Wednesday October 2, 2013

Last time: Blocking; Capture-Recapture; Work on Projects

Today: Transitivity; Clustering Records; Linking more than 2 lists

Next Time: Building Supervised Models when you have too much data

Transitivity:

Clustering: grouping records that correspond to same entity;
unlike most clustering problems, interested in lots of small clusters

Hierarchical Agglomerative Clustering:

Example

```
iris[1,]; help(iris)
pairs(iris,col=as.numeric(iris$Species))
dist.iris<-dist(iris[,1:4])  ##needs to be distance object

hc.sing<-hclust(dist.iris,method="single")
plclust(hc.sing,labels=c(rep(1,50),rep(2,50),rep(3,50)))
group.assn<-cutree(hc.sing, h=1.0)
table(group.assn,iris$Species)
group.assn.2<-cutree(hc.sing,k = 3)
table(group.assn.2,iris$Species)

hc.comp<-hclust(dist.iris,method="complete")
```

How do we do this for record linkage?

Only using a small subset of the RL500; keeping the indices

```
data<-read.table("JWLabel.txt"); dim(data)

ones<-sample(which(data$identity.vec==1),20)
zeros<-sample(which(data$identity.vec==0),70)
obs<-rbind(cbind(index1.vec[ones],index2.vec[ones]),
cbind(index1.vec[zeros],index2.vec[zeros]))
data2<-rbind(data[ones,],data[zeros,])

mod<-glm(identity.vec~by.vec+bd.vec+bm.vec,data=data2,family="binomial")
mod$fit ##prob values

##Building "distance" matrix
prob.matrix<-matrix(0,90,90)
for(i in 1:90){
for(j in 1:90){
by.val<-JW(as.character(RLdata500[obs[i,1],5]),as.character(RLdata500[obs[j,2],5]))
bm.val<-JW(as.character(RLdata500[obs[i,1],6]),as.character(RLdata500[obs[j,2],6]))
bd.val<-JW(as.character(RLdata500[obs[i,1],7]),as.character(RLdata500[obs[j,2],7]))
pred.log.odds<-predict(mod,data.frame(by.vec=by.val,bd.vec=bd.val,bm.vec=bm.val))
prob.matrix[i,j]<-exp(pred.log.odds)/(1+exp(pred.log.odds))
}
print(i)
}

dist.m<-1-prob.matrix
hc<-hclust(as.dist(dist.m),method="single")
plclust(hc)
groups<-cutree(hc,h=0.10)
table(groups)
```