Classification: LDA and QDA Approaches

Shane T. Mueller shanem@mtu.edu

2019-02-28

Classification and Categorization

General regression approaches we have taken so far have typically had the goal of modeling how a dependent variable (usually continuous, but in the case of logistic regression, binary, or with multinomial regression multiple levels) is predicted by a set of independent or predictor variables. The goals of this are often prediction, inference, forecasting, and understanding the underlying processes.

Many times, we might care less about making a numerical prediction, and more about classifying into categories to be able to take action, based on measureable properties. This is very similar to the goals of logistic regression, but probably with a preference for assessing the most likely categorization, rather than the probability of a category. Furthermore, logistic regression will typically ignore the prior likelihoods of the different outcomes. Classification methods would typically want to consider the overall base rate as well as the relative influence of different factors, in order to maximize the probability of getting the right prediction. Typically, for classification approaches, we need to select a criterion or cut-off of some sort to make a decision—hopefully one that will give the greatest chance of getting a new classification right.

There are many applications for classification, and recently the work on this has migrated from statisticians to computer scientists and specialists in machine classification, machine learning, and artificial intelligence.

Historically, there are many uses for classification, and the goals differ somewhat from approaches like logistic regression and MANOVA. For example, either approach could be used as a tool to help diagnose a disease. A logistic regression would produce the odds that a person has a disease, whereas the a machine classification might be used to determine whether a disease is present—perhaps with some margin of error. A regression might be used to determine the likelihood a company’s will declare bankruptcy, while a classification might be used to identify a set of ‘at-risk’ companies. Overall, these are very similar goals, and although there are many different approaches to achieve this, the solutions end up looking very similar.

Some of the terminology tends to differ across these two approaches. A regression approach might discuss variables, fitting, and inference; a classification might call these features, learning, and cross-validation. There are some aspects of applying regression models and classification models that have similar goals, but different methods. For example, in regression modeling, we often use anova, analysis of deviance, or a criterion such as AIC to compare models and determine whether a predictor should be used. In classification, the ultimate criterion is often classification accuracy, and a huge aspect of ‘machine learning’ is selecting and removing features to use in a model. In general, the attitude for regression models is typically interpretability, leading to small models with comprehensible measures. Oftentimes, classification approaches might start with hundreds or thousands of features, and combination rules and decision rules might be very abstract and difficult to comprehend. Furthermore, classification is more likely to use separate data sets or cross-validation to do variable selection (feature learning). Regression is more likely to use AIC or BIC methods to select variables on the whole data set.

Classification as Logistic Regression

Logistic regression provides a method for making a prediction about the binary classification of something based on a set of predictors.

In the following data set, we asked participants to judge whether concepts were related. Some of the concepts were engineering terms; others were psychology terms. We also polled both engineering and psychology
students, and we wanted to know if we could tell them apart based on the speed and accuracy of their responses. Terms were classified into different bins, with e=engineering, p=psychology, r=related, and u=unrelated.

```r
library(dplyr)

data.raw <- read.csv("samediff-pooled.csv")
data <- dplyr::filter(data.raw, cond %in% c("eerc", "eer", "eeu", "ep", "ppr", "ppu"))

stim <- c(as.character(data$word1), as.character(data$word2))
acc <- c(data$corr, data$corr)
rt <- c(data$rt, data$rt)
sub <- c(data$subnum, data$subnum)
data$pairs <- paste(data$word1, data$word2, sep = "-")

dat.corr <- as.data.frame(tapply(data$corr, list(sub = data$sub, type = factor(data$cond)), mean))
dat.rt <- as.data.frame(tapply(data$rt, list(sub = data$sub, type = factor(data$cond)), function(x) { exp(mean(log(x), na.rm = T)) }))

dat.joint <- cbind(dat.corr, dat.rt)

survey <- read.csv("survey.csv")
surv2 <- data.frame(sub = survey$subnum, eng = survey/engineering)

joint <- data.frame(eng = surv2$eng, dat.joint)

modell <- glm(eng ~ ., data = joint, family = binomial)

summary(modell)
```

**Call:**

```
glm(formula = eng ~ ., family = binomial, data = joint)
```

**Deviance Residuals:**

```
          Min        1Q  Median        3Q        Max
-1.65694  -1.07432  -0.08598   1.09348   1.76901
```

**Coefficients:**

```
                Estimate Std. Error   z value  Pr(>|z|)
(Intercept)  -3.322e+00  2.018e+00  -1.646   0.0997 .
eer        2.279e-02  9.498e-01   0.024   0.9809
eeu       6.761e-01  8.583e-01   0.788   0.4309
ep        9.017e-01  1.916e+00   0.471   0.6380
ppr      3.123e-01  9.844e-01   0.317   0.7510
ppu     1.564e+00  1.043e+00   1.499   0.1337
eer.1  -3.658e-05  2.387e-04  -0.153   0.8782
eeu.1  -1.627e-04  2.979e-04  -0.546   0.5849
ep.1   3.758e-04  3.804e-04   0.988   0.3232
ppr.1  -1.092e-04  2.485e-04  -0.439   0.6605
```

2
ppu.1   2.380e-04  2.761e-04   0.862   0.3886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 105.36  on 75  degrees of freedom
Residual deviance: 97.54  on 65  degrees of freedom
AIC: 119.54

Number of Fisher Scoring iterations: 4

We can make a prediction about the log-odds of each person being an engineer, and compare it to the actual values

\[
\text{table}(\text{predict}(\text{model1, joint}) > 0, \text{joint}$\text{eng}$)
\]

\[
\begin{array}{cc}
0 & 1 \\
FALSE & 26 14 \\
TRUE & 12 24 \\
\end{array}
\]

This is not bad—about 2/3 correct in each category. Maybe there is a better place to put the criterion. This could be true if we had mostly engineers or mostly psychologists.

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint}) > -0.5, \text{joint}$\text{eng}$)))
\]

[1] 48

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint}) > -0.25, \text{joint}$\text{eng}$)))
\]

[1] 50

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint}) > -0.1, \text{joint}$\text{eng}$)))
\]

[1] 50

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint}) > 0, \text{joint}$\text{eng}$)))
\]

[1] 50

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint}) > 0.25, \text{joint}$\text{eng}$)))
\]

[1] 47

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint}) > 0.5, \text{joint}$\text{eng}$)))
\]

[1] 40

If we looked at just the first 30 participants, the best criterion is different:

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint[1:50, ]}) > -0.5, \text{joint}$\text{eng[1:50]}$)))
\]

[1] 35

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint[1:50, ]}) > -0.25, \text{joint}$\text{eng[1:50]}$)))
\]

[1] 36

\[
\text{sum} (\text{diag} (\text{table}(\text{predict}(\text{model1, joint[1:50, ]}) > -0.1, \text{joint}$\text{eng[1:50]}$)))
\]

[1] 31
Here, a criterion of -.25 gives us better classification performance. This is just a consequence of base rate, because more engineers were stored in the first half of the data file.

This is OK, but a bit worrisome. None of the predictors were significant, so we might have gotten here just by chance. We might be over-fitting the data. In regression, we’d do inferential test or information criteria to identify which variables to use. Later, we’ll see how this is done for machine classification.

**Basics of Machine Classification**

We can see how logistic regression can be used as a classifier, as long as we can determine a resonable criterion. Prior to the widespread use of logistic regression (which requires maximum-likelihood estimation and computerized approaches), various approaches to classification were developed that use simplified models. One traditional model assumes that two groups have multivariate normal distributions with equal variance. In the figure below, we see two such 2-dimensional distributions, with a line drawn between the centers of each, and contours showing the basic data.

```r
library(ggplot2)
n <- 500
class <- rep(c("A", "B"), each = n)
x <- rnorm(n * 2, mean = rep(c(3, 4), each = n))
y <- rnorm(n * 2, mean = rep(c(2.5, 1.5), each = n))
ggplot(data.frame(class, y, x), aes(x = x, y = y, colour = class)) + geom_point(size = 1.5) +
  geom_density_2d() + geom_segment(x = -1, y = 6.5, xend = 7, yend = -1.5,
  col = "grey60", lwd = 1.2) + geom_segment(x = 3, y = 2.5, xend = 4, yend = 1.5,
  col = "black", lwd = 1.2) + geom_segment(x = 1.5, y = 0, xend = 5.5, yend = 4,
  col = "darkgreen") + coord_fixed(ratio = 1, xlim = c(0, 6), ylim = c(0, 6))
```
Under our assumptions, if we draw a line between the centers and extend it out (the grey line), we can project any points onto this line (i.e., the closest point on the line to each point). We could use position along this line to predict category membership in a regression or logistic regression—and this is essentially what regression does. This mapping from input variables to a single function is called the discriminant function, and is equivalent to the weighted sum in regression. Now, if we want to classify any observation, we just need to determine which case is more likely. Given the assumptions of equal variance and normality, this can be shown to be a single criterion that maximizes our chances of being correct. In this case, that corresponds to where the green line intersects the block line, and if we move back to the original data, the entire green line is a good rule discriminating the two groups.

Under these assumptions, there are a number of approaches that can be used. In fact, MANOVA is essentially equivalent, but framing the model backwards. The most common approach used is referred to as linear discriminant analysis (LDA), or sometimes multivariate discrimination analysis (MDA). The assumptions of this approach bit stronger than regression (requiring normal input predictors and equal variances). If these assumptions hold or we can transform them so they work, we can get improved classification results over other methods. In practice, the methods are likely to be almost equivalent to logistic regression.

# Linear discriminant analysis

Using the fake data from the figure, we can fit an lda model, using the MASS library:

```r
library(MASS)
model0 <- lda(as.factor(class) ~ x + y)
model0
```

Call:
`lda(as.factor(class) ~ x + y)`

Prior probabilities of groups:

```
A   B
0.5 0.5
```

Group means:
Coefficients of linear discriminants:
LD1
x 0.6722027
y -0.7126982

We can see that the simple LDA finds the mean of group A and B along the two measured dimensions, and then reports 'coefficients of linear discriminants'. This is the direction in XY space that best discriminates the two groups. If we could map each observation onto this line, we can easily make a decision that optimally classifies the two groups. We can simply multiply each observed variable by the coefficient to get a sum value, which is the value used to discriminate the two classes, once an optimal threshold is chosen. You can see that if you do predict() on a model, the $class tells you predicted class, and $x tells you the exact values we calculated in ldout.

```r
ldout <- x * 0.62 - y * 0.7857
p <- predict(model0)
plot(p$x, ldout, col = p$class)
```

This is a bit clearer if we visualize, which we can do via the klaR library, which has a number of classification schemes available, including a number of visualization methods. Let’s look at a ‘partimat’ plot:

```r
library(klaR)
partimat(as.factor(class) ~ x + y, method = "lda")
```
```
partimat(as.factor(class) ~ x + y, method = "lda", plot.matrix = TRUE, imageplot = FALSE) # takes some
```
If you think about the line connecting the centers of the two groups, it goes from the upper left to lower right. This direction can be defined by a vector: (.62, -.78).

Let’s look at the engineering data set, which has more than two predictors

```r
partimat(as.factor(eng) ~ ., data = joint, method = "lda", plot.matrix = TRUE, imageplot = FALSE)
```

This shows the classification along each pair of dimensions.

```r
library(MASS)
library(DAAG)
ll <- lda(eng ~ ., data = joint)
ll
```

Call:
lda(eng ~ ., data = joint)

Prior probabilities of groups:
0 1
0.5 0.5

Group means:

<table>
<thead>
<tr>
<th></th>
<th>eer</th>
<th>eeu</th>
<th>ep</th>
<th>ppr</th>
<th>ppu</th>
<th>eer.1</th>
<th>eeu.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.6140351</td>
<td>0.6140351</td>
<td>0.6951754</td>
<td>0.5789474</td>
<td>0.6052632</td>
<td>3038.008</td>
<td>2962.742</td>
</tr>
<tr>
<td>1</td>
<td>0.6228070</td>
<td>0.6842105</td>
<td>0.6973684</td>
<td>0.6754386</td>
<td>0.7017544</td>
<td>3387.195</td>
<td>3120.454</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ep.1</th>
<th>ppr.1</th>
<th>ppu.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3005.707</td>
<td>2921.188</td>
<td>3007.688</td>
</tr>
<tr>
<td>1</td>
<td>3428.785</td>
<td>3024.660</td>
<td>3432.817</td>
</tr>
</tbody>
</table>

Coefficients of linear discriminants:

Coefficient of LD1:

- eer: 3.672284e-02
- eeu: 9.989394e-01
- ep: 1.295582e+00
- ppr: 4.695055e-01
- ppu: 2.385932e+00
- eer.1: -5.396259e-05
- eeu.1: -2.316937e-04
- ep.1: 4.827826e-04
- ppr.1: -1.456996e-04
- ppu.1: 3.812870e-04

If we look at group means and the coefficients, we can see that a few of the measures differ substantially between the two groups, but not all. The largest coefficients typically map onto the dimensions with the greatest difference between groups.

### Predicting class from an LDA model

```r
predict(l1)
```

$class

```
[1] 0 1 0 1 1 0 1 0 1 1 0 0 1 0 1 0 1 1 1 0 0 1 1 1 0 1 0 1 0 1 0 0 0 0 1
[36] 1 1 0 1 0 0 1 0 1 0 0 1 1 0 1 0 0 1 1 1 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1
[71] 1 0 0 0 0 0
Levels: 0 1
```

$posterior

```
0 1
11 0.5512784 0.4487216
12 0.4907813 0.5092187
13 0.5540551 0.4459449
22 0.1602546 0.8497454
31 0.1237940 0.8762060
32 0.6055351 0.3944649
33 0.4234531 0.5765469
41 0.6275495 0.3724505
51 0.1367011 0.8632989
52 0.4797138 0.5202862
```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>0.5544694</td>
<td>0.4455306</td>
</tr>
<tr>
<td>62</td>
<td>0.6745057</td>
<td>0.3254943</td>
</tr>
<tr>
<td>63</td>
<td>0.4469692</td>
<td>0.5530308</td>
</tr>
<tr>
<td>64</td>
<td>0.6752190</td>
<td>0.3247810</td>
</tr>
<tr>
<td>71</td>
<td>0.3398060</td>
<td>0.6601940</td>
</tr>
<tr>
<td>81</td>
<td>0.5468117</td>
<td>0.4531883</td>
</tr>
<tr>
<td>101</td>
<td>0.4014884</td>
<td>0.5988116</td>
</tr>
<tr>
<td>102</td>
<td>0.3694719</td>
<td>0.6305281</td>
</tr>
<tr>
<td>104</td>
<td>0.3822059</td>
<td>0.6177941</td>
</tr>
<tr>
<td>201</td>
<td>0.8183543</td>
<td>0.1816457</td>
</tr>
<tr>
<td>202</td>
<td>0.5981471</td>
<td>0.4018529</td>
</tr>
<tr>
<td>203</td>
<td>0.3306295</td>
<td>0.6693705</td>
</tr>
<tr>
<td>301</td>
<td>0.4570754</td>
<td>0.5429246</td>
</tr>
<tr>
<td>302</td>
<td>0.4452991</td>
<td>0.5547009</td>
</tr>
<tr>
<td>303</td>
<td>0.6392452</td>
<td>0.3607548</td>
</tr>
<tr>
<td>304</td>
<td>0.2763718</td>
<td>0.7236282</td>
</tr>
<tr>
<td>401</td>
<td>0.5196067</td>
<td>0.4803933</td>
</tr>
<tr>
<td>402</td>
<td>0.4912082</td>
<td>0.5087918</td>
</tr>
<tr>
<td>501</td>
<td>0.5067966</td>
<td>0.4932034</td>
</tr>
<tr>
<td>502</td>
<td>0.4105879</td>
<td>0.5894121</td>
</tr>
<tr>
<td>503</td>
<td>0.6803426</td>
<td>0.3196574</td>
</tr>
<tr>
<td>504</td>
<td>0.5564842</td>
<td>0.4435158</td>
</tr>
<tr>
<td>505</td>
<td>0.5814196</td>
<td>0.4185804</td>
</tr>
<tr>
<td>507</td>
<td>0.6919807</td>
<td>0.3080193</td>
</tr>
<tr>
<td>508</td>
<td>0.3626906</td>
<td>0.6373094</td>
</tr>
<tr>
<td>509</td>
<td>0.4488615</td>
<td>0.5511385</td>
</tr>
<tr>
<td>510</td>
<td>0.4027089</td>
<td>0.5972911</td>
</tr>
</tbody>
</table>

[reached getOption("max.print") -- omitted 39 rows]

```r
$x
  
  LD1
11  -0.319286006
12   0.057204773
13  -0.336707250
22   2.687541519
31   3.035585511
32  -0.664798373
33   0.478708503
41  -0.809266716
51   2.858725700
52   0.125937361
61  -0.339308577
62  -1.130227198
63   0.330278399
64  -1.135269580
71   1.030214376
81  -0.291302314
101  0.619325999
102   0.829066967
104   0.744858679
201  -2.334858402
202  -0.616973413
203   1.094091479
301   0.266988525
```
<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>302</td>
<td>0.340762460</td>
<td>303</td>
<td>-0.887400074</td>
<td>304</td>
<td>1.493035179</td>
<td>401</td>
<td>-0.121714764</td>
<td>402</td>
<td>0.054555562</td>
<td>501</td>
<td>-0.042172729</td>
</tr>
<tr>
<td>502</td>
<td>0.560798315</td>
<td>503</td>
<td>-1.171660463</td>
<td>504</td>
<td>-0.351965570</td>
<td>505</td>
<td>-0.509715466</td>
<td>507</td>
<td>-1.255499384</td>
<td>508</td>
<td>0.874394450</td>
</tr>
<tr>
<td>510</td>
<td>0.611451229</td>
<td>602</td>
<td>-1.123001743</td>
<td>603</td>
<td>1.264762328</td>
<td>605</td>
<td>-1.974224117</td>
<td>606</td>
<td>-1.222998498</td>
<td>607</td>
<td>0.924897502</td>
</tr>
<tr>
<td>609</td>
<td>0.548135292</td>
<td>611</td>
<td>-0.063963944</td>
<td>701</td>
<td>-1.050595289</td>
<td>702</td>
<td>0.109848246</td>
<td>703</td>
<td>1.802930456</td>
<td>704</td>
<td>-0.836690773</td>
</tr>
<tr>
<td>706</td>
<td>-1.153642814</td>
<td>707</td>
<td>-0.238165272</td>
<td>708</td>
<td>0.510908846</td>
<td>709</td>
<td>0.283663800</td>
<td>801</td>
<td>0.590030309</td>
<td>802</td>
<td>1.517968113</td>
</tr>
<tr>
<td>804</td>
<td>-0.350171265</td>
<td>805</td>
<td>1.234031870</td>
<td>806</td>
<td>-0.169898973</td>
<td>807</td>
<td>-1.018771624</td>
<td>808</td>
<td>-0.060496899</td>
<td>809</td>
<td>-0.336577470</td>
</tr>
<tr>
<td>903</td>
<td>-0.835666615</td>
<td>905</td>
<td>-1.652252375</td>
<td>906</td>
<td>-0.008369771</td>
<td>907</td>
<td>-0.003339867</td>
<td>908</td>
<td>-1.266912360</td>
<td>909</td>
<td>1.161982264</td>
</tr>
<tr>
<td>911</td>
<td>-1.372632710</td>
<td>913</td>
<td>-0.295670105</td>
<td>914</td>
<td>-0.355635080</td>
<td>916</td>
<td>-1.285565045</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[ reached getOption("max.print") -- omitted 1 row ]
```r
library(GGally)
ggplot(data.frame(Logistic = model1$coefficients[-1], LDA = (ll$scaling)[, 1]),
       aes(x = Logistic, y = LDA)) + geom_point()
```

```r
table(predict(ll)$class, joint$eng)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>24</td>
</tr>
</tbody>
</table>

confusion(joint$eng, predict(ll)$class)

Overall accuracy = 0.671

Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>[,1]</th>
<th>[,2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.711</td>
<td>0.289</td>
</tr>
<tr>
<td>[2,]</td>
<td>0.368</td>
<td>0.632</td>
</tr>
</tbody>
</table>

sum(diag(table(predict(ll)$class, joint$eng)))

[1] 51

Let’s examine the different aspects of the results.

First, the model reports the prior probabilities of groups. This will be the number of each type of input value. Note that if the true base rate differs from the training set (something that might be very likely), we might want to set this explicitly when we predict the data. Next, we see the mean values across each of the variables measured. This is the center of the two normal distributions. Then, we see coefficients of linear discriminants—these are equivalent to the beta weight used to create the discriminant function. It shows the best guess classifications, and finally posterior likelihoods of each—this should be very similar to estimated probabilities from the logistic model. Finally, $x$ shows the discriminant function value, which we could use to choose a different decision criterion. There are several methods for determining the best decision rule.
Here, the coefficients from the logistic and lda models are not identical, but their correlation is 1.0! If we compare the predicted probability vs. the lda likelihood, we see that they are highly correlated.

```r
logit <- function(lo) {
  1/(1 + exp(-lo))
}
##This is the inverse of the logodds function. The
df <- data.frame(logistic = predict(model1), lda = (predict(ll)$x)[, 1], logisticprob = logit(predict(model1)), ldaprob = predict(ll)$posterior[, 2])

ggplot(df, aes(x = logistic, y = lda)) + geom_point()
```
ggplot(df, aes(x = logisticprob, y = ldaprob)) + geom_point()
Notice how the discriminant value is almost the same as the log-odds predicted value in logistic regression, and transforming these to probabilities also produces almost identical values.

In terms of classification performance, the prediction for LDA was 1 error less than logistic regression, but the two models are really essentially identical, with the main difference being how the parameters are fit and the assumptions being made.

## Cross-validation

It is easy to overfit classification data, and so we must be careful to avoid this. Generally, just as with variable selection in regression models, we are worried about determining the best subset of predictors to use. Since using more variables will never make the model worse at fitting its own data, it is useful to hold some data out and test the model on the held-out data. A common approach is to use leave-one-out cross-validation. This approach will fit the model N times for N observations, fitting each left-out case in each model.

The LDA model allows you to do this automatically using the CV=TRUE option. This will do the classification automatically, instead of embedding it within the predict function and doing it manually.

```r
ll2 <- lda(eng ~ ., data = joint, CV = TRUE)
## Original model
confusion(joint$eng, predict(ll)$class)
```

Overall accuracy = 0.671

Confusion matrix

<table>
<thead>
<tr>
<th>Predicted (cv)</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>[1,1] 0.711 0.289</td>
</tr>
<tr>
<td>[2,]</td>
<td>[2,] 0.368 0.632</td>
</tr>
</tbody>
</table>
sum(diag(table(joint$eng, predict(ll)$class)))

[1] 51

## cross-validation:
confusion(ll2$class, joint$eng)

Overall accuracy = 0.461

Confusion matrix
   Predicted (cv)
Actual  0   1
  0 0.465 0.535
  1 0.545 0.455

sum(diag(table(ll2$class, joint$eng)))

[1] 35

## correspondence between predictions
confusion(ll2$class, predict(ll)$class)

Overall accuracy = 0.789

Confusion matrix
   Predicted (cv)
Actual  0   1
  0 0.791 0.209
  1 0.212 0.788

sum(diag(table(ll2$class, predict(ll)$class)))

[1] 60

In contrast to the 51 cases we got correct before, the cross-validation gets just 35 correct (out of 76)–this is actually worse than chance! This is in spite of the fact that there is a lot of agreement between the two models.

Notice that we no longer have a single lda model to look at. the output of the cross-validation is simpler. We can’t look at the linear discriminant values or means because there is no longer one model–we tested N models.

ll2

$class
[1] 0 0 0 1 1 0 0 0 1 0 0 0 1 0 1 0 1 1 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0
[36] 1 1 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 1 1 0 1 1 1 0 1 0 1 0 0 1 1 0 1
[71] 1 0 0 1 0 0
Levels: 0 1

$posterior
          0   1
11  0.7932692 0.2067308
12  0.5469743 0.4530257
13  0.6111616 0.3888384
22  0.1626239 0.8373761
31  0.1460898 0.8539102
32  0.6590130 0.3409870
33  0.5343306 0.4656694
41 0.6910390 0.3089610
51 0.1772707 0.8227293
52 0.5496802 0.4504198
61 0.5799697 0.4200303
62 0.7698881 0.2401019
63 0.4867715 0.5132285
64 0.7598981 0.2401019
71 0.4180372 0.5819628
81 0.5583747 0.4416253
101 0.4326410 0.5673590
102 0.3218178 0.6781822
104 0.2705828 0.7294172
201 0.7983831 0.2016169
202 0.5647722 0.4352278
203 0.2553252 0.7446748
301 0.5082852 0.4917148
302 0.6851039 0.3148961
303 0.5797293 0.4202707
304 0.2312416 0.7687584
401 0.5506530 0.4493470
402 0.5486566 0.4513434
501 0.5358578 0.4641422
502 0.3694875 0.6305125
503 0.8212536 0.1787464
504 0.5295038 0.4704962
505 0.4794108 0.5205892
507 0.6428161 0.3571839
508 0.5490367 0.4509633
509 0.4697326 0.5302674
510 0.4345844 0.5654156

$terms
eng ~ eer + eeu + ep + ppr + ppu + eer.1 + eeu.1 + ep.1 + ppr.1 +
ppu.1
attr(,"variables")
list(eng, eer, eeu, ep, ppr, ppu, eer.1, eeu.1, ep.1, ppr.1,
ppu.1)
attr(,"factors")
eer eeu ep ppr ppu eer.1 eeu.1 ep.1 ppr.1 ppu.1
eer  0  0  0  0  0  0  0  0  0  0
eeu  1  0  0  0  0  0  0  0  0  0
eep  0  1  0  0  0  0  0  0  0  0
ppr  0  0  0  1  0  0  0  0  0  0
ppu  0  0  0  1  0  0  0  0  0  0
eer.1 0  0  0  0  0  1  0  0  0  0
[ reached getOption("max.print") -- omitted 39 rows ]
attr(,"term.labels")
[1] "eer" "eeu" "ep" "ppr" "ppu" "eer.1" "eeu.1" "ep.1"
[9] "ppr.1" "ppu.1"
attr(,"order")
[1] 1 1 1 1 1 1 1 1 1 1
attr(,"intercept")
The best practice in a situation like this might be to use cross-validation accuracy to help guide variable selection. You might use a stepwise procedure, and only include a variable if it improves cross-validation accuracy. You might use the single best model at the end, but still acknowledge cross-validation performance. In this case, results such as this led our research lab to conclude that there was no substantial difference between groups, and we developed new behavioral methods that were more powerful.

## LDA with multiple categories

The other advantage of LDA over regression is that it handles multiple categories directly. Here, just as with the multinom() model, it creates $N - 1$ discriminant functions for $N$ classes, each compared against a baseline. Let’s look at the iris data for which we examined previously under the multinomial model.

```r
m.iris <- lda(Species ~ ., data = iris)
m.iris
```

Call:
```
lda(Species ~ ., data = iris)
```

Prior probabilities of groups:
```
setosa versicolor virginica
0.3333333 0.3333333 0.3333333
```

Group means:
```
     Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa     5.006       3.428       1.462      0.246
versicolor 5.936       2.770       4.260      1.326
virginica  6.588       2.974       5.552      2.026
```

Coefficients of linear discriminants:
```
     LD1     LD2
Sepal.Length  0.8293776 0.02410215
Sepal.Width   1.5344731 2.16452123
Petal.Length  -2.2012117 -0.93192121
Petal.Width   -2.8104603 2.83918785
```

Proportion of trace:
LD1   LD2
0.9912 0.0088

m.irisCV <- lda(Species ~ ., data = iris, CV = TRUE)

table(iris$Species, m.irisCV$class)

<table>
<thead>
<tr>
<th></th>
<th>setosa</th>
<th>versicolor</th>
<th>virginica</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>versicolor</td>
<td>0</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>virginica</td>
<td>0</td>
<td>1</td>
<td>49</td>
</tr>
</tbody>
</table>

Now, the classification is very good, even with cross-validation. We can see two sets of coefficients—the two discriminant functions distinguishing each pairing of outcomes.

**Quadratic Discriminant Analysis**

One of the assumptions of LDA is that the two distributions have equal variance. If we relax this assumption, the best classification no longer has to be a line separating the space. We can get curved boundaries, or even a small region within a larger region. For example:

qda1 <- qda(class ~ x + y, CV = TRUE)

table(class, qda1$class)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>387</td>
<td>113</td>
</tr>
<tr>
<td>B</td>
<td>123</td>
<td>377</td>
</tr>
</tbody>
</table>

n <- 500
class <- rep(c("A", "B"), each = n)
x <- rnorm(n * 2, mean = rep(c(3, 4), each = n), sd = rep(c(0.25, 1), each = n))
y <- rnorm(n * 2, mean = rep(c(2.5, 1.5), each = n), sd = rep(c(0.25, 1), each = n))

ggplot(data.frame(class, y, x), aes(x = x, y = y, colour = class)) + geom_point(size = 1.5) + geom_density_2d() + geom_segment(x = -1, y = 6.5, xend = 7, yend = -1.5, col = "grey60", lwd = 1.2) + geom_segment(x = 3, y = 2.5, xend = 4, yend = 1.5, col = "black", lwd = 1.2) + coord_fixed(ratio = 1, xlim = c(0, 6), ylim = c(0, 6)) + ggtitle("Color by true class")
```r
ggplot(data.frame(class, y, x), aes(x = x, y = y, colour = qda1$classification)) + geom_point(size = 1.5) + geom_density_2d() + geom_segment(x = -1, y = 6.5, xend = 7, yend = -1.5, col = "grey60", lwd = 1.2) + geom_segment(x = 3, y = 2.5, xend = 4, yend = 1.5, col = "black", lwd = 1.2) + coord_fixed(ratio = 1, xlim = c(0, 6), ylim = c(0, 6)) + ggtitle("Colored by QDA classification")
```
We can see how an LDA model will suffer. If all points are projected onto the discriminant line, the single boundary on that line will not be ideal. If we were able to make a curved boundary in this xy space, we could capture correct classifications. Quadratic Discriminant Analysis (QDA) permits this. It provides a more powerful classifier that can capture non-linear boundaries in the feature space. Thus, it is also less constrained, so requires more careful analysis to ensure we don’t overfit the model. How does it work with our real data set?

```r
library(MASS)
q <- qda(eng ~ ., data = joint, CV = TRUE)
q
```

```
$class
 [1] 0 1 0 1 1 1 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 0 0 1 1 1 1 0 0 1 0 0 0 1 1
[36] 0 0 0 1 1 1 0 0 0 1 0 0 1 1 0 1 1 0 1 0 1 0 1 0 1 1 1 1 1 0 0 0 1 0
[71] 1 0 0 1 0 0
Levels: 0 1

$posterior
     0         1
11 9.974693e-01 2.530667e-03
12 8.815744e-02 9.118426e-01
13 9.472289e-01 5.277114e-02
22 1.370973e-01 8.629027e-01
31 7.048625e-01 2.951375e-01
32 7.048625e-01 2.951375e-01
```
$terms
eng - eer + eeu + ep + ppr + ppu + eer.1 + eeu.1 + ep.1 + ppr.1 +
  ppu.1
attr("variables")
list(eng, eer, eeu, ep, ppr, ppu, eer.1, eeu.1, ep.1, ppr.1,
  ppu.1)
attr("factors")
eer eeu ep ppr ppu eer.1 eeu.1 ep.1 ppr.1 ppu.1
  eng  0  0  0  0  0  0  0  0  0  0
eer  1  0  0  0  0  0  0  0  0  0
eeu  0  1  0  0  0  0  0  0  0  0
eep  0  0  1  0  0  0  0  0  0  0
ppr  0  0  0  1  0  0  0  0  0  0
ppu  0  0  0  0  1  0  0  0  0  0
eer.1  0  0  0  0  1  0  0  0  0  0
attr("term.labels")
[1] "eer"  "eeu"  "ep"  "ppr"  "ppu"  "eer.1"  "eeu.1"  "ep.1"
[9] "ppr.1"  "ppu.1"
attr("order")
[1] 1 1 1 1 1 1 1 1 1 1 1
attr("intercept")
[1] 1
attr("response")
attr(".Environment")
<environment: R_GlobalEnv>
attr("predvars")
list(eng, eer, eeu, ep, ppr, ppu, eer.1, eeu.1, ep.1, ppr.1,  
ppu.1)
attr("dataClasses")
  eng  eer  eeu  ep  ppr  ppu  eer.1  
  "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"  
eeu.1  ep.1  ppr.1  ppu.1  
  "numeric" "numeric" "numeric" "numeric"

$call
qda(formula = eng ~ ., data = joint, CV = TRUE)

$xlevels
named list()

confusion(q$class, joint$eng)

Overall accuracy = 0.566

Confusion matrix
   Predicted (cv)
Actual   0 1
   0 0.556 0.444
   1 0.419 0.581

sum(diag(table(q$class, joint$eng)))

[1] 43

Now, the qda model is a reasonable improvement over the LDA model—even with Cross-validation. We were
at 46% accuracy with cross-validation, and now we are at 57%. This increased cross-validation accuracy from
35 to 43 accurate cases.

##Variable Selection in LDA
We now have a good measure of how well this model is doing. But we suspect
that—at least for LDA, the predictors might be over-fitting. We’d like to try removing variables to see if we
get a better cross-validation performance. We could do this by hand, or using some tools built for this. The
stepclass function within klaR package will do this:

library(klaR)
modelstepL <- stepclass(eng ~ ., "lda", direction = "both", data = joint)

correctness rate: 0.56786; in: "ppr"; variables (1): ppr

hr.elapsed min.elapsed sec.elapsed
  0.000  0.000  0.456

modelstepL

method : lda
final model : eng ~ ppr
<environment: 0x55f7184a8260>

correctness rate = 0.5679
modelstepQ <- stepclass(eng ~ ., "qda", direction = "both", data = joint)
correctness rate: 0.59286; in: "ppu"; variables (1): ppu

hr. elapsed min. elapsed sec. elapsed
0.000 0.000 0.455

modelstepQ

method : qda
final model : eng ~ ppu
<environment: 0x55f71c7e6590>
correctness rate = 0.5929

If you run this several times, you will find that you get a slightly different model each time. The best models have 1 to 2 predictors, and vary in accuracy from 55 to 65%. This is happening because the cross-validation the method uses is somewhat random, so the best model will depend on how the cross-validation is initialized. Perhaps if we reduce the improvement required, and use a higher cross-validation value, we will end up at a more stable result. Using fold=76 should be similar to doing leave-one-out cross-validation, and using a smaller improvement criterion will avoid stopping early.

modelstepL <- stepclass(eng ~ ., "lda", direction = "both", data = joint, improvement = 0.001, fold = 76)
correctness rate: 0.57895; in: "ppr"; variables (1): ppr
correctness rate: 0.59211; in: "ppu"; variables (2): ppr, ppu
correctness rate: 0.60526; in: "ppu.1"; variables (3): ppr, ppu, ppu.1

hr. elapsed min. elapsed sec. elapsed
0.000 0.000 5.333

modelstepL

method : lda
final model : eng ~ ppr + ppu + ppu.1
<environment: 0x55f7184eeee8>
correctness rate = 0.6053

modelstepQ <- stepclass(eng ~ ., "qda", direction = "both", data = joint, improvement = 0.001, fold = 76)
correctness rate: 0.57895; in: "ppr"; variables (1): ppr
correctness rate: 0.64474; in: "ep"; variables (2): ppr, ep

hr. elapsed min. elapsed sec. elapsed
0.000 0.000 3.255

modelstepQ

method : qda
final model : eng ~ ep + ppr
<environment: 0x55f71c2ee18>
correctness rate = 0.6447

Now, the each model tends to converge on the same result each time. The variables selected are different for the two models, but that is probably fine. We can refit the best models using lda and qda to get more details...
about the fit:

```r
l.final <- lda(eng ~ ppr + ppu + ppu.1, data = joint)
l.final
```

Call:
`lda(eng ~ ppr + ppu + ppu.1, data = joint)`

Prior probabilities of groups:
```
  0  1
0.5 0.5
```

Group means:
```
  ppr  ppu  ppu.1
0 0.5789474 0.6052632 3007.688
1 0.6754386 0.7017544 3432.817
```

Coefficients of linear discriminants:
```
LD1
  ppr  1.2417704238
  ppu  2.5194768800
  ppu.1 0.0004672492
```

```r
q.final <- qda(eng ~ ep + ppr, data = joint)
q.final
```

Call:
`qda(eng ~ ep + ppr, data = joint)`

Prior probabilities of groups:
```
  0  1
0.5 0.5
```

Group means:
```
  ep   ppr
0 0.6951754 0.5789474
1 0.6973684 0.6754386
```

**Example: LDA on the iphone data set**

The following work through all the steps of LDA and QDA again with the iphone data set. ## Data Preprocessing

```r
phone_ds <- read.csv("data_study1.csv")
phone_type <- phone_ds[, c(1, 3:13)]
phone_type[, 2:12] <- scale(phone_type[, 2:12], center = TRUE, scale = TRUE)
```

**Loading library for LDA**

```r
library(MASS)
```
Compute LDA without Cross validation

```r
lda_mod1 <- lda(Smartphone ~ ., data = phone_type)
lda_mod1

Call:
lda(Smartphone ~ ., data = phone_type)

Prior probabilities of groups:
  Android     iPhone
0.4139887 0.5860113

Group means:

<table>
<thead>
<tr>
<th>Age</th>
<th>Honesty.Humility</th>
<th>Emotionality</th>
<th>Extraversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android</td>
<td>0.2071126</td>
<td>0.2304728</td>
<td>-0.1885209</td>
</tr>
<tr>
<td>iPhone</td>
<td>-0.1463150</td>
<td>-0.1628179</td>
<td>0.1331809</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agreeableness</th>
<th>Conscientiousness</th>
<th>Openness</th>
<th>Avoidance.Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android</td>
<td>0.04975669</td>
<td>0.11922875</td>
<td>0.1678358</td>
</tr>
<tr>
<td>iPhone</td>
<td>-0.03515069</td>
<td>-0.08422934</td>
<td>-0.1185679</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phone.as.status.object</th>
<th>Social.Economic.Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android</td>
<td>0.04975669</td>
</tr>
<tr>
<td>iPhone</td>
<td>0.2013865</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time.owned.current.phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android</td>
</tr>
<tr>
<td>iPhone</td>
</tr>
</tbody>
</table>

Coefficients of linear discriminants:

<table>
<thead>
<tr>
<th>LD1</th>
<th>Age</th>
<th>Honesty.Humility</th>
<th>Emotionality</th>
<th>Extraversion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.24833268</td>
<td>-0.45905620</td>
<td>0.34275770</td>
<td>0.32566258</td>
</tr>
<tr>
<td></td>
<td>-0.1463150</td>
<td>-0.1628179</td>
<td>0.1331809</td>
<td>0.26431245</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LD1</th>
<th>Agreeableness</th>
<th>Conscientiousness</th>
<th>Openness</th>
<th>Avoidance.Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01021013</td>
<td>0.26431245</td>
<td>-0.15490974</td>
<td>-0.29824541</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LD1</th>
<th>Phone.as.status.object</th>
<th>Social.Economic.Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.38353264</td>
<td></td>
</tr>
</tbody>
</table>

plot(lda_mod1)
```
Predict on `phone_type`

```r
library(DAAG)
plda1 <- predict(object = lda_mod1)
confusion(phone_type$Smartphone, plda1$class)
```

Overall accuracy = 0.673

Confusion matrix

```
  Actual  Android iPhone
Predicted (cv)
  Android 0.484 0.516
  iPhone  0.194 0.806
```

Compute LDA with Cross-validation

```r
lda_mod2 <- lda(Smartphone ~ ., data = phone_type, CV = TRUE)
lda_mod2
```

```
$class
 [1] iPhone iPhone Android iPhone Android iPhone iPhone Android
 [9] iPhone iPhone Android iPhone iPhone Android iPhone iPhone
[17] Android iPhone Android Android Android iPhone iPhone iPhone
[25] iPhone Android Android iPhone Android iPhone iPhone iPhone
[33] iPhone iPhone iPhone iPhone iPhone iPhone iPhone iPhone
[41] iPhone iPhone iPhone iPhone iPhone iPhone iPhone iPhone
[49] iPhone iPhone iPhone iPhone iPhone iPhone iPhone iPhone
[57] Android iPhone Android iPhone iPhone Android Android iPhone
[65] Android iPhone iPhone iPhone Android iPhone Android Android
```

27
$posterior

Levels: Android iPhone

$terms

[73] iPhone iPhone iPhone
 [ reached getOption("max.print") -- omitted 454 entries ]
Levels: Android iPhone

Smartphone ~ Age + Honesty.Humility + Emotionality + Extraversion +
    Agreeableness + Conscientiousness + Openness + Avoidance.Similarity +
    Phone.as.status.object + Social.Economic.Status + Time.owned.current.phone
attr(,"variables")
list(Smartphone, Age, Honesty.Humility, Emotionality, Extraversion,
    Agreeableness, Conscientiousness, Openness, Avoidance.Similarity,
    Phone.as.status.object, Social.Economic.Status, Time.owned.current.phone)
attr(,"factors")
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Honesty.Humility</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Emotionality</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Openness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Avoidance.Similarity</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phone.as.status.object</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Social.Economic.Status</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time.owned.current.phone</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

[reached getOption("max.print") -- omitted 6 rows]

attr("term.labels")
[1] "Age"                 "Honesty.Humility"
[3] "Emotionality"         "Extraversion"
[5] "Agreeableness"        "Conscientiousness"
[7] "Openness"             "Avoidance.Similarity"

attr("order")
[1] 1 1 1 1 1 1 1 1 1 1 1 1

attr("intercept")
[1] 1

attr("response")
[1] 1

attr(".Environment")
<environment: R_GlobalEnv>

attr("predvars")

attr("dataClasses")

<table>
<thead>
<tr>
<th>Smartphone</th>
<th>Age</th>
<th>Honesty.Humility</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;factor&quot;</td>
<td>&quot;numeric&quot;</td>
<td>&quot;numeric&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Emotionality</th>
<th>Extraversion</th>
<th>Agreeableness</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;numeric&quot;</td>
<td>&quot;numeric&quot;</td>
<td>&quot;numeric&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conscientiousness</th>
<th>Openness</th>
<th>Avoidance.Similarity</th>
</tr>
</thead>
</table>
$\text{call}$

\texttt{lda(formula = Smartphone - ., data = phone_type, CV = TRUE)}

$xlevels$

named list()

\textbf{confusion() on lda_mod2}

\texttt{confusion(\texttt{phone_type$Smartphone, lda_mod2$\texttt{class}})}

Overall accuracy = 0.648

Confusion matrix

\begin{tabular}{lcc}
  & \text{Predicted (cv)} & \\
  \text{Actual} & \text{Android} & \text{iPhone} \\
  \text{Android} & 0.466 & 0.534 \\
  \text{iPhone} & 0.223 & 0.777 \\
\end{tabular}

We can see that some of the accuracy comes from overfitting.

\textbf{Compute QDA without CV}

\texttt{qda_mod1 <- qda(Smartphone - ., data = phone_type)}

\texttt{qda_mod1}

Call:

\texttt{qda(Smartphone - ., data = phone_type)}

Prior probabilities of groups:

\begin{tabular}{ll}
  \text{Android} & 0.4139887 \\
  \text{iPhone} & 0.5860113 \\
\end{tabular}

Group means:

\begin{tabular}{lccccc}
  & \text{Age} & \text{Honesty.Humility} & \text{Emotionality} & \text{Extraversion} \\
  \text{Android} & 0.2071126 & 0.2304728 & -0.1885209 & -0.08233383 \\
  \text{iPhone} & -0.1463150 & -0.1628179 & 0.1331809 & 0.05816486 \\
\end{tabular}

\begin{tabular}{lccccc}
  & \text{Agreeableness} & \text{Conscientiousness} & \text{Openness} & \text{Avoidance.Similarity} \\
  \text{Android} & 0.04975669 & -0.02859348 & 0.11922875 & 0.1678358 \\
  \text{iPhone} & -0.03515069 & 0.02019991 & -0.08422934 & -0.1185679 \\
\end{tabular}

\begin{tabular}{lcc}
  & \text{Phone.as.status.object} & \text{Social.Economic.Status} \\
  \text{Android} & -0.2850676 & -0.02423018 \\
  \text{iPhone} & 0.2013865 & 0.01711745 \\
\end{tabular}

\begin{tabular}{lcc}
  & \text{Time.owned.current.phone} \\
  \text{Android} & 0.06110396 \\
  \text{iPhone} & -0.04316699 \\
\end{tabular}

\texttt{qlda1 <- predict(object = qda_mod1)}

\texttt{confusion(\texttt{phone_type$Smartphone, qlda1$\texttt{class}})}
Overall accuracy = 0.682

Confusion matrix

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted (cv)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android</td>
<td>0.53 0.47</td>
</tr>
<tr>
<td>iPhone</td>
<td>0.21 0.79</td>
</tr>
</tbody>
</table>

Compute QDA with LOOC

```r
qda_mod2 <- qda(Smartphone ~ ., data = phone_type, CV = TRUE)
qda_mod2
```

$class

```
  [1] Android iPhone Android iPhone Android iPhone iPhone Android
  [9] iPhone iPhone iPhone Android iPhone Android iPhone iPhone
 [17] iPhone Android Android Android iPhone iPhone iPhone iPhone
 [25] iPhone Android Android Android iPhone iPhone iPhone iPhone
 [33] iPhone iPhone iPhone iPhone iPhone iPhone iPhone iPhone
 [41] iPhone Android iPhone iPhone iPhone iPhone iPhone iPhone
 [49] iPhone iPhone iPhone iPhone iPhone iPhone iPhone iPhone
 [57] Android Android Android iPhone iPhone iPhone iPhone iPhone
 [65] iPhone iPhone iPhone iPhone iPhone iPhone iPhone iPhone
 [73] iPhone Android iPhone

[ reached getOption("max.print") -- omitted 454 entries ]
```

Levels: Android iPhone

$posterior

```
Android iPhone
1 0.550069334 4.499307e-01
2 0.459988979 5.400110e-01
3 0.758731421 2.412686e-01
4 0.373753872 6.262461e-01
5 0.509395253 4.906047e-01
6 0.223179114 7.768209e-01
7 0.046547074 9.534529e-01
8 0.629604274 3.703957e-01
9 0.155903266 8.440967e-01
10 0.404783055 5.952169e-01
11 0.39001107 6.099999e-01
12 0.527163938 4.728361e-01
13 0.277148516 7.228515e-01
14 0.668604563 3.319354e-01
15 0.352077947 6.479221e-01
16 0.241226583 7.587734e-01
17 0.442156078 5.784393e-01
18 0.712822007 2.871780e-01
19 0.629898021 3.701020e-01
20 0.653601501 3.463985e-01
21 0.352077947 6.479221e-01
22 0.365360338 6.346397e-01
23 0.486570165 5.134298e-01
24 0.053057264 9.469427e-01
```
25  0.223667837 7.763322e-01
26  0.791062415 2.089376e-01
27  0.682312320 3.176877e-01
28  0.338804289 6.611957e-01
29  0.541435101 4.585649e-01
30  0.159967108 8.400329e-01
31  0.283987762 7.160122e-01
32  0.251433765 7.485662e-01
33  0.188424179 8.115758e-01
34  0.203354196 7.966458e-01
35  0.149893921 8.501061e-01
36  0.504465263 4.955347e-01
37  0.057381722 9.426183e-01
[ reached getOption("max.print") -- omitted 492 rows ]

$terms
Smartphone ~ Age + Honesty.Humility + Emotionality + Extraversion +
   Agreeableness + Conscientiousness + Openness + Avoidance.Similarity +
   Phone.as.status.object + Social.Economic.Status + Time.owned.current.phone
attr("variables")
list(Smartphone, Age, Honesty.Humility, Emotionality, Extraversion,
   Agreeableness, Conscientiousness, Openness, Avoidance.Similarity,
   Phone.as.status.object, Social.Economic.Status, Time.owned.current.phone)
attr("factors")

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Honesty.Humility</th>
<th>Emotionality</th>
<th>Extraversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Honesty.Humility</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Emotionality</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Agreeableness</th>
<th>Conscientiousness</th>
<th>Openness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Honesty.Humility</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Emotionality</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Avoidance.Similarity</th>
<th>Phone.as.status.object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Honesty.Humility</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Emotionality</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Social.Economic.Status</th>
<th>Time.owned.current.phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Age</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Honesty.Humility</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Emotionality</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
[ reached getOption("max.print") -- omitted 6 rows ]
attr("term.labels")
$call
qda(formula = Smartphone ~ ., data = phone_type, CV = TRUE)

$xlevels
named list()

## confusion() on qda_mod2

```
confusion(phone_type$Smartphone, qda_mod2$class)
```

Overall accuracy = 0.594

Confusion matrix

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted (cv)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android</td>
<td>0.420 0.580</td>
</tr>
<tr>
<td>iPhone</td>
<td>0.284 0.716</td>
</tr>
</tbody>
</table>

## Split-half Cross Validation

This function will compute an lda and qda model, randomly picking half each time. The model fits on 1/2 of
the data, and then predicts the other half.
cv.lda <- function(class, predictors) {
  selection <- rep(c(T, F), length.out = length(class))[order(runif(length(class)))]
  out1 <- class[selection]
  pred1 <- predictors[selection, ]
  joint1 <- data.frame(out = out1, pred1)
  out2 <- class[!selection]
  pred2 <- predictors[!selection, ]
  joint2 <- data.frame(out = NA, pred2)
  ll <- lda(out ~ ., data = joint1)
  table(predict(ll, newdata = joint2)$class, out2)
  out.ll <- sum(diag(table(predict(ll, newdata = joint2)$class, out2)))/length(out2)
  qq <- qda(out ~ ., data = joint1)
  table(predict(qq, newdata = joint2)$class, out2)
  out.qq <- sum(diag(table(predict(qq, newdata = joint2)$class, out2)))/length(out2)
  c(out.ll, out.qq)
}

cv.lda(phone_type$Smartphone, phone_type[, 2:12])

[1] 0.6098485 0.5871212

There is no advantage for the qda here, and the overall split-half cross-validation numbers are only 64%.

Step from klaR

We can automate variable selection with klaR

library(klaR)

model <- stepclass(Smartphone ~ ., data = phone_type, method = "lda", fold = 2,
  start.vars = 1:11, direction = "both", output = T)


modell2 <- lda(model$formula, data = phone_type)

confusion(phone_type$Smartphone, predict(modell2)$class)

Overall accuracy = 0.675

Confusion matrix

  Actual    Android iPhone
  Android  0.502  0.498
  iPhone  0.203  0.797

modelq <- stepclass(Smartphone ~ ., data = phone_type, method = "qda", fold = 2,
  start.vars = 1:11, direction = "both", output = T)

correctness rate: 0.63141; starting variables (11): Age, Honesty.Humility, Emotionality, Extraversion,

correctness rate: 0.6541; out: "Social.Economic.Status"; variables (9): Age, Emotionality, Extraversion, Agreeableness, Conscientiousness, Openness, Avoidance.Similarity, Phone.as.status.object, Time.owned.current.phone

hr.elapsed min.elapsed sec.elapsed
0.000 0.000 0.501

modelq2 <- qda(modelq$qformula, data = phone_type)

confusion(phone_type$Smartphone, predict(modelq2)$class)

Overall accuracy = 0.675

Confusion matrix
  Predicted (cv)
  Actual       Android iPhone
    Android   0.493 0.507
    iPhone     0.197 0.803

This appears to improve things; by fitting a smaller model we actually do better.

Applications of LDA

Although the performance of LDA can often be surpassed by more modern machine learning methods, there are several reasons it still sees widespread use.

- It is simple to use and understand. Like logistic regression, it can be used to make a simple model or decision tool that is both easy to implement and transparent.
- It is sufficient for many situations. Many times, the benefit you might get from using a more complex model is negligible, at the cost of complexity or (worse yet) the possibility of making large mistakes because of strange interactions that you might not be able to predict.

Some of the most widely-used LDA models are within finance. For example, Altman’s (1968) bankruptcy model is based on LDA, predicting bankruptcy of firms within the next two years based on a handful of publicly-available statistics (see Altman, 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.) This is nice because the model can be implemented in a spreadsheet and decisions can be made by individuals evaluating stock purchases.

Alternatives and extensions in Machine Classification

There are hundreds of special-purpose methods available for machine classification, many of which are developed for special kinds of situations or that work under different assumptions. We will cover several of these in this class, and here is a partial listing of methods you might want to be familiar with:

Within the klaR library, there are several implementations of related methods

- rda: Regularized discriminant Analysis. Attempts to build a discriminat model that is more robust to correlation between predictors (multi-collinearity)
- Probabilistic LDA. This frames the LDA problem in a Bayesian and/or maximum likelihood format, and is increasingly used as part of deep neural nets as a ‘fair’ final decision that does not hide complexity.
- loclda: Makes a local lda for each point, based on its nearby neighbors.
- sknn: simple k-nearest-neighbors classification. Makes classification based on a vote of the nearest observations
- NaiveBayes: A common and simple classifier based on bayes rule
• **svmlight**: a lightweight ‘support vector machine’, which generalizes lda, focusing especially on identifying a good decision rule that separates the two groups.

The klaR library also has a lot of functions to help with variable selection and cross-validation.

Within the nnet library:

• **nnet**: a neural net classifier—essentially a network of LDA classifiers or logistic regressions.
• **multinom**: an extension of generalized linear regression for multiple groups.