ST 790 Assignment 4

David Elsheimer and Jimmy Hickey

11/12/2020

Instruction

This assignment consists of 3 problems. Choose 2 out of the 3 problems. The assignment is due on **Friday, November 12** at 11:59pm EDT. Please submit your assignment electronically through the **Moodle** webpage. The assignment can be done as a group with at most 3 members per group (please include the name of the group members on the front page of the assignment).

Problem 1

Skim the following article. In particular take a look at Algorithm 1 in the referenced paper. Next download the **DBLP4** dataset from here. Now try to analyze this dataset using the Algorithm 1 (SVM-Cone) form the referenced paper (see also section 4.1.2 of the referenced paper). For the evaluation metric, use rank correlation (see section 4.1.2. of the referenced paper).

Note This problem might be a tad tricky. It is perfectly fine if your results are not as good as that presented in the paper. As long as you are not getting a performance that is worse than chance, it should be fine.

```
A_edge <- as.matrix(read.delim("DBLP4_adjacency.txt", header=FALSE)) #edgelist
A_graph <- graph.data.frame(A_edge, directed=FALSE) #graph object
A<- as_adjacency_matrix(A_graph) #Adjacency matrix
#A_f <- norm(as.matrix(A), type="F") #frobenius norm

theta<- as.matrix(read.delim("DBLP4_community.txt", sep= " ", header=FALSE))

colnames(theta) <- c("row", "community", "val")

decomp<- spectrum(A_graph, which=list(pos="LM", howmany=3)) #500 works!
eigval <- decomp$values
eigvec <- decomp$vectors

#sqrt(sum(eigval^2))/A_f #Needs to be >.8

Zhat <- eigvec

###Need to perform row normalization
rownorm <- function(i){
  (Zhat[i,]-mean(Zhat[i,]))/sd(Zhat[i,])
  }

Yhat <- t(sapply(1:nrow(Zhat), rownorm))
```

```
model <- svm(Yhat, type='one-classification', kernel="linear")

w <- t(model$coefs) %*% model$SV
b <- -model$rho

Yhat_w <- Yhat %*% t(w)

delta <- 1
filtered <- Yhat_w[(Yhat %*% t(w) < b)==TRUE]

which((Yhat %*% t(w) < b)==TRUE)[c(3,232,925)]
```

```
## [1]    7  332 1960
```

```
##clustering

k3 <- kmeans(filtered, centers = 3, nstart = 25)

###Need a point from each cluster
#k3$cluster
#lets just choose 1, 925, 232

###Feeding that back into Yhat to get M...

Yhat_c <- Yhat[c(3,232,925),]
Mhat <- Zhat %*% t(Yhat_c)%*%solve(Yhat_c%*%t(Yhat_c)+0.00000001)

#theorem 3.2
#need proper indices again
which((Yhat %*% t(w) < b)==TRUE)[c(3,232,925)]
```

```
## [1]    7  332 1960
```

```
Vhat_c <- Zhat[c(7, 332, 1960),]
Nhat_c <- t(sapply(1:nrow(Vhat_c), rownorm))

D <-sqrt(diag(Nhat_c%*%Vhat_c%*%diag(eigval)%*% t(Vhat_c)%*%t(Nhat_c)))
Dhat <- diag(D)

Fhat <-  Mhat%*%Dhat%*%as.matrix(rep(1,3))


thetahat <- solve(Diagonal(length(Fhat),Fhat))%*%Mhat%*%Dhat

theta_empty <- Matrix(0,nrow(thetahat),3)


thetaval_j <- function(j){
  #theta_empty[which(theta[,2]==j),j] <-
  temp <- theta[which(theta[,2]==j),c(1,3)]
  theta_empty[temp[,1],j] <- temp[,2]
  theta_empty
```

```
}

theta_empty <- thetaval_j(1)
theta_empty <- thetaval_j(2)
theta_empty <- thetaval_j(3)

theta_proper <- theta_empty

avgrankcorr <- function(sigma){
  summand <- cor(thetahat[,1], theta_proper[,sigma[1]], method="spearman")
  summand <- c(summand,cor(thetahat[,2], theta_proper[,sigma[2]], method="spearman"))
  summand <- c(summand,cor(thetahat[,3], theta_proper[,sigma[3]], method="spearman"))
  mean(summand)
}


sigmas <- list(c(1,2,3),
               c(2,3,1),
               c(3,1,2),
               c(1,3,2),
               c(2,1,3),
               c(3,2,1))
sums <- sapply(sigmas,avgrankcorr)
```

Using the code above $\hat{\Theta}$ was generated using algorithm 1 and additional definitions as laid out in the paper. Based on how rank correlation is defined in the paper, the average rank correlation between $\hat{\Theta}, \Theta$ here is 0.0464011, which corresponds to the permuation c(3, 2, 1).

Problem 2

Skim the following article (maybe only the first 16 pages, unless you really have time and care about the theory). Now take a look at section 4.1 on model selection for stochastic blockmodels. Next try to reproduce Table 1 (or a part of Table 1) but only for the ECV algorithm (Algorithm 3) with L_2 loss; you don't need to consider L_2 loss with stability.

Problem 3

Let $\mathbf{A}_1$ and $\mathbf{A}_2$ be two-blocks stochastic blockmodel graphs, with block probability matrices

$$\mathbf{B}_1 = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix}$$

Suppose for simplicity that the first $n/2$ vertices of both $\mathbf{A}_1$ and $\mathbf{A}_2$ are assigned to block 1 and that the last $n/2$ vertices of both $\mathbf{A}_1$ and $\mathbf{A} - 2$ are assigned to block 2. Assume that $\mathbf{A}_1(i,j)$ and $\mathbf{A}_2(k,\ell)$ are independent of one another if $\{i,j\} \neq \{k,\ell\}$. Finally, $\mathbf{A}_1(i,j)$ and $\mathbf{A}_2(i,j)$ are correlated with correlation $\rho \in [-1,1]$. We assume that ρ is the same for all $\{i,j\}$ pairs.

- Given $\mathbf{A}_1$ and $\mathbf{A}_2$, formulate a test statistic for testing $\mathbb{H}_0 : \rho = 0$ against the alternative hypothesis that $\mathbb{H}_1 : \rho \neq 0$. In other words, test the hypothesis that $\mathbf{A}_1$ is independent of $\mathbf{A}_2$.

- Do you think your test procedure is valid and consistent as $n \to \infty$? (assuming that the parameters of $\mathbf{B}_1$ and $\mathbf{B}_2$ and ρ are kept constant).
- How would you adapt this procedure when the block assignments are unknown, or, in addition, if the graphs involved are degree corrected SBMs instead of the vanilla SBMs ?