Group 7

2022/5/28

```
set.seed(1082)
data=read.csv("crop2x2.csv",header = T)
data.2=read.csv("testCrop2x2.csv",header = T)
data$Label <- factor(data$Label)
```

```
trControl=trainControl(method  = "cv",number  = 5)
```

PCA(training data)

15 PCs can explain 89.14 variation.

```
pca = princomp(data[,1:52],cor=T)
#summary(pca)
```

```
z1 <- pca$scores[,1]
z2 <- pca$scores[,2]
z3 <- pca$scores[,3]
z4 <- pca$scores[,4]
z5 <- pca$scores[,5]
z6 <- pca$scores[,6]
z7 <- pca$scores[,7]
z8 <- pca$scores[,8]
z9 <- pca$scores[,9]
z10 <- pca$scores[,10]
z11 <- pca$scores[,11]
z12 <- pca$scores[,12]
z13 <- pca$scores[,13]
z14 <- pca$scores[,14]
z15<- pca$scores[,15]
pca_data_train = data.frame(z1 = z1, z2 = z2, z3 = z3, z4 = z4, z5 = z5,z6 = z6, z7 = z7, z8 = z8, z9 =
pca_data_train$Label = data$Label
```

PCA(testing data)

```
pca.test <- predict(pca, newdata = data.2[,1:52])
pca.test=pca.test[,1:15]
colnames(pca.test) <- c("z1", "z2", "z3", "z4", "z5", "z6", "z7",
                        "z8", "z9", "z10", "z11", "z12", "z13","z14","z15")
```

QDA

```
qda.fit <- train(Label ~ ., method = "qda"
                 , trControl = trControl
                 ,metric = "Accuracy"
                 , data = pca_data_train)
confusionMatrix(qda.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##             Reference
## Prediction    0   1   2   3   4   5
##          0 343   5   9  64   2   0
##          1  26 280   0  15   6  16
##          2  23   0  47   0   0   0
##          3   3  11   2 144  16   1
##          4   0   0   0   4 122   0
##          5   0   7   0   0   0 354
##
##  Accuracy (average) : 0.86
```

```
pred.qda = predict(qda.fit,newdata = pca.test)
#pred.qda
```

```
write.csv(pred.qda, "QDA_Label.csv", row.names = FALSE)
```

LDA

```
lda.fit <- train(Label ~ .
                 , method = "lda"
                 , trControl = trControl
                 ,metric = "Accuracy"
                 , data = pca_data_train)
confusionMatrix(lda.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
```

```
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 296   3  15  68   7   0
##          1  41 250   0  19   8   0
##          2  40   0  38   0   0   0
##          3  17  12   4 112  11   2
##          4   1   9   1  28 120   0
##          5   0  29   0   0   0 369
##
##  Accuracy (average) : 0.79
```

```r
pred.lda = predict(lda.fit,newdata = pca.test)
#pred.lda
```

```r
write.csv(pred.lda, "LDA_Label.csv", row.names = FALSE)
```

KNN

```r
knn.fit <- train(Label ~ .
                , method = "knn"
                , tuneGrid = expand.grid(k = 5)
                ,trControl = trControl
                , metric = "Accuracy"
                , data = data)
confusionMatrix(knn.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 345  14  39  60   7  15
##          1   3 102   0  29  27  98
##          2   9   0  17   0   0   0
##          3  32   7   2  95  13   3
##          4   0  19   0  20  76  22
##          5   6 161   0  23  23 233
##
##  Accuracy (average) : 0.5787
```

```r
pred.knn = predict(knn.fit,newdata = data.2)
#pred.knn
```

```r
write.csv(pred.knn, "KNN_Label.csv", row.names = FALSE)
```

Random Forest

```r
rf.fit <- train(Label ~ .,method = "rf"
                ,trControl= trControl
                ,metric = "Accuracy"
                ,data = data)
rf.fit
```

```
## Random Forest
##
## 1500 samples
##   52 predictor
##    6 classes: '0', '1', '2', '3', '4', '5'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1200, 1199, 1200, 1200, 1201
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9246863  0.9047260
##   27    0.9299930  0.9116122
##   52    0.9179819  0.8964271
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 27.
```

```r
confusionMatrix(rf.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 385  11  10  17   1   0
##          1   1 269   0  14   2   3
##          2   1   0  48   0   0   0
##          3   8  12   0 189   7   0
##          4   0   3   0   7 136   0
##          5   0   8   0   0   0 368
##
##  Accuracy (average) : 0.93
```

```r
pred.rf=predict(rf.fit,newdata = data.2)
#pred.rf
```

```r
write.csv(pred.rf,"Random forest_Label.csv", row.names = FALSE)
```

Boosting Tree

```r
boosttree.fit <- train(Label ~ .,method = "gbm"
                       ,verbose = FALSE
                       ,trControl= trControl
                       ,metric = "Accuracy"
                       ,data = data)
boosttree.fit
```

```
## Stochastic Gradient Boosting
##
## 1500 samples
##   52 predictor
##    6 classes: '0', '1', '2', '3', '4', '5'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1199, 1202, 1200, 1199, 1200
## Resampling results across tuning parameters:
##
##   interaction.depth  n.trees  Accuracy   Kappa
##   1                   50      0.8880209  0.8585212
##   1                  100      0.9120055  0.8889550
##   1                  150      0.9126633  0.8897279
##   2                   50      0.9260124  0.9064887
##   2                  100      0.9306702  0.9124288
##   2                  150      0.9366725  0.9200309
##   3                   50      0.9353482  0.9185145
##   3                  100      0.9440193  0.9294121
##   3                  150      0.9440082  0.9294072
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were n.trees = 100, interaction.depth =
##  3, shrinkage = 0.1 and n.minobsinnode = 10.
```

```r
confusionMatrix(boosttree.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 383   6   7  11   2   0
##          1   2 276   0  13   3   2
##          2   3   0  51   0   0   0
##          3   7   9   0 199   3   0
##          4   0   1   0   4 138   0
##          5   0  11   0   0   0 369
```

```
##
##  Accuracy (average) : 0.944
```

```r
pred.boosttree=predict(boosttree.fit,newdata = data.2)
#pred.boosttree
```

```r
write.csv(pred.boosttree,"Boosting Tree_label.csv", row.names = FALSE)
```

Naive Bayes

```r
naive.fit=train(Label ~ .,method = "naive_bayes",trControl= trControl,metric = "Accuracy",data = data)
naive.fit
```

```
## Naive Bayes
##
## 1500 samples
##   52 predictor
##    6 classes: '0', '1', '2', '3', '4', '5'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1199, 1200, 1200, 1201, 1200
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.7626967  0.7055388
##    TRUE      0.7740324  0.7188678
##
## Tuning parameter 'laplace' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were laplace = 0, usekernel = TRUE
##  and adjust = 1.
```

```r
confusionMatrix(naive.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 212   2   5  40   5   0
##          1 101 257   0  15   5   6
##          2  44   0  51   0   0   0
##          3  38  22   2 153  13   0
##          4   0   4   0  19 123   0
##          5   0  18   0   0   0 365
##
##  Accuracy (average) : 0.774
```

```r
pred.naive=predict(naive.fit,newdata = data.2)
#pred.naive
```

```r
write.csv(pred.naive,"Naive Bayes_label.csv", row.names = FALSE)
```

LASSO

```r
grid=seq (0,10,0.1)
x =model.matrix(Label ~ ., data)[,-1]
x.new=as.matrix(data.2)
y =data$Label
cv.out=cv.glmnet(x, y,family ="multinomial"
                 ,alpha =1,nfolds=5
                 ,type.multinomial="grouped")
bestlam=cv.out$lambda.min
bestlam
```

```
## [1] 0.00306664
```

```r
train_pred.lasso <- predict(cv.out,s=bestlam,type = "class",newx =x)
# Confusion Matrix and Accuracy
table(train_pred.lasso,data[,53]) ; mean(train_pred.lasso==data[,53])
```

```
##
## train_pred.lasso   0   1   2   3   4   5
##                0 375   7  11  29   3   0
##                1   5 279   0  13   2   1
##                2   3   0  47   0   0   0
##                3  12   5   0 182  12   0
##                4   0   3   0   3 129   0
##                5   0   9   0   0   0 370
```

```
## [1] 0.9213333
```

```r
lasso.pred=predict(cv.out,s=bestlam,type = "class",newx =x.new)
```

```r
write.csv(lasso.pred,"LASSO_label.csv", row.names = FALSE)
```

Forward Selection

```
# allNames <- names(data[,1:52])
# allVar <- paste("~", paste(allNames, collapse=" + "))
#
# multi.fit=multinom(Label~1,data=data, trace = F)
# stepAIC(multi.fit, direction = "forward",trace = FALSE,scope = allVar)
```

```
multi.fit.aic=multinom(formula = Label ~ X2 + X19 + X18 + X5 + X11 + X4 + X40 +X47 + X23 + X41 + X21 + X

train_pred.forward <- predict(multi.fit.aic, data = data)
# Confusion Matrix and Accuracy
table(train_pred.forward,data[,53]) ; mean(train_pred.forward==data[,53])
```

```
##
## train_pred.forward   0   1   2   3   4   5
##                  0 379   8   3  19   0   0
##                  1   4 281   0  12   1   0
##                  2   1   0  54   0   0   0
##                  3  11   6   1 194   7   0
##                  4   0   4   0   2 138   0
##                  5   0   4   0   0   0 371

## [1] 0.9446667
```

```
step.multi.pred=predict(multi.fit.aic,newdata=data.2)
```

```
write.csv(step.multi.pred,"Forward Selection_label.csv", row.names = FALSE)
```

Penalized Multinomial Regression(Cross Validation)

```
multi.fit.2=train(Label ~ .
                  ,method = "multinom"
                  ,trControl=trControl
                  ,metric = "Accuracy"
                  , trace = F
                  ,data = data)
multi.fit.2
```

```
## Penalized Multinomial Regression
##
## 1500 samples
##   52 predictor
##    6 classes: '0', '1', '2', '3', '4', '5'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1200, 1200, 1202, 1201, 1197
## Resampling results across tuning parameters:
##
##   decay  Accuracy   Kappa
##   0e+00  0.8920101  0.8644047
##   1e-04  0.8913434  0.8636007
##   1e-01  0.9033835  0.8780765
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was decay = 0.1.
```

```
confusionMatrix(multi.fit.2,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 373  10   9  28   2   1
##          1   6 269   0  11   6   5
##          2   3   0  48   1   1   0
##          3  13  12   1 176  12   1
##          4   0   4   0  11 125   0
##          5   0   8   0   0   0 364
##
##  Accuracy (average) : 0.9033
```

```
multi.fit.2.pred=predict(multi.fit.2,newdata=data.2)
```

```
write.csv(multi.fit.2.pred,"Penalized Multinomial Regression_label.csv", row.names = FALSE)
```

SVM

```r
svm.fit <- train(Label~.,method= "svmRadial",
                 trControl = trControl,
                 metric= "Accuracy",
                 data= data)
pred.svm=predict(svm.fit,newdata=data.2)
#pred.svm
confusionMatrix(svm.fit,norm="none")
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are un-normalized aggregated counts)
##
##           Reference
## Prediction   0   1   2   3   4   5
##          0 374   8  22  39   3   0
##          1   9 270   0  14   2   0
##          2   4   0  36   0   0   0
##          3   8   6   0 172   4   4
##          4   0   1   0   2 137   0
##          5   0  18   0   0   0 367
##
##  Accuracy (average) : 0.904
```

```r
write.csv(pred.svm,"SVM-radial_Label.csv", row.names = FALSE)
```

mode

```
train_pred_response <- cbind(
  matrix(predict(qda.fit, data = pca_data_train), ncol=1),
  matrix(predict(lda.fit, data = pca_data_train), ncol=1),
  matrix(predict(knn.fit, data = data), ncol=1),
  matrix(predict(rf.fit, data = data), ncol=1),
  matrix(predict(boosttree.fit, data = data), ncol=1),
  matrix(predict(naive.fit, data = data), ncol=1),
  matrix(predict(cv.out,s=bestlam,type = "class",newx =x), ncol=1),
  matrix(predict(multi.fit.aic, data = data), ncol=1),
  matrix(rep(NA,1500), ncol=1), # !!!
  #matrix(predict(multi.fit.2, data = data), ncol=1),
  matrix(predict(svm.fit, data = data), ncol=1))

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

train_pred.mode <- apply(train_pred_response, 1, Mode)
# Confusion Matrix and Accuracy
table(train_pred.mode,data[,53]) ; mean(train_pred.mode==data[,53])
```

```
##
## train_pred.mode   0   1   2   3   4   5
##               0 382   2   1  22   2   0
##               1   5 290   0  11   0   0
##               2   2   0  57   0   0   0
##               3   6   6   0 194   5   0
##               4   0   0   0   0 139   0
##               5   0   5   0   0   0 371
```

```
## [1] 0.9553333
```

```
test_pred_response <- cbind(matrix(pred.qda, ncol=1),
                            matrix(pred.lda, ncol=1),
                            matrix(pred.knn, ncol=1),
                            matrix(pred.rf, ncol=1),
                            matrix(pred.boosttree, ncol=1),
                            matrix(pred.naive, ncol=1),
                            lasso.pred,
                            matrix(step.multi.pred, ncol=1),
                            matrix(multi.fit.2.pred, ncol=1),
                            matrix(pred.svm, ncol=1))

pred.mode <- apply(test_pred_response, 1, Mode)
#pred.mode
test_pred_response <- cbind(test_pred_response, pred.mode)
```

all model

```r
colnames(test_pred_response)<- paste(c("QDA_Label", "LDA_Label", "KNN_Label", "Random forest_Label", "Bo
head(test_pred_response)
```

```
##      QDA_Label_2x2 LDA_Label_2x2 KNN_Label_2x2 Random forest_Label_2x2
## [1,] "5"           "5"           "0"           "5"
## [2,] "0"           "0"           "0"           "0"
## [3,] "0"           "0"           "0"           "3"
## [4,] "0"           "0"           "0"           "0"
## [5,] "0"           "0"           "0"           "0"
## [6,] "1"           "1"           "5"           "1"
##      Boosting Tree_Label_2x2 Naive Bayes_Label_2x2 LASSO_Label_2x2
## [1,] "5"                     "5"                   "5"
## [2,] "0"                     "2"                   "0"
## [3,] "3"                     "0"                   "0"
## [4,] "0"                     "1"                   "0"
## [5,] "0"                     "0"                   "0"
## [6,] "1"                     "1"                   "1"
##      Forward Selection_Label_2x2 Penalized Multinomial Regression_Label_2x2
## [1,] "5"                         "5"
## [2,] "0"                         "0"
## [3,] "3"                         "0"
## [4,] "0"                         "0"
## [5,] "0"                         "0"
## [6,] "1"                         "1"
##      SVM-radial_Label_2x2 Mode_Label_2x2
## [1,] "5"                  "5"
## [2,] "0"                  "0"
## [3,] "0"                  "0"
## [4,] "0"                  "0"
## [5,] "0"                  "0"
## [6,] "1"                  "1"
```

```r
write.csv(test_pred_response,"2x2_label.csv", row.names = FALSE)
```