EEBL - Business Statistics

Assignment 3

The dataset considered for this assignment originated from 4601 emails sent to an individual (named George, at HP labs, before 2000). For each email message the true outcome, labeled as 1 if the message is `spam` or 0 if the message is a valid `email`, is available, along with 57 input variables measuring the relative frequencies with which selected words or punctuation marks, listed in table 1, occurred. Variables 55-57 measure the length of sequences of consecutive capital letters. See Hastie, Tibshirani and Freedman (HTF), page 2 and section 9.1.2 for further details.

The data set is available as the file spam.csv and is accessed in R using the following statements.

```
rm(list = ls())   # remove previous objects
spam = read.csv("spam.csv", sep = ",", header = T);
attach(spam)
table(y)
```

The objective of the assignment is to construct and validate a classifier of whether an incoming email is 'spam' or a valid message (this is the basis of an anti-spam filter).

The 4601 observations are divided into a training set, consisting of 3200 observations, while the rest is used for validation. Use the same seed as in assignment 2. This could be achieved by the following statements:

```
Nobs = nrow(spam)
names(spam)
set.seed(     ) # set the seed as in assignment 2
N = 3200
s = sample( 1:Nobs, N)
train = spam[s,]
test = spam[-s,]
```

You should address the following points:

a. Fit the full logistic model using all the variables as inputs. Use the following code:

```
model_full = factor(y)~
  A.1+A.2+A.3+A.4+A.5+A.6+A.7+A.8+A.9+A.10+A.11+A.12+A.13+A.14+A.15+A.16+A.17+A.18+A.19+
  A.20+A.21+A.22+A.23+A.24+A.25+A.26+A.27+A.28+A.29+A.30+A.31+A.32+A.33+A.34+A.35+A.36+
  A.37+A.38+A.39+A.40+A.41+A.42+A.43+A.44+A.45+A.46+A.47+A.48+A.49+A.50+A.51+A.52+A.53+
  A.54+A.55+A.56+A.57;
full = glm(model_full,  family=binomial(link="logit"), data = train);
summary(full);
```

Comment on the results and on the goodness of fit as measured by the deviance and the missclassification rate.

b. Perform variable selection by Forward Stepwise, which can be implemented here as

```
# stepwise selection (using the standard stat library)
sel.f = step(restr, scope=formula(full), direction="forward",  steps = 50, k=log(N))
summary(sel.f)
```

here restr is the null model containing only the intercept, i.e.

```
restr = glm(factor(y)~1, family=binomial(link="logit"), data = train).
```

c. Evaluate the missclassification rate in the test sample: do you achieve about the same rate by using the selected model compared to the full model?

```
# missclassification error in training sample
pred.full = predict.glm( full, train[,-58], type="response" )
merror.full.train = mean( as.numeric(pred.full >0.5) != train$y)
merror.full.train
# missclassification error in test sample
pred.full = predict( full, test[,-58], type="response" )
merror.full = mean( as.numeric(pred.full >0.5) != test$y)
merror.full
# missclassification error in test sample
pred.sel = predict(sel.f, test[,-58], type="response" )
merror.sel = mean( as.numeric(pred.sel >0.5) != test$y)
merror.sel
```

c. Perform estimation by Lasso using the R statements below. Read the Glm Vignette.

```
xmat = as.matrix(train[,-58])
yvar = as.factor(train$y)
glmod = glmnet(xmat, yvar, alpha=1, family="binomial")
plot(glmod, xvar="lambda")
cvfit = cv.glmnet(xmat, as.factor(y[s]), family = "binomial", type.measure = "class")
plot(cvfit)
cvfit$lambda.min
coef(cvfit, s = "lambda.min")
```

Is the estimated model similar to that selected by forward stepwise?

Is the predictive performance in the test sample improved with respect to the full model? To predict the out of sample observations use:

```
predict(cvfit, newx = as.matrix(test[,-58]), s = "lambda.min", type = "class")
```

This assignment is due by 11:00 p.m. of October 17. You should upload in

https://www.dropbox.com/request/90mzH2xBRAbqYhwhu0Y0

a unique pdf file with the results and your comments. The filename must be YourSurname_YourName_Assignment3.pdf.

Table 1: Words and punctuation marks considered by variables 1-54

Variable	Word		Variable	Word		Variable	Word		Variable	Word		Variable	Word		Variable	Word
1.	make		10.	mail		19.	you		28.	650		37.	1999		46.	edu
2.	address		11.	receive		20.	credit		29.	lab		38.	parts		47.	table
3.	all		12.	will		21.	your		30.	labs		39.	pm		48.	conference
4.	3d		13.	people		22.	font		31.	telnet		40.	direct		49.	;
5.	our		14.	report		23.	0		32.	857		41.	cs		50.	(
6.	over		15.	addresses		24.	money		33.	data		42.	meeting		51.	[
7.	remove		16.	free		25.	hp		34.	415		43.	original		52.	!
8.	internet		17.	business		26.	hpl		35.	85		44.	project		53.	$
9.	order		18.	email		27.	george		36.	technology		45.	re		54.	#