Choosing and Evaluating Models

Mapping business problems to machine learning tasks.

Evaluating model quality

Explaining model predictions

Mapping business problems to machine learning tasks.

```
**** Classification - Assigning labels to datums
**** Scoring - Assigning numerical values to datums
**** Grouping - Discovering patterns and commonalities in data
```

```r
#House-keeping
rm(list = ls(all=TRUE))
#dev.off()
cat('\014')
```

```r
#set working directory
path <- 'C:/Users/mwm9/Desktop/100DaysOfCode/100DaysCodeChallenge/Day 5'
getwd()
```

```
## [1] "C:/Users/mwm9/Desktop/100DaysOfCode/100DaysCodeChallenge/Day 5"
```

```r
setwd(path)
list.files()
```

```
## [1] "Choosing_Evaluating_Models.knit.md" "Choosing_Evaluating_Models.nb.html"
## [3] "Choosing_Evaluating_Models.Rmd"     "Choosing_Evaluating_Models.utf8.md"
## [5] "pandocc086b3253c5.html"             "spamD.tsv"
```

```r
#install and load libraries
#install.packages(c('ggplot2'))

#library(ggplot2)
```

Data to be used for Chapter 5 and 6

```r
spamD <- read.table('http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.data', s

spamCols <- c(
  'word.freq.make', 'word.freq.address', 'word.freq.all',
   'word.freq.3d', 'word.freq.our', 'word.freq.over', 'word.freq.remove',
   'word.freq.internet', 'word.freq.order', 'word.freq.mail',
   'word.freq.receive', 'word.freq.will', 'word.freq.people',
   'word.freq.report', 'word.freq.addresses', 'word.freq.free',
   'word.freq.business', 'word.freq.email', 'word.freq.you',
   'word.freq.credit', 'word.freq.your', 'word.freq.font',
   'word.freq.000', 'word.freq.money', 'word.freq.hp', 'word.freq.hpl',
   'word.freq.george', 'word.freq.650', 'word.freq.lab',
   'word.freq.labs', 'word.freq.telnet', 'word.freq.857',
   'word.freq.data', 'word.freq.415', 'word.freq.85',
   'word.freq.technology', 'word.freq.1999', 'word.freq.parts',
   'word.freq.pm', 'word.freq.direct', 'word.freq.cs',
   'word.freq.meeting', 'word.freq.original', 'word.freq.project',
   'word.freq.re', 'word.freq.edu', 'word.freq.table',
   'word.freq.conference', 'char.freq.semi', 'char.freq.lparen',
   'char.freq.lbrack', 'char.freq.bang', 'char.freq.dollar',
   'char.freq.hash', 'capital.run.length.average',
   'capital.run.length.longest', 'capital.run.length.total',
   'spam'
)

colnames(spamD) <- spamCols
spamD$spam <- as.factor(ifelse(spamD$spam>0.5, 'spam', 'non-spam'))
set.seed(18012020)
spamD$rgroup <- floor(100*runif(dim(spamD)[[1]]))
write.table(spamD, file='spamD.tsv',quote = F, sep = '\t', row.names = F)
```

Classification problems - Multicategory | Two-category Classification

```r
#read data into R
spamD <- read.table('spamD.tsv', header= TRUE, sep = '\t')


#partion data into training and test datasets
spamTrain <- subset(spamD,spamD$rgroup >= 10)
spamTest <- subset(spamD, spamD$rgroup < 10)


#Create a formula that describes the model
spamVars <- setdiff(colnames(spamD), list('rgroup','spam'))

spamFormula <- as.formula(paste('spam == ""',
                                paste(spamVars, collapse = '+'), sep = '~'))


#Fit the logistic regression model
spamModel <- glm(spamFormula, family = binomial(link = 'logit'),
                 data = spamTrain, maxit = 100)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
#Make predictions on the training and test sets
spamTrain$pred <- predict(spamModel, newdata = spamTrain,
                          type = 'response')
spamTest$pred <- predict(spamModel, newdata = spamTest,
                         type = 'response')
```

Title: Spam classifications

```r
sample <- spamTest[c(7,35,224,327), c('spam','pred')]
print(sample)
```

```
##          spam         pred
## 130      spam 2.220446e-16
## 494      spam 2.220446e-16
## 2442 non-spam 2.220446e-16
## 3416 non-spam 2.220446e-16
```

```r
#Spam Confusion Matrix
confmat_spam <- table(truth = spamTest$spam,
                      prediction = ifelse(spamTest$pred > 0.5,
                                          "spam", "non-spam"))
print(confmat_spam)
```

```
##           prediction
## truth      non-spam
##   non-spam      290
##   spam          170
```