

Les nuées dynamiques ou les kmeans généralisés

L'algorithme des nuées dynamiques ou kmeans généralisés remplace la notion de centre de classe par la notion de représentant la classe et minimise le critère suivant:

$$J(C, D) = \sum_i \sum_{k=1}^{K} c_{ik} d^2(\boldsymbol{x}_i, \mathcal{D}_k),$$

où $C = (c_{ik})$ est une matrice de classification, $D = \{\mathcal{D}_1, \cdots, \mathcal{D}_K\}$, et d^2 la distance entre $\boldsymbol{x}_i$ et représentant.

- Décrire et coder un algorithme des kmeans généralisés dans $\mathbb{R}^p$ où le représentant de classe est une droite et d^2 la distance entre un point et son projeté orthogonal sur $\mathcal{D}_k$.
- Appliquer votre algorithme au jeu de données `iris` et comparer aux k-means et à l'algorithme aux mélange des gaussiennes (`mclust` en R et `sklearn.mixture.GaussianMixture` en python).
- Visualiser les différentes partitions sur les premiers plans d'une analyse en composantes principale.
- Simuler des données sont plus adaptées à votre algorithmes qu'aux kmeans. Commenter.

Questions subsidiaires

Etendre à un algorithme des kmeans généralisés dans $\mathbb{R}^p$ où les représentant de classe sont des hyperplans de dimension $q < p$ vos:

- calculs,
- codes,
- illustrations.

Références

Diday, E. (1973). The dynamic clusters method in nonhierarchical clustering. International Journal of Computer & Information Sciences, 2(1), 61-88.

Bock, H. H. (2007). Clustering methods: a history of k-means algorithms. Selected contributions in data analysis and classification, 161-172. http://www.modulad.fr/dac/Slides/Bock/Ahistoryofthek-means.pdf

The project should be submitted in the form of an `Rmd` file or a Python notebook and its compiled version in PDF format sent to the email address christophe.ambroise@univ-evry.fr.

The project must be submitted no later than December 18, 2023.

Orthogonal Projection

Consider a line $\mathcal{D}$ and a point $\boldsymbol{x}$ in $\mathbb{R}^p$. Write the explicit formulas and the associated computer code that gives the coordinates of the orthogonal projection of $\boldsymbol{x}$ onto $\mathcal{D}$ the distance between $\boldsymbol{x}$ and $\mathcal{D}$. Let K lines $\mathcal{D}_1, \cdots, \mathcal{D}_K$ and a point $\boldsymbol{x}$ in $\mathbb{R}^p$. Write the code that determines the closest line to $\boldsymbol{x}$. In case of equal distances, a random selection will be used.

Dynamic Clouds or Generalized k-means

The dynamic clouds algorithm or generalized k-means replaces the notion of class center with the notion of class representative and minimizes the following criterion:

$$J(C, D) = \sum_i \sum_{k=1}^{K} c_{ik} d^2(\boldsymbol{x}_i, \mathcal{D}_k),$$

where $C = (c_{ik})$ is a classification matrix, $D = \mathcal{D}_1, \cdots, \mathcal{D}_K$, and d^2 is the distance between $\boldsymbol{x}_i$ and the representative.

Describe and code a generalized k-means algorithm in $\mathbb{R}^p$ where the class representative is a line and d^2 is the distance between a point and its orthogonal projection onto $\mathcal{D}_k$. Apply your algorithm to the iris dataset and compare it to k-means and Gaussian Mixture Model algorithm (mclust in R and sklearn.mixture.GaussianMixture in Python). Visualize the different partitions on the first few principal components. Simulate data that are better suited for your algorithm than for k-means. Comment on the results.

Additional Questions

Extend to a generalized k-means algorithm in $\mathbb{R}^p$ where the class representatives are hyperplanes of dimension $q < p$:

- calculations,
- codes,
- illustrations.

References

Diday, E. (1973). The dynamic clusters method in nonhierarchical clustering. International Journal of Computer & Information Sciences, 2(1), 61-88.

Bock, H. H. (2007). Clustering methods: a history of k-means algorithms. Selected contributions in data analysis and classification, 161-172. http://www.modulad.fr/dac/Slides/Bock/Ahistoryofthek-means.pdf