

Multivariate Statistical Analysis

Homework 4

Lucas Fellmeth, Helen Kafka, Sven Bergmann

03/13/24

```
set.seed(42)
```

Problem 1

Consider the `bupa_liver_disorder` data set, where a number of individuals are classified into two classes according to blood test results (see full description on p. 258 of the book).

```
bupa_liver_disorder <- read.csv(file = "Data_csv/bupa_liver_disorder.csv")
bupa_liver_disorder$class <- factor(bupa_liver_disorder$class)
```

a)

Split the data set into training and test sets (roughly a 70/30 split). Compute the logistic classifier using the training set.

```
i_test <- sample(seq_along(bupa_liver_disorder[, 1]), size = length(bupa_liver_disorder[, 1]) * 0.3)
bupa_liver_disorder_test <- bupa_liver_disorder[i_test, ]
bupa_liver_disorder_train <- bupa_liver_disorder[-i_test, ]
```

```
out <- glm(formula = class ~ mcv + alkphos + sgpt + sgot + gammagt +
           drinks, data = bupa_liver_disorder_train, family = binomial)
out$coefficients
```

```
## (Intercept)          mcv      alkphos      sgpt      sgot      gammagt
##  5.01564297 -0.05348287 -0.01975076 -0.06741025  0.13150896  0.01971347
##      drinks
## -0.08709490
```

b)

Construct the misclassification table and compute the misclassification rate for the test set.

```
pil_hat <- predict(out, type = "response", newdata = bupa_liver_disorder_test)
gr_hat <- ifelse(pil_hat > 0.5, 1, 2)
mctable <- table(gr_hat, bupa_liver_disorder_test$class)
mctable
```

```
##
## gr_hat  1  2
##      1 19 44
##      2 28 12
```

```
1 - sum(diag(mctable))/length(bupa_liver_disorder_test$class)
```

```
## [1] 0.6990291
```

Problem 2

Consider the `ecoli` data set. These data were obtained in a study of protein localization sites for 336 examples of *E. coli*. There are 7 predictor variables and a class variable, `localization_site`, which indicates the protein localization.

```
ecoli <- read.csv(file = "Data_csv/ecoli.csv")
ecoli$localization_site <- factor(ecoli$localization_site)
```

a)

Split the data set into training and test sets (roughly a 80/20 split). Compute the multinomial logistic classifier using the training set.

```
i_test <- sample(seq_along(ecoli[, 1]), size = length(ecoli[,
1]) * 0.2)
ecoli_test <- ecoli[i_test, ]
ecoli_train <- ecoli[-i_test, ]
```

```
out <- nnet::multinom(localization_site ~ mvg + gvh + lip + chg +
aac + alm1 + alm2, data = ecoli, maxit = 100)
```

```
## # weights: 72 (56 variable)
## initial value 698.692358
## iter 10 value 175.691378
## iter 20 value 112.464952
## iter 30 value 107.295386
## iter 40 value 105.913978
## iter 50 value 105.487896
## iter 60 value 105.353525
## iter 70 value 105.242112
## iter 80 value 105.109689
## iter 90 value 105.014041
## iter 100 value 104.969071
## final value 104.969071
## stopped after 100 iterations
```

```
out
```

```
## Call:
```

```
## nnet::multinom(formula = localization_site ~ mvg + gvh + lip +
## chg + aac + alm1 + alm2, data = ecoli, maxit = 100)
```

```
##
```

```
## Coefficients:
```

	(Intercept)	mvg	gvh	lip	chg	aac
## im	-14.179741	-1.198919	5.8227058	-1.671139	-7.791876	-0.7240761
## imL	-42.802584	37.770240	-49.7078462	33.274120	18.860327	-23.7806465
## imS	-15.930816	14.647235	3.9962416	-7.935507	-7.823256	3.0728426
## imU	-17.740856	10.842884	0.8374873	2.285835	-9.110530	-1.1427249
## om	-29.106326	7.400551	22.5668026	1.679383	-14.606608	39.1260334
## omL	-14.537074	13.877294	-14.1367197	28.299112	-21.495756	8.0589385
## pp	-4.044233	6.299159	15.6413210	-20.001416	-4.885281	0.4754381

```
##          alm1          alm2
## im  36.37688  -4.653052
## imL 18.57044   4.957740
## imS 26.74233  -9.901364
## imU 29.40923  -2.259731
## om   11.01966 -27.030325
## omL 21.53240 -31.368635
## pp   20.56965 -12.007667
##
## Residual Deviance: 209.9381
## AIC: 321.9381

mctable <- table(predict(out, newdata = ecoli_test), ecoli_test$localization_site)

1 - sum(diag(mctable))/length(ecoli_test$localization_site)

## [1] 0.04477612
```