Lab Report 2

By: Abhishek Shah, Ravi Seth, Tanya Ralliaram

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
# Insert necessary packages
library('glmnet')
library('caret')
library('ISLR')
```

Question 1: Nonlinear Regression

Question 2: Text Classification

```
# read in data
health <- read.csv("mental_health.csv")[,-1]
```

2.1 Train / Test Split

```
set.seed(123)

train_inds <- sample(1:nrow(health), floor(nrow(health)*0.8))
train <- health[ train_inds, ]
test  <- health[-train_inds, ]

X_train   <- model.matrix(IsMentalHealthRelated ~ .,train)
y_train   <- train$IsMentalHealthRelated
X_test   <- model.matrix(IsMentalHealthRelated ~ .,test)
y_test   <- test$IsMentalHealthRelated

cat('train: ', dim(train), ', test: ', dim(test))
```

```
## train:  5049 489 , test:  1263 489
```

2.2 Fit models

```
# Logistic Regression model
fit.logreg <- glm(formula = IsMentalHealthRelated ~ ., data=train, family = binomial())

# L1 Model
cv.fit  <- cv.glmnet(X_train, y_train, alpha=1, family="binomial", nfolds = 5)
lambda.l1 <- cv.fit$lambda.min
fit.l1 <- glmnet(X_train, y_train, alpha=1, family="binomial", lambda=lambda.l1)

# L2 Model
cv.fit  <- cv.glmnet(X_train, y_train, alpha=0, family="binomial", nfolds = 5)
lambda.l2 = cv.fit$lambda.min
fit.l2 <- glmnet(X_train, y_train, alpha=0, family="binomial", lambda=lambda.l2)
```

2.3 Compare Performances

```
# Logistic Regression (LR)
probs.logreg <- predict(fit.logreg, as.data.frame(X_test), type="response")
preds.logreg  <- ifelse(probs.logreg >= 0.5, 1, 0)
acc.logreg <- mean(preds.logreg == y_test)
```

```r
# L1 Model
probs.l1 <- predict(fit.l1, X_test, type="response")
preds.l1  <- ifelse(probs.l1 >= 0.5, 1, 0)
acc.l1 <- mean(preds.l1 == y_test)

# L2 Model
probs.l2 <- predict(fit.l2, X_test, type="response")
preds.l2  <- ifelse(probs.l2 >= 0.5, 1, 0)
acc.l2 <- mean(preds.l2 == y_test)

cat(sprintf("Logisitc Regression Accuracy: %f \nL1 Accuracy: %f \nL2 Accuracy: %f", acc.logreg, ac
```

```
## Logisitc Regression Accuracy: 0.855107
## L1 Accuracy: 0.866983
## L2 Accuracy: 0.869359
```

The L2 model had the best accuracy and L1 has the second best accuracy. Logistic Regression without any regularization had the worst accuracy out of the three.

2.4 Interpret the models

```r
sorted.l1 <- sort(coef(fit.l1)[,1])
cat('The words that have the highest coefficients with L1 are: \n')
```

```
## The words that have the highest coefficients with L1 are:
```

```r
sort(tail(sorted.l1, 5), decreasing=TRUE)
```

```
##          term       counsel mental.health            op    university
##      7.943078      6.930450      4.722108      4.580570      3.966913
```

```r
cat('\nThe words that have the smallest coefficients with L1 are: \n')
```

```
##
## The words that have the smallest coefficients with L1 are:
```

```r
head(sorted.l1, 5)
```

```
##    fitness    workout     muscle      squat   workouts
## -11.431782 -10.222683  -9.380444  -7.980440  -7.517964
```

```r
sorted.l2 <- sort(coef(fit.l2)[,1])
cat('\nThe words that have the highest coefficients with L2 are: \n')
```

```
##
## The words that have the highest coefficients with L2 are:
```

```r
sort(tail(sorted.l2, 5), decreasing=TRUE)
```

```
##       term    counsel university         op    service
##   4.485622   3.948127   3.516435   2.921839   2.837861
```

```r
cat('\nThe words that have the smallest coefficients with L2 are: \n')
```

```
##
## The words that have the smallest coefficients with L2 are:
```

```r
head(sorted.l2, 5)
```

```
##   fitness   workout time.week     sugar  workouts
## -6.076931 -5.249456 -5.140047 -4.867361 -4.832190
```

L1 tends to tends to zero many coefficients while keeping the rest as they are. L2 tends to shrink all the coefficients and doesn't zero any.

Question 3: Subset Selection