TER | UNIVERSITÉ PARIS DESCARTES

Machine Learning for Data Science

Factorisation matricielle non-négative et classification d'images

Mickael Febrissy et Mohamed Nadif

1 Contexte

La classification automatique ou *clustering* consiste à partionner un ensemble d'objets (instances) décrits par un ensemble de variables en groupes (classes) homogènes. Avec l'avènement du Big Data et de la science des données, le *clustering* est devenu une tâche encore très importante dans divers domaines dont l'imagerie.

Les images sont des données très répandues notamment sur le web et les réseaux sociaux (Instagram, Pinterest, Flickr, Google, etc). Le but sera de proposer un système de classification pour des images provenant de divers bases de données (photos, peintures, bandes dessinées, etc).

La factorisation matricielle non négative permet d'approximer une matrice de données positive par le produit de deux matrices de dimensions inférieures et positives. Par sa simplicité, cette méthode est devenue populaire et est utilisée à la fois dans la réduction de la dimension et également dans la classification automatique (clustering) en un un nombre de classes k fixé par l'utilisateur.

Soit $\mathbf{X}$ une matrice de données positives $\in \mathbb{R}^{n\times p}$, on recherche une matrice positive $\mathbf{W} \in \mathbb{R}^{n\times k}$ et une matrice positive $\mathbf{H} \in \mathbb{R}^{p\times k}$ tel que :

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}^{\top}$$

$\mathbf{W}$ et $\mathbf{H}$ sont respectivement les matrices des bases et des coefficients ayant en commun un rang fixé k. Ces matrices peuvent être estimées sous couvert de différentes fonctions objectives.

2 Travail à réalisé

Ci-après une description des différentes étapes à réaliser.

1. Dans un premier temps, il conviendra de construire une matrice numérique retranscrivant plusieurs milliers d'images. Elle pourra être dupliquée sous plusieurs versions en fonction des différents filtres appliqués (binarisation, niveaux de gris, etc).
2. Une étude comparative entre différentes méthodes d'initialisation et d'estimation de paramètres pour la NMF impliquant la SVD (Singular Vaue Decomposition), la PCA (Principal Component Analysis) et le Spherical K-means sera établie. Une visualisation nécessitant la retranscription en images des différents résultats obtenues sera nécessaire.
3. Troisièmement, la classification des données sera effectuée et une étude comparative des résultats des différents algorithmes sera réalisée en utilisant des critères d'évaluation appropriés.

Une description des données (taille, dimension etc.) doit être précisée, les méthodes utilisées doivent être maitrisées et enfin tous les résultats obtenus doivent être rigoureusement commentés.

3 Mots-clés

Apprentissage non supervisé, Spherical K-means, K-means, R, NMF, PCA, Python, Scikit-Learn, Nimfa, Pillow, Jython.

4 Références

`https://pypi.python.org/pypi/Pillow/5.0.0`
`https://www.flickr.com/services/api/`
`https://github.com/BathVisArtData/PeopleArt`