

TP de Machine Learning : Classification de Fruits avec K-Means

Objectif

L'objectif de ce TP est de se familiariser avec l'utilisation de l'algorithme K-Means pour la classification d'images. Nous classerons des fruits. Vous devrez appliquer K-Means sur un ensemble d'images de fruits, augmenter leur jeu de données en utilisant des transformations d'images, regrouper les fruits similaires en clusters et évaluer la qualité de la classification.

Prérequis

En plus de scikit-learn, vous aurez besoin des bibliothèques suivantes : Pillow, matplotlib et numpy.

1. Téléchargement du jeu de données

Téléchargez l'ensemble de données "fruits.zip" depuis et extrayez-le dans un dossier local. Cet ensemble de données contient des images de fruits dans différentes catégories.

2. Chargement des images

Écrivez un script Python pour charger les images du jeu de données dans un format adapté à l'utilisation de scikit-learn. Utilisez la bibliothèque Pillow pour cela. Assurez-vous que les images sont chargées sous forme de matrices de pixels.

3. Prétraitement des images

Afin de réduire la complexité, redimensionnez toutes les images à une taille fixe (par exemple, 100x100 pixels). Vous pouvez également normaliser les valeurs de pixel.

4. Augmentation du jeu de données

Utilisez des techniques d'augmentation de données pour générer de nouvelles variations des images de fruits. Appliquez des transformations telles que la rotation, le changement d'échelle, le recadrage, le changement de luminosité, pour augmenter la diversité du jeu de données. Vous aurez par exemple besoin des fonctions suivantes de la bibliothèque Pillow : `rotate()`, `resize()`, `crop()`.

5. Application de K-Means

Appliquez l'algorithme K-Means sur les images prétraitées et augmentées pour regrouper les fruits en clusters. Choisissez un nombre de clusters K approprié. Vous pouvez expérimenter avec différentes valeurs de K.

6. Visualisation des clusters

Utilisez Matplotlib pour afficher quelques images de chaque cluster pour évaluer visuellement la qualité de la classification.

7. **Évaluation de la classification**

En utilisant l'inertie, évaluez la qualité de la classification. L'inertie est une mesure de la cohésion des clusters. Plus l'inertie est faible, plus les clusters sont cohérents. L'inertie est calculée comme la somme des distances au carré de chaque exemple au centre de son cluster. Vous pouvez utiliser la fonction `KMeans.inertia_` de scikit-learn pour calculer l'inertie.

8. **Deuxième jeu de données**

Répétez les étapes 1 à 7 avec l'ensemble de données "fruits-vegetables.zip" qui contient plus des fruits et des légumes.