Classification

Fabian Peri

October 9, 2018

```r
# packages needed for chapter 5
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(FNN)
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
```

```r
library(rpart)
library(klaR)
```

```r
# Import the datasets needed for chapter 5
PSDS_PATH <- file.path('C:/Users/fabia/Desktop', 'psds_data')

## Import datasets needed for chapter 5
loan3000 <- read.csv(file.path(PSDS_PATH, 'data', 'loan3000.csv'))
loan_data <- read.csv(file.path(PSDS_PATH, 'data', 'loan_data.csv'))
loan_data$outcome <- ordered(loan_data$outcome, levels=c('paid off', 'default'))
full_train_set <- read.csv(file.path(PSDS_PATH, 'data', 'full_train_set.csv'))
full_train_set$outcome <- ordered(full_train_set$outcome, levels=c('paid off', 'default'))
```

```r
## Naive Bayes
naive_model <- NaiveBayes(outcome ~ purpose_ + home_ + emp_len_,
                          data = na.omit(loan_data))
naive_model$table
```

```
## $purpose_
##           var
## grouping   credit_card debt_consolidation home_improvement major_purchase
##   paid off  0.18759649         0.55215915       0.07150104     0.05359270
##   default   0.15151515         0.57571347       0.05981209     0.03727229
##           var
## grouping      medical      other small_business
##   paid off 0.01424728 0.09990737     0.02099599
##   default  0.01433549 0.11561025     0.04574126
##
## $home_
##           var
## grouping    MORTGAGE       OWN      RENT
##   paid off 0.4894800 0.0808963 0.4296237
##   default  0.4313440 0.0832782 0.4853778
##
## $emp_len_
##           var
## grouping     < 1 Year   > 1 Year
##   paid off 0.03105289 0.96894711
##   default  0.04728508 0.95271492
```

```r
new_loan <- loan_data[147, c('purpose_', 'home_', 'emp_len_')]
row.names(new_loan) <- NULL
new_loan
```

```
##         purpose_    home_  emp_len_
## 1 small_business MORTGAGE  > 1 Year
```

```r
predict(naive_model, new_loan)
```

```
## $class
## [1] default
## Levels: paid off default
##
## $posterior
##       paid off   default
## [1,] 0.3463013 0.6536987
```

```r
## example not in book
less_naive <- NaiveBayes(outcome ~ borrower_score + payment_inc_ratio +
                           purpose_ + home_ + emp_len_, data = loan_data)
less_naive$table[1:2]
```

```
## $borrower_score
##               [,1]      [,2]
## paid off 0.5347933 0.1238649
## default  0.4632195 0.1233597
##
## $payment_inc_ratio
##              [,1]     [,2]
## paid off 7.294367 4.018183
## default  8.770084 4.373793
```

```r
png(filename=file.path(PSDS_PATH, 'figures', 'psds_naive_bayes.png'),  width = 4, height=3, units='in',

stats <- less_naive$table[[1]]
```

```
ggplot(data.frame(borrower_score=c(0,1)), aes(borrower_score)) +
  stat_function(fun = dnorm, color='blue', linetype=1,
                arg=list(mean=stats[1, 1], sd=stats[1, 2])) +
  stat_function(fun = dnorm, color='red', linetype=2,
                arg=list(mean=stats[2, 1], sd=stats[2, 2])) +
  labs(y='probability')
```

```
## Warning: Ignoring unknown parameters: arg

## Warning: Ignoring unknown parameters: arg
```

```
dev.off()
```

```
## pdf
##   2
```

```
#
```

```
## Code for LDA
loan_lda <- lda(outcome ~ borrower_score + payment_inc_ratio,
                data=loan3000)
loan_lda$scaling
```

```
##                           LD1
## borrower_score     7.17583880
## payment_inc_ratio -0.09967559
```

```
## Code snippet 4.2
pred <- predict(loan_lda)
head(pred$posterior)
```

```
##     default  paid off
## 1 0.5535437 0.4464563
## 2 0.5589534 0.4410466
## 3 0.2726962 0.7273038
## 4 0.5062538 0.4937462
## 5 0.6099525 0.3900475
## 6 0.4107406 0.5892594
```

```
## LDA
## Code for Figure 5-1
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0501.png'),  width = 4, height=3, units='in', res=30

pred <- predict(loan_lda)
lda_df <- cbind(loan3000, prob_default=pred$posterior[,'default'])

x <- seq(from=.33, to=.73, length=100)
y <- seq(from=0, to=20, length=100)
newdata <- data.frame(borrower_score=x, payment_inc_ratio=y)
pred <- predict(loan_lda, newdata=newdata)
lda_df0 <- cbind(newdata, outcome=pred$class)

ggplot(data=lda_df, aes(x=borrower_score, y=payment_inc_ratio, color=prob_default)) +
  geom_point(alpha=.6) +
  scale_color_gradient2(low='white', high='blue') +
  scale_x_continuous(expand=c(0,0)) +
  scale_y_continuous(expand=c(0,0), lim=c(0, 20)) +
```

```r
  geom_line(data=lda_df0, col='green', size=2, alpha=.8) +
  theme_bw()
```

```
## Warning: Removed 18 rows containing missing values (geom_point).
```

```r
dev.off()
```

```
## pdf
##   2
```

```r
## Logistic regression
logistic_model <- glm(outcome ~ payment_inc_ratio + purpose_ +
                        home_ + emp_len_ + borrower_score,
                      data=loan_data, family='binomial')
logistic_model
```

```
##
## Call:  glm(formula = outcome ~ payment_inc_ratio + purpose_ + home_ +
##     emp_len_ + borrower_score, family = "binomial", data = loan_data)
##
## Coefficients:
##                (Intercept)           payment_inc_ratio
##                    1.63809                     0.07974
## purpose_debt_consolidation    purpose_home_improvement
##                    0.24937                     0.40774
##     purpose_major_purchase             purpose_medical
##                    0.22963                     0.51048
##              purpose_other      purpose_small_business
##                    0.62066                     1.21526
##                   home_OWN                   home_RENT
##                    0.04833                     0.15732
##         emp_len_ > 1 Year              borrower_score
##                   -0.35673                    -4.61264
##
## Degrees of Freedom: 45341 Total (i.e. Null);  45330 Residual
## Null Deviance:      62860
## Residual Deviance: 57510     AIC: 57540
```

```r
summary(logistic_model)
```

```
##
## Call:
## glm(formula = outcome ~ payment_inc_ratio + purpose_ + home_ +
##     emp_len_ + borrower_score, family = "binomial", data = loan_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.51951  -1.06908  -0.05853   1.07421   2.15528
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.638092   0.073708  22.224  < 2e-16 ***
## payment_inc_ratio           0.079737   0.002487  32.058  < 2e-16 ***
## purpose_debt_consolidation  0.249373   0.027615   9.030  < 2e-16 ***
## purpose_home_improvement    0.407743   0.046615   8.747  < 2e-16 ***
## purpose_major_purchase      0.229628   0.053683   4.277 1.89e-05 ***
```

```
## purpose_medical             0.510479   0.086780   5.882 4.04e-09 ***
## purpose_other               0.620663   0.039436  15.738  < 2e-16 ***
## purpose_small_business      1.215261   0.063320  19.192  < 2e-16 ***
## home_OWN                    0.048330   0.038036   1.271    0.204
## home_RENT                   0.157320   0.021203   7.420 1.17e-13 ***
## emp_len_ > 1 Year          -0.356731   0.052622  -6.779 1.21e-11 ***
## borrower_score             -4.612638   0.083558 -55.203  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 62857  on 45341  degrees of freedom
## Residual deviance: 57515  on 45330  degrees of freedom
## AIC: 57539
##
## Number of Fisher Scoring iterations: 4
```

```
p <- seq(from=0.01, to=.99, by=.01)
df <- data.frame(p = p ,
                 logit = log(p/(1-p)),
                 odds = p/(1-p))
## Figure 5-2

png(filename=file.path(PSDS_PATH, 'figures', 'psds_0502.png'),  width = 5, height=4, units='in', res=30
ggplot(data=df, aes(x=p, y=logit)) +
  geom_line() +
  labs(x = 'p', y='logit(p)') +
  theme_bw()
dev.off()
```

```
## pdf
##   2
```

```
## Figure 5-3
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0503.png'),  width = 5, height=4, units='in', res=30
ggplot(data=df, aes(x=logit, y=odds)) +
  geom_line() +
  labs(x = 'log(odds ratio)', y='odds ratio') +
  ylim(1, 100) +
  xlim(0, 5) +
  theme_bw()
```

```
## Warning: Removed 49 rows containing missing values (geom_path).
```

```
dev.off()
```

```
## pdf
##   2
```

```
pred <- predict(logistic_model)
summary(pred)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.704774 -0.518825 -0.008539  0.002564  0.505061  3.509606
```

```
prob <- 1/(1 + exp(-pred))
summary(prob)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06269 0.37313 0.49787 0.50000 0.62365 0.97096
```

```
#

logistic_gam <- gam(outcome ~ s(payment_inc_ratio) + purpose_ +
                      home_ + emp_len_ + s(borrower_score),
                    data=loan_data, family='binomial')
logistic_gam
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## outcome ~ s(payment_inc_ratio) + purpose_ + home_ + emp_len_ +
##     s(borrower_score)
##
## Estimated degrees of freedom:
## 7.45 4.17  total = 21.61
##
## UBRE score: 0.2681413
```

```
terms <- predict(logistic_gam, type='terms')
partial_resid <- resid(logistic_gam) + terms
df <- data.frame(payment_inc_ratio = loan_data[, 'payment_inc_ratio'],
                 terms = terms[, 's(payment_inc_ratio)'],
                 partial_resid = partial_resid[, 's(payment_inc_ratio)'])


## Code for Figure 5-4
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0504.png'),  width = 5, height=4, units='in', res=30

ggplot(df, aes(x=payment_inc_ratio, y=partial_resid, solid = FALSE)) +
  geom_point(shape=46, alpha=.4) +
  geom_line(aes(x=payment_inc_ratio, y=terms),
            color='red', alpha=.5, size=1.5) +
  labs(y='Partial Residual') +
  xlim(0, 25) +
  theme_bw()
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```

```
## Warning: Removed 9 rows containing missing values (geom_path).
```

```
dev.off()
```

```
## pdf
##   2
```

```
# Confusion matrix
pred <- predict(logistic_gam, newdata=loan_data)
pred_y <- as.numeric(pred > 0)
true_y <- as.numeric(loan_data$outcome=='default')
true_pos <- (true_y==1) & (pred_y==1)
true_neg <- (true_y==0) & (pred_y==0)
```

```
false_pos <- (true_y==0) & (pred_y==1)
false_neg <- (true_y==1) & (pred_y==0)
conf_mat <- matrix(c(sum(true_pos), sum(false_pos),
                     sum(false_neg), sum(true_neg)), 2, 2)
colnames(conf_mat) <- c('Yhat = 1', 'Yhat = 0')
rownames(conf_mat) <- c('Y = 1', 'Y = 0')
conf_mat
```

```
##       Yhat = 1 Yhat = 0
## Y = 1    14295     8376
## Y = 0     8052    14619
```

```
# precision
conf_mat[1,1]/sum(conf_mat[,1])
```

```
## [1] 0.6396832
```

```
# recall
conf_mat[1,1]/sum(conf_mat[1,])
```

```
## [1] 0.6305412
```

```
# specificity
conf_mat[2,2]/sum(conf_mat[2,])
```

```
## [1] 0.6448326
```

```
## Code for Figure 5-6
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0506.png'),  width = 4, height=4, units='in', res=30

idx <- order(-pred)
recall <- cumsum(true_y[idx]==1)/sum(true_y==1)
specificity <- (sum(true_y==0) - cumsum(true_y[idx]==0))/sum(true_y==0)
roc_df <- data.frame(recall = recall, specificity = specificity)
ggplot(roc_df, aes(x=specificity, y=recall)) +
  geom_line(color='blue') +
  scale_x_reverse(expand=c(0, 0)) +
  scale_y_continuous(expand=c(0, 0)) +
  geom_line(data=data.frame(x=(0:100)/100), aes(x=x, y=1-x),
            linetype='dotted', color='red') +
  theme_bw()

dev.off()
```

```
## pdf
##   2
```

```
## Code for Figure 5-7
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0507.png'),  width = 4, height=4, units='in', res=30

ggplot(roc_df, aes(specificity)) +
  geom_ribbon(aes(ymin=0, ymax=recall), fill='blue', alpha=.3) +
  scale_x_reverse(expand=c(0, 0)) +
  scale_y_continuous(expand=c(0, 0)) +
  labs(y='recall') +
  theme_bw()

dev.off()
```

```
## pdf
##   2
```

```
## AUC calculation
sum(roc_df$recall[-1] * diff(1-roc_df$specificity))
```

```
## [1] 0.6926172
```

```
head(roc_df)
```

```
##         recall specificity
## 1 4.410921e-05   1.0000000
## 2 8.821843e-05   1.0000000
## 3 8.821843e-05   0.9999559
## 4 1.323276e-04   0.9999559
## 5 1.764369e-04   0.9999559
## 6 2.205461e-04   0.9999559
```

```
## Code for Undersampling
mean(full_train_set$outcome=='default')
```

```
## [1] 0.1889455
```

```
full_model <- glm(outcome ~ payment_inc_ratio + purpose_ +
                    home_ + emp_len_+ dti + revol_bal + revol_util,
                  data=full_train_set, family='binomial')
pred <- predict(full_model)
mean(pred > 0)
```

```
## [1] 0.003942094
```

```
## Code for oversampling/up weighting
wt <- ifelse(full_train_set$outcome=='default', 1/mean(full_train_set$outcome == 'default'), 1)
full_model <- glm(outcome ~ payment_inc_ratio + purpose_ +
                    home_ + emp_len_+ dti + revol_bal + revol_util,
                  data=full_train_set, weight=wt, family='quasibinomial')
pred <- predict(full_model)
mean(pred > 0)
```

```
## [1] 0.5767208
```

```
# Code for Figure 5-8: comparison of methods
loan_tree <- rpart(outcome ~ borrower_score + payment_inc_ratio,
                   data=loan3000,
                   control = rpart.control(cp=.005))

lda_pred <- lda_df0[, c('borrower_score', 'payment_inc_ratio')]
lda_pred$method = 'LDA'

tree_pred <- data.frame(borrower_score = c(0.375, 0.375, 0.525, 0.525, 0.625, 0.625),
                        payment_inc_ratio = c(0, 9.732,  9.732, 8.772, 8.772, 20),
                        method = rep('Tree', 6))

glm0 <- glm(outcome ~ (payment_inc_ratio) +  (borrower_score),
            data=loan3000, family='binomial')
y <- seq(from=0, to=20, length=100)
x <- (-glm0$coefficients[1] - glm0$coefficients[2]*y)/glm0$coefficients[3]
glm0_pred <- data.frame(borrower_score=x, payment_inc_ratio=y, method='Logistic')
```

```
gam1 <- gam(outcome ~ s(payment_inc_ratio) +  s(borrower_score),
            data=loan3000, family='binomial')
# newdata = gam0_pred

gam_fun <- function(x){
  rss <- sum(predict(gam1, newdata=data.frame(borrower_score=x, payment_inc_ratio=y))^2)
}
est_x <- nlminb(newdata$borrower_score, gam_fun )
gam1_pred <- data.frame(borrower_score=est_x$par, payment_inc_ratio=y, method="GAM")

loan_fits <- rbind(lda_pred,
                   tree_pred,
                   glm0_pred,
                   gam1_pred)


## Code for Figure 5-8
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0508.png'),  width = 6, height=4, units='in', res=300
ggplot(data=loan_fits, aes(x=borrower_score, y=payment_inc_ratio, color=method, linetype=method)) +
  geom_line(size=1.5) +
  theme(legend.key.width = unit(2,"cm")) +
  guides(linetype = guide_legend(override.aes = list(size = 1)))
dev.off()
```

```
## pdf
##   2
```