Convex Clustering with ADMM implementation via R package

Please see the enclosed pdf for correct rendering of all the math equations.

Description of Package

This R package implements ADMM algorithm to solve the convex clustering problem. Given n data points x_i, $i = 1, \ldots, n$, in p dimensions, $x_i \in \mathbb{R}^p$, the key idea behind the convex clustering model is that if two observations x_i and x_j belong to the same cluster, then their corresponding centroids u_i^* and u_j^* should be the same. The optimal solution $U^* = [u_i^*, \ldots, u_n^*]$.

It aims to minimize

$$\text{minimize}_{U \in \mathbb{R}^{n \times p}} \frac{1}{2} \sum_{i=1}^{n} \|x_i - u_i\|_2^2 + \gamma \sum_{i<j} w_{ij} \|u_i - u_j\|$$

where

$$w_{ij} = exp(-mu \times \|x_i - x_j\|_2^2)$$

are the non-negative gaussian kernel weights set between the ith and jth points used for implementation in the package. *gamma* is the positive tuning parameter, u_i is the cluster centroid which is u^{th} column of matrix U associated to the point x_i.The first term is the fidelity term while the second term is the regularization term to penalize the differences between different centroids in order to make sure that centroids for observations in the same cluster should be identical.

Reference:

E. C. Chi and K. Lange. Splitting methods for convex clustering. J. Computational and Graphical Statistics, 24(4):994–1013, 2015.

The algorithm implemented in the package is taken from the above paper. The main function takes p by n data matrix X gamma, (optional) initial p by n matrix of cluster centers U, p by nk matrix of cluster adjacency matrix V, p by nk matrix of starting value of H as arguments.

The algorithm can be run for the specified maximum number of iterations. The algorithm is not guaranteed to find the absolute minimizer. The function returns three matrices after ADMM algorithm is run, which are the centroid matrix for each data point(U), centroid difference matrix(V), and Lagrange multipliers matrix.

Algorithm's Implementation

To minimize the above objective, ADMM algorithm which employs variable splitting in order to account for the shrinkage penalties in the convex clustering problem.

The minimization problem is made into a equivalent constrained problem as follows:

$$\text{minimize} \frac{1}{2} \sum_{i=1}^{n} \|x_i - u_i\|_2^2 + \gamma \sum_{l \in \epsilon} w_l \|v_l\|$$

subject to

$$u_{l1} - u_{l2} - v_l = 0$$

where l is a centroid pair ($l1$,$l2$) and

$$\epsilon = \{l = (l1, l2) : w_l > 0\}$$

ADMM updates

Augmented Lagrangian: given $\tau > 0$

$$L_\tau(U, V, H) = \frac{1}{2}\sum_{i=1}^{n}\|x_i - u_i\|_2^2 + \gamma\sum_{l\in\epsilon} w_l\|v_l\| + \sum_{l\in\epsilon}(H_l, v_l - u_{l1} + u_{l2}) + \frac{\tau}{2}\sum_{l\in\epsilon}(\|v_l - u_{l1} + u_{l2}\|_2)$$

Update of U:

$$U_i = \frac{1}{1+n\tau}y_i + \frac{1}{1+n\tau}\bar{x}$$

where

$$y_i = x_i + \sum_{l1=i}[H_l^t + \tau V_l^t] - \sum_{l2=i}[H_l^t + \tau V_l^t]$$

$$U^{t+1} = \frac{1}{1+n\tau}Y + \frac{1}{1+n\tau}\bar{x}$$

Update of V:

$$V_l^{t+1} = \arg\min_{V_l}\{\frac{\gamma w_l}{\tau}\|v_l\| + \frac{1}{2\tau}(\|V_l - (u_{l1} - u_{l2} - \frac{H_l}{\tau})\|_2^2\}$$

$$V_l^{t+1} = prox_{\frac{\gamma w_l}{\tau}}(u_{l1}^{t+1} - u_{l2}^{t+1} + \frac{H^t}{\tau})$$

Update of H:

$$H_l^{t+1} = H_l^t + \tau(v_l^{t+1} - u_{l1}^{t+1} + u_{l2}^{t+1})$$

Installation

devtools::install_github("sowmyakolluru/ConvexCluster")

Example usage Of Package

Test on multidimensional case

```
X = matrix(c(rep(1, 10), rep(0, 10)))
multipleDimensionX = t(cbind(X, X, X))
w <- ComputeWeights(multipleDimensionX, mu = 1)
n <- ncol(multipleDimensionX)
p <- nrow(multipleDimensionX)
nk <- nrow(w)
H <- matrix(0,p,nk)
V <- matrix(0, nrow = p, ncol = nk)
gamma <- 0.5
tau <- 1.0
convexADMM(multipleDimensionX,H,U = NULL,V,w,gamma,tau,num_iter=100)
```