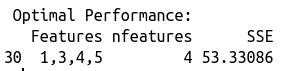
Group\_A11

Sridhar, Naveen, Obaid

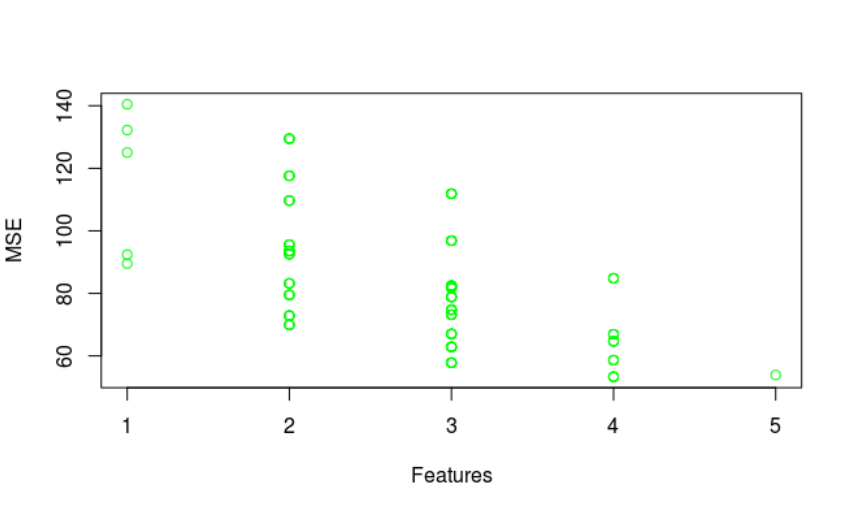
24 November 2018

### Assignment 3

As we increase the number of features, the MSE shows a decreasing trend. The optimal subset are features 1,3,4 and 5 as they evaluate to lowest MSE of 53.33 than any other feature subset. Therefore, these features have major impact on the target.



MSE



Feature Subset

### Appendix

#Assignment 3  
linreg <- function(x, y){  
   
 if(!is.matrix(x)){  
 x = as.matrix(x)  
 }  
 if(!is.matrix(y)){  
 y = as.matrix(y)  
 }  
   
 x = cbind(x,1)  
   
 lmlist = list()  
   
 lmlist[['coeffs']] = solve(t(x) %\*% x) %\*% t(x) %\*% y  
 lmlist[['preds']]=x %\*% lmlist[['coeffs']]  
 lmlist[['residuals']]=lmlist[['preds']]-y  
 lmlist[['mse']]=mean(lmlist[['residuals']]^2)  
 attr(lmlist, "class") <- "customlm"  
 return(lmlist)  
}  
  
predict.customlm<-function(model,x,y)  
{  
 if (!is.matrix(x))  
 {  
 x=as.matrix(x)  
 }  
   
 x=cbind(x,1)  
   
 lmlist = list()  
 preds <- x%\*%model[["coeffs"]]  
 lmlist[['preds']]=preds  
 lmlist[['residuals']]=preds-y  
 lmlist[['mse']]=mean(lmlist[['residuals']]^2)  
 return(lmlist)  
}  
  
plotcv<-function(data,bins=30,..)  
{  
 library(ggplot2)  
   
 qplot(my\_lm[["residuals"]], geom="histogram",bins=bins) + xlab('Residuals values') + ggtitle('Residual distribution')  
   
}  
  
customCV <- function(x,y,Nfolds){  
   
 cvInfo <- data.frame(matrix(ncol = 3, nrow = 0))  
 headers <- c("SelectedFeature","features", "CVScore")  
 colnames(cvInfo) <- headers  
   
 dataset = x  
 totalFeatures <- ncol(x)  
   
 featureList = list()  
 for(i in 1:totalFeatures){  
 features <- c(1:totalFeatures)  
 featureList[[i]]<-features  
 }  
 allcomb <- as.data.frame(expand.grid(featureList))  
   
 allcomb <- t(apply(allcomb, 1, sort)) # sorted  
 allcomb <- allcomb[!duplicated(allcomb), ] # remove duplicate features  
   
 for (i in 1:dim(allcomb)[1]){  
 selectedFeatures <- c()  
 flist = allcomb[i,]  
 names(flist) <- NULL  
 flist <- as.list(flist)  
 flist <- unlist(unique(flist)) # got true features  
 selectedFeatures <- c(selectedFeatures,flist)  
 nfeatures <- length(selectedFeatures)  
   
 X <- as.matrix(dataset[,flist])  
 Y <- y  
   
 folds <- rep\_len(1:Nfolds, nrow(dataset))  
   
 # actual cross validation  
 SSE <- 0  
 for(k in 1:Nfolds) {  
 # actual split of the data  
 fold <- which(folds == k)  
 X.train <- X[-fold,]  
 X.test <- X[fold,]  
   
 Y.train <- Y[-fold,]  
 Y.test <- Y[fold,]  
   
 # train and test your model with data.train and data.test  
   
 model = linreg(X.train,Y.train)  
 results = predict(model,X.test,Y.test)  
 SSE <- SSE + results[["mse"]]  
   
 }  
 SSE <- SSE / Nfolds  
 cvdata<-data.frame(Features=paste(selectedFeatures, collapse = ','),nfeatures,SSE)  
 cvInfo <- rbind(cvInfo,cvdata)  
 }  
   
 print(cvInfo)  
 minCv = cvInfo[which.min(cvInfo$SSE),]  
 cat("\n Optimal Performance: \n")  
 print(minCv)  
 plot(cvInfo$nfeatures,cvInfo$SSE,xlab = "Features",ylab = "MSE",type = "p",col="green")  
}  
  
x <- swiss[,2:6]  
y<- swiss["Fertility"]  
customCV(x,y,5)