STAT 154 Lab 10: Logistic regression and support vector machines

Yuansi Chen and Raaz Dwivedi

Apr 22, 2019

1 White-board discussion on Logistic regression and support vector machines

1. Quick derivation of the two methods and Comparison to LDA.
2. Discuss the possibility of kernel versions.
3. Discuss the regularized versions.

2 Parameter estimation in Logistic regression

3 The stock market smarket data with logistic regression

You will be working with the **Smarket** data, which is part of the "ISLR" package. This data consists of percentage returns fro the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005. For each date, the percentage returns for each of the five previous tradings has been records, **Lag1** through **Lag5**. Other variables are:

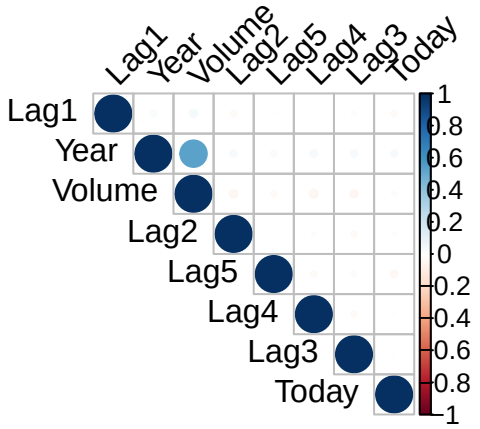
- **Volume** = the number of shares traded on the prevous day, in billions
- **Today** = the percentage return on the data in question
- **Direction** = whether the market was Up or Down on this date

```
# remember to load package ISLR
library(ISLR)
names(Smarket)

## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"

dim(Smarket)

## [1] 1250    9

summary(Smarket)

##       Year           Lag1                Lag2
##  Min.   :2001   Min.   :-4.922000   Min.   :-4.922000
##  1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500
##  Median :2003   Median : 0.039000   Median : 0.039000
```

```
##  Mean   :2003   Mean   : 0.003834   Mean   : 0.003919
##  3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750
##  Max.   :2005   Max.   : 5.733000   Max.   : 5.733000
##       Lag3                Lag4                Lag5
##  Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.92200
##  1st Qu.:-0.640000   1st Qu.:-0.640000   1st Qu.:-0.64000
##  Median : 0.038500   Median : 0.038500   Median : 0.03850
##  Mean   : 0.001716   Mean   : 0.001636   Mean   : 0.00561
##  3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.59700
##  Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.73300
##      Volume            Today           Direction
##  Min.   :0.3561   Min.   :-4.922000   Down:602
##  1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648
##  Median :1.4229   Median : 0.038500
##  Mean   :1.4783   Mean   : 0.003138
##  3rd Qu.:1.6417   3rd Qu.: 0.596750
##  Max.   :3.1525   Max.   : 5.733000
```

1. Compute the matrix of correlations of the variables in Smarket, excluding the variable **Direction**.

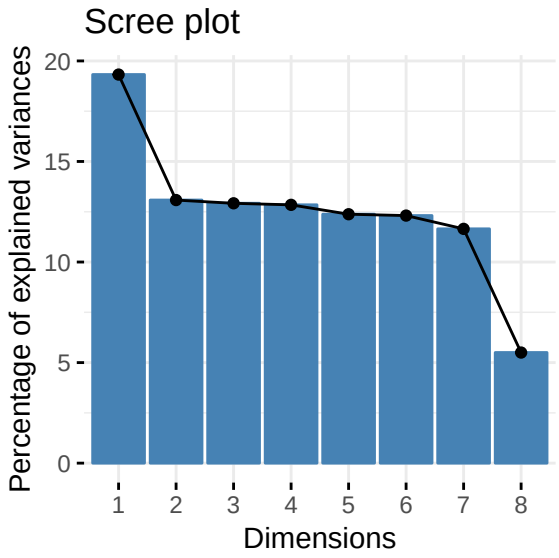
```
cor.mat <- cor(Smarket[, -9])
library(corrplot)
corrplot(cor.mat, type="upper", order="hclust",
         tl.col="black", tl.srt=45)
```



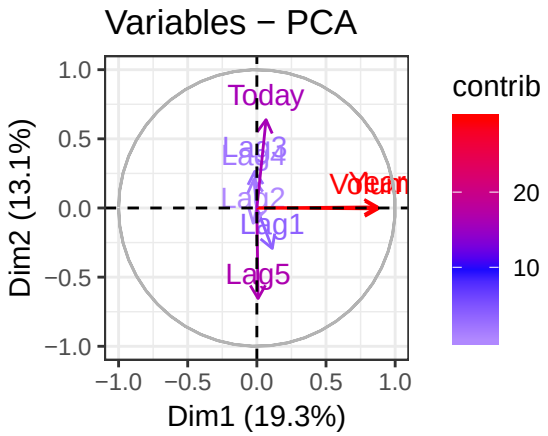
2. Perform a PCA on Smarket[,-9] to get a visual display of the variables. You can accomplish this with the function PCA() from the "FactoMineR" package. By default, it plots a circle of correlations.

```
library(FactoMineR)
library(factoextra)
res.pca <- PCA(Smarket[, -9], graph = FALSE)
print(res.pca)
```

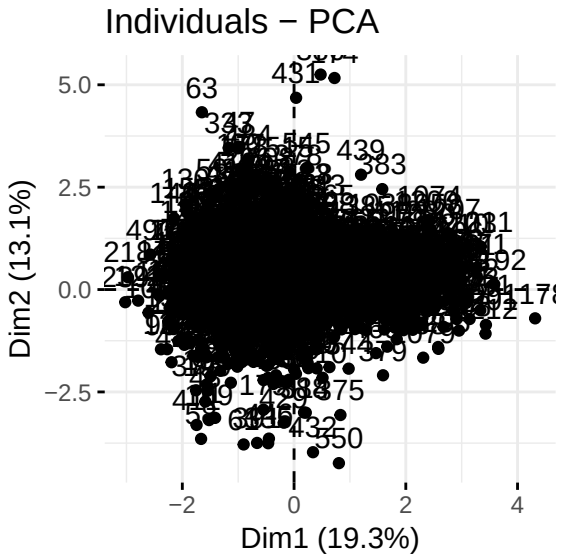
```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 1250 individuals, described by 8 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"

eigenvalues <- res.pca$eig
head(eigenvalues[, 1:2])

##        eigenvalue percentage of variance
## comp 1  1.5456704              19.32088
## comp 2  1.0464672              13.08084
## comp 3  1.0336139              12.92017
## comp 4  1.0274707              12.84338
## comp 5  0.9901911              12.37739
## comp 6  0.9847526              12.30941

# screeplot
fviz_screeplot(res.pca, ncp=10)
```




```
# PCA graph of variables
# Control variable colors using their contribution
# Possible values for the argument col.var are :
  # "cos2", "contrib", "coord", "x", "y"
fviz_pca_var(res.pca, col.var="contrib")+
scale_color_gradient2(low="white", mid="blue",
                      high="red", midpoint=10)+theme_bw()
```




```
# PCA graph of individuals
fviz_pca_ind(res.pca)
```



3. How correlated are the lag variables with today's returns? Are previous day's returns highly correlated with today's returns?

4. Make a scatterplot of **Year** and **Volume**.

5. Use the **glm()** function to fit a logistic regression model in order to predict **Direction** using **Lag1** through **Lag5** and **Volume**.

```
res.logistic <- glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Smarket, family = 'binomial')
summary(res.logistic)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = Smarket)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.446  -1.203   1.065   1.145   1.326
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.126000   0.240736  -0.523    0.601
## Lag1        -0.073074   0.050167  -1.457    0.145
## Lag2        -0.042301   0.050086  -0.845    0.398
## Lag3         0.011085   0.049939   0.222    0.824
## Lag4         0.009359   0.049974   0.187    0.851
## Lag5         0.010313   0.049511   0.208    0.835
## Volume       0.135441   0.158360   0.855    0.392
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1731.2  on 1249  degrees of freedom
```

```
## Residual deviance: 1727.6  on 1243  degrees of freedom
## AIC: 1741.6
##
## Number of Fisher Scoring iterations: 3

predicted_train <- predict(res.logistic, Smarket, type="response")
predicted_train[1:10]

##         1         2         3         4         5         6         7
## 0.5070841 0.4814679 0.4811388 0.5152224 0.5107812 0.5069565 0.4926509
##         8         9        10
## 0.5092292 0.5176135 0.4888378
```

6. Inspect the summary() of the "glm" object containing the output of the logistic regression.
7. Looking at the p-values of the regression coefficients, which coefficient seems to be significant?
8. What is the coefficient value of **Lag1**? How would you interpret the sign of this coefficient?
9. Use the **predict()** function to predict the probability that the market will go up, given values of the predictors. Use the argument type = "response" which tells R to output probabilities of the form P(Y = 1—X), as oppose to other information such as the logit.

4 A Comparison of Classification Methods (optional)

Recommended reading: Chapter 4.5 A Comparison of Classification Methods in ISL book. You should be able to generate the plots in Figure 4.10 and Figure 4.11.