Assignment_1

Keshav Padiyar Manuru

03/01/2021

Assignment 1. Variable selection with randomized LASSO

Here we are trying to enhance the chances of right variable selection there by minimizing elimination of highly correlated variables. As mentioned in the paper (https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2010.00740.x)s

The randomized lasso is a new generalization of the lasso. Whereas the lasso penalizes the absolute value $|\beta_k|$ of every component with a penalty term proportional to λ, the randomized lasso changes the penalty λ to a randomly chosen value in the range $[\lambda, \frac{\lambda}{\alpha}]$.

In-order to achieve the said goal following steps are performed:

1. Take samples from dataset without replacement using non parametric bootstrap (here 100 iterations).
2. Generate weights in-order to randomly penalize the $|\beta_k|$. Where weight W_k be IID random variables in $[\alpha, 1]$ here $\alpha = 0.5$ and its called as weakness parameter.

$$\hat{\beta}^{\lambda,W} = arg\ min(||Y - X\beta||_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k})$$

3. Run the lasso model with bootstrapped data for every single $\lambda\ \epsilon\ 0.05 - 1$ over step of 0.05 also, add above generated weights to the model.
4. For every λ value calculate the probability of the significance of the variable. It can be achieved by, assigning the value $= 1$ all the variables for every iteration whose coefficients are non zero for a given λ. Then by taking the sum over the variable values and dividing it by total number of iterations we get the probabilities.
5. Repeat the step 4 for all λ values, then finally take the maximum probability for all the variables irrespective of λ
6. Inorder to select the stable variables, we have set the threshold to be **0.7**. Variables with their maximum probability > 0.7 are selected as stable variables

1. Reading data excluding ***total_UPDRS***, scaling to zero and unit variance.

```
park <- read.csv("parkinsons.csv")
park <- park %>% select(-total_UPDRS) %>% scale() %>% data.frame()
lambda <- seq(0.05,1,by=0.05)
```

2. Defining function to implement randomized lasso.

In the function *f1* glmnet package is used to implement lasso regression with weakness parameter $\alpha = 0.5$

$$W_k\ be\ IID\ random\ variables\ in[\alpha, 1]$$

```
f1 <- function(data,ind,lambda,alpha){

  data1 <- data[ind,]

  x_train <- as.matrix(data1 %>%select(-motor_UPDRS))

  y_train <- data1$motor_UPDRS

  # generating w_k
  w <- runif(ncol(x_train),alpha,1)

  m1 <- glmnet(x=x_train,y=y_train,lambda=lambda,alpha=1,
               family = "gaussian", panalty.factor = 1/w)

  return(matrix(as.vector(t(coef(m1))),ncol=ncol(x_train)+1,nrow=1))

}
```

3 Executing the model for different values of lambda and on bootstrap samples

```
for (l in 1:length(lambda)){

  m2 <- boot(park, f1, R =100,lambda=lambda[l],alpha=0.5)

  coefs<-m2$t

  if (l ==1){

    max_probabilities <- matrix(colSums
                                (apply(coefs,2
                                    ,function(x) ifelse(x!=0,1,0)))/100,nrow=1)

  }else{

    max_probabilities <- rbind(max_probabilities
                               ,matrix(colSums(apply(coefs,2
                              ,function(x) ifelse(x!=0,1,0)))/100,nrow=1))
  }

}

max_probabilities <- max_probabilities[,-1]

rownames(max_probabilities) <- lambda

colnames(max_probabilities) <- colnames(park %>% select(-motor_UPDRS ))
```

Table 1: Probabilities of Vairable Significance for different Lambda values

	subject.	age	sex	test_time	Jitter...	Jitter.Abs.	Jitter.RAP	Jitter.PPQ5	Jitter.DDP	Shimmer	Shimmer.dB.	Shimmer.APQ3	Shimmer.APQ5	Shimmer.APQ11	Shimmer.DDA	NHR	HNR	RPDE	DFA	PPE
0.05	1.00	1.00	1.00	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0.63	0.13	1.00	1.00
0.1	1.00	1.00	0.01	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0.02	0.97	0.98
0.15	1.00	1.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0.00	0.00	0.08
0.2	1.00	1.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.25	0.55	0.99	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.3	0.00	0.02	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.35	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.4	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.45	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.5	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.55	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.6	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.65	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.7	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.75	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.8	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.85	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.9	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
0.95	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00

```r
max_probabilities <- matrix(apply(max_probabilities,2,max),nrow=1,
                            dimnames = list("probability",colnames(park %>% select(-motor_UPDRS ))))

stable <- max_probabilities[,which(apply(max_probabilities,2,function(x) x>0.7))]

kable(t(stable), caption = "Stable Variable Set and their Probabilities")
```

Table 2: Stable Variable Set and their Probabilities

subject.	age	sex	test_time	DFA	PPE
1	1	1	0.95	1	1