KML Practical Work

Maria Gkotsopoulou & Ricard Monge

01/12/2019

Abstract

Introduction

Previous work

Theory

The k-means clustering algorithm is one of the most commonly used clustering methods providing solid results but also having some drawbacks. Having a sample $\{x_1, ..., x_n\}$ and k clusters π_j for $j = 1, ..., k$, we assign a point x_i to a cluster iff:

$$k = argmin_j\{d(x_i, c_j)\} = argmin_j\{\|x_i - c_j\|^2\}$$

$$where\ c_j = \frac{1}{|\pi_j|}\sum_{x\in\pi_j} x$$

As such, this algorihtm minimizes the within clusters sum of squares:

$$WSS = \sum_{j=1}^{k}\sum_{x\in\pi_j}\|x - c_j\|^2$$

A major drawback of k-means is that it cannot separate clusters that are not linearly separable in input space. Both kernel k-means and spectral clustering address this problem

Kernel K-means

Our aim is to use the kernel trick to project the data points in the input space onto a higher dimensional feature space in which the points are linearly separable. What's more, we can apply kernels to non-numeric types of data and use the k-means algorithm to the projected versions directly.

To kernelize the method, we look at the expression of the distance between data points and centers of the clusters in feature space. We denote a scalar product with $< \cdot\ ; \cdot >$ and the feature map as ϕ. It is known that it can be expressed as:

$$d(x, y) = \|x - y\|^2 =< x - y\ ;\ x - y >= ... =$$
$$= k(x, x) + k(y, y) - 2k(x, y)$$

where $k(x, y) =< \phi(x), \phi(y) >$ is the kernel function, inner product of points in the feature space. Next, we express the distance of a point x_i to the center c_j of the cluster ϕ_j:

$$\|\phi(x_i)-\phi(c_j)\|^2 =< \phi(x_i)-\phi(c_j)\ ;\ \phi(x_i)-\phi(c_j) >= ... =$$
$$= k(x_i, x_i) + f(x_i, c_j) + g(c_j)$$

where the functions f represent a sort of point-cluster kernel distance

$$f(x_i, c_j) = \frac{-2}{|\pi_j|}\sum_{l=1}^{n} z(x_l, \pi_j)k(x_i, x_l)$$

with $z(x_l, c_j)$ indicator function which is 1 if x_l belongs to π_j and 0 otherwise. On the other hand, g represent a sort of within-cluster kernel distance:

$$g(c_j) = \frac{1}{|\pi_j|^2}\sum_{l=1}^{n}\sum_{m=1}^{n} z(x_l, \pi_j)z(x_m, \pi_j)k(x_l, x_m)$$

From the expression of $\|\phi(x_i) - \phi(c_j)\|^2$ we see that $k(x_i, x_i)$ does not change with the cluster considered. Furthermore we see g is just cluster depdendent and can be precomputed at each iteration of the algorithm. With these considerations, in the kernelized k-means algorithm we will assign a point x_i to a cluster π_k iff:

$$k = argmin_j\{f(x_i, c_j) + g(c_j)\}$$

We define an indicator matrix Z with as many rows as data points and as many columns as clusters with elements $z_ij = z(x_i, \pi_j)$ given by the previous indicator function. Moreover, we define the kernel matrix K that contains elements $k_ij = k(x_i, x_j)$ given by the kernel function of each pair of points. Finally, we define diagonal matrix L which contains in the diagonal the elements $l_{jj} = 1/|\pi_j|$ with the inverse size of the clusters at any given point. With these definitions, the previous functions f can be easily computed for all points in a matrix F by:

$$F = K \cdot Z \cdot -2L$$

in addition, if we restrict the Kernel matrix to the points of a given cluster π_j in a given time in a matrix G_j, the functions g can be easily compute for all clusters in a vector g by:

$$g = \left(\sum G_j\right) \cdot l_{jj}^2$$

where $\sum G_j$ denots the sum of all elements of G_j.

Spectral clustering

Given the previous data sample $\{x_1, ..., x_n\}$ and a similarity matrix S with elements s_{ij} that measure similarity between points x_i and x_j, one can view the dataset as a graph with each point as a node and each edge as the directed connections from x_i to x_j with the weight given by s_{ij}. If two points have similarity $s_{ij} = 0$ they won't be connected. We assume the similiarity matrices in the following discusion are symmetric and thus the graph is undirected.

We define the degree of a node (point) x_i of this graph as the sum of it's outgoing edge's weights, given by:

$$d_i = \sum_{j=1}^{n} s_{ij}$$

and build the diagonal matrix D with the degrees of each node at the diagonal, and 0 elsewhere.

Given the undirected graph associated with the previous data point sample and its similarity matrix S, we define the following laplacian matrices:

- Unnormalized laplacian matrix $L = D - S$ which considers the similarities between each pair of points without considering the similarities of the points with themselves.
- Normalized Symmetric laplacian matrix $L_{sym} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}SD^{-1/2}$
- Normalized Random Walk laplacian matrix $L_{rw} = D^{-1}L = I - D^{-1}S$

this matrices give insights on the properties of the graph.

The problem of clustering the data points $\{x_1, ..., x_n\}$ can be viewed as graph partition problem into groups such that the similarities between points of different groups are small and between points of the same group are big.

Experiments

Results

Conclusions & future work

References

"A Large Scale Clustering Scheme for Kernel K-Means." 2002. In *Proceedings of the 16 Th International Conference on Pattern Recognition (Icpr'02) Volume 4 - Volume 4*, 40289. ICPR '02. Washington, DC, USA: IEEE Computer Society. http://dl.acm.org/citation.cfm?id=846227.848565.

Cardoso-Cachopo, Ana. 2007. "Improving Methods for Single-label Text Categorization." PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa.

Dhillon, Inderjit S., Yuqiang Guan, and Brian Kulis. 2004. "Kernel K-Means: Spectral Clustering and Normalized Cuts." In *Proceedings of the Tenth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 551–56. KDD '04. New York, NY, USA: ACM. https://doi.org/10.1145/1014052.1014118.

Karatzoglou, Alexandros, and Ingo Feinerer. 2006. "Text Clustering with String Kernels in R." In *GfKl*.

Luxburg, Ulrike. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17 (4): 395–416. https://doi.org/10.1007/s11222-007-9033-z.

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. 2001. "On Spectral Clustering: Analysis and an Algorithm." In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 849–56. NIPS'01. Cambridge, MA, USA: MIT Press. http://dl.acm.org/citation.cfm?id=2980539.2980649.

Shi, Jianbo, and Jitendra Malik. 2000. "Normalized Cuts and Image Segmentation." *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (8): 888–905. https://doi.org/10.1109/34.868688.

Welling, Max. 2013. "Kernel K-Means and Spectral Clustering." *2013-03-15]. Http://Www, Ics. Uci. Edu/-Welling/Teaching/273 ASpring09/SpectralClustering. Pdf.*