Convex relaxations of K-means

Nicolas Verzelen

INRA, UMR 729 MISTEA, F-34060 Montpellier, FRANCE

Résumé. Le problème du clustering revient à regrouper des objets (points dans l'espace, noeuds d'un graphe, lignes d'un tableau) partageant des propriétés similaires. Parmi les nombreuses méthodes dédiées au clustering, le critère du K-means, ainsi que sa minimisation par l'algorithme de Lloyd sont particulièremet populaires. Dans cet exposé de synthèse, je détaillerai des arguments de Peng et Wei [8] étudiant différent relaxations convexes du problème du K-means. Une première relaxation aboutit à un programme semi-défini (SDP), tandis qu'une seconde relaxation correspond aux méthodes de clustering spectral.

Mots-clés: clustering, clustering spectral, relaxation convexe, SDP

Abstract. The problem of clustering is that of grouping objects (points, nodes of a graph, lines of a matrix) that share similar properties. In an Euclidean space, the K-means criterion and its minimization based on Lloyd's algorithm are really popular. In this talk, it will detail some arguments of Peng and Wei [8] that pinpoint the connections between some classical clustering algorithms such as the minimization of the K-means criterion, spectral algorithms and SDPs.

Keywords: clustering, spectral clustering spectral, convex relaxations, semi-definite programming

1 Introduction

Let $X_1, \ldots, X_n$ be n points in $\mathbb{R}^p$. Given a positive integer $K > 0$, the problem of clustering amounts to build a partition $G = (G_1, \ldots, G_K)$ of $[n]$ into K subsets in such a way that points in the same group are closer than points in different groups.

One classical approach for this problem amounts to pick the partition $\widehat{G}$ minimizing the K-means criterion defined by

$$\widehat{G} := \arg\min_{G} \sum_{k=1}^{K} \sum_{a \in G_k} \|\mathbf{X}_a - \overline{\mathbf{X}}_{G_k}\|_2^2 \ , \tag{1}$$

where $\overline{\mathbf{X}}_{G_k} = \frac{1}{|G_k|}\sum_{b \in G_k} \mathbf{X}_b$ is the mean of the points $\mathbf{X}_b$ in group G_k. unfortunately, minimizing (1) over the collection of all partitions is known to be hard [3]. Iterative methods such as Lloyd's algorithm [6] or its variants [2] may only converge to local minima of the objective function. There is therefore a need for tractable clustering procedures. The purpose of this talk is to explain how spectral algorithms [1] and SDPs [7] formulations can be interpreted as convex relaxations of the K-means problem.

2 From K-means to Spectral algorithm and SDP

To describe these relaxations, it is helpful to consider the clustering matrix defined by

$$B^*_{ab} = \begin{cases} \frac{1}{|G_k|} & \text{if } a \text{ and } b \text{ are in the same group } G_k, \\ 0 & \text{if } a \text{ and } b \text{ are in a different group.} \end{cases}$$

The groups in a partition G are in a one-to-one correspondence with the non-zero blocks of B^*. Up to a permutation, B^* is block diagonal and is constant on each block. Note that defining B^* allows, in some way, to circumvent the problem of label switching. We start by rewriting the K-means criterion as a minimization problem involving matrices B. In the sequel, $\widehat{\Sigma} := \frac{1}{p}\mathbf{X}\mathbf{X}^T$ stands for the empirical covariance matrix associated to $\mathbf{X}$. Simple algebra leads to the following result.

Proposition 1. *The partition $\widehat{G}_{K-M}$ associated to the minimization of the K-means criterion also corresponds to*

$$\widehat{B}_{K-M} := \arg\min_{B \in \mathcal{C}_0} \langle \widehat{\Sigma}, B \rangle \ ,$$

where $\mathcal{C}_0 := \{B \in \mathbb{R}^{n \times n} : B \text{ is a clustering matrix}\}$.

Obviously, the domain $\mathcal{C}_0$ is not convex. Upon defining the convex set

$$\mathcal{C}_1 := \left\{ B \in \mathbb{R}^{n\times n} : \begin{array}{l} \bullet\ B \succcurlyeq 0 \ \ \text{(symmetric and positive semidefinite)} \\ \bullet\ \sum_a B_{ab} = 1,\ \forall b \\ \bullet\ B_{ab} \geq 0,\ \forall a, b \\ \bullet\ \mathrm{tr}(B) = K \end{array} \right\} ,$$

we consider the estimator

$$\widehat{B}_{SDP} := \arg\min_{B \in \mathcal{C}_1} \langle \widehat{\Sigma}, B \rangle \ ,$$

which is the solution of a SDP. We also consider the second domain

$$\mathcal{C}_2 := \left\{ B \in \mathbb{R}^{n\times n} : \begin{array}{l} \bullet\ I \succcurlyeq B \succcurlyeq 0 \ \ \text{(symmetric and positive semidefinite)} \\ \bullet\ \mathrm{tr}(B) = K \end{array} \right\} ,$$

to define the estimator

$$\widehat{B}_{Spec} := \arg\min_{B \in \mathcal{C}_2} \langle \widehat{\Sigma}, B \rangle$$

Note that the solution $\widehat{B}_{Spec}$ is the orthogonal projection matrix onto the space spanned the eigenvectors of $\widehat{\Sigma}$ associated to its K largest eigenvalues. As a consequence, $\widehat{B}_{Spec}$ can be easily computed by a singular value decomposition, the construction of $\widehat{B}_{Spec}$ corresponds to the first step of a spectral clustering algorithm.

Proposition 2. *[8] We have the following inclusions*

$$\mathcal{C}_0 \subset \mathcal{C}_1 \subset \mathcal{C}_2$$

This entails that the spectral estimator $\widehat{B}_{Spec}$ is a convex relaxation of the SDP estimator $\widehat{B}_{SDP}$, which is in turn is a convex relaxation of the K-means estimator $\widehat{B}_{K-M}$.

Note that the estimators $\widehat{B}_{SDP}$ and $\widehat{B}_{Spec}$ do not necessarily belong to $\mathcal{C}_0$ and therefore may not define a proper partition of $[n]$. Nevertheless, one can easily estimate a partition from any of these two matrices by applying for instance Lloyd's algorithm to $\widehat{B}_{SDP}$ or $\widehat{B}_{Spec}$.

Analyses of the SDP estimator $\widehat{B}_{SDP}$ may be found in [7] or [4] whereas clustering bounds for $\widehat{B}_{Spec}$ have been derived e.g. in [1].

3 Extension to other models

One can follow a similar strategy to interpret some graph clustering algorithms such as those of [5] as convex relaxations of a suitably defined K-means criterion.

References

[1] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[2] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[3] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. Kemal Sinop. The Hardness of Approximation of Euclidean k-means. *ArXiv e-prints*, February 2015.

[4] S. Chrétien, C. Dombry, and A. Faivre. A Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering. *ArXiv e-prints*, June 2016.

[5] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 2015.

[6] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[7] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *ArXiv e-prints*, February 2016.

[8] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. on Optimization*, 18(1):186–205, February 2007.