Assignment 4 - Classification

Conceptual

1. Discuss the differences between LDA and QDA in terms of their main assumptions about classes, decision boundaries, number of samples, and overfitting.

Your answer here

2. Regarding KNN:

(a) How does the choice of distance metric affect the performance of k-NN classification?

Your answer here

(b) Please also discuss the concept of the curse of dimensionality and its implications for k-NN algorithm.

Your answer here

Practical

Overview of the steps

1. Load the data and get an overview of the data
2. Perform a logistic regression
3. Use the logistic regression models
4. Perform an LDA
5. Use the LDA regression model
6. Perform an QDA
7. Use the QDA regression model
8. Use k-Nearest Neighbors (KNN)

Steps in detail

Load the data and get an overview of the data

Load the data file `Smarket.rda` or `Smarket.csv` .

This data set consists of percentage returns for a stock index over $1,250$ days. For each date, it contains the percentage returns for each of the five previous trading days, `Lag1` through `Lag5`. It also contains `Volume` (the number of shares traded on the previous day, in billions), `Today` (the percentage return on the date in question) and `Direction` (whether the market was Up or Down on this date).

In [85]:
```
load(file = "../ISLR/data/Smarket.rda")
```

Display the number of predictors and possible responses and their names:

In [86]:
```
dim(Smarket)[2]
names(Smarket)
```

9

'Year' 'Lag1' 'Lag2' 'Lag3' 'Lag4' 'Lag5' 'Volume' 'Today' 'Direction'

Print a statistic summary of the predictors and responses:

In [87]:
```
summary(Smarket)
```

```
      Year           Lag1                Lag2                Lag3          
 Min.   :2001   Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.9220
00
 1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500   1st Qu.:-0.6400
00
 Median :2003   Median : 0.039000   Median : 0.039000   Median : 0.0385
00
 Mean   :2003   Mean   : 0.003834   Mean   : 0.003919   Mean   : 0.0017
16
 3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.5967
50
 Max.   :2005   Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.7330
00
      Lag4                Lag5              Volume           Today        
 Min.   :-4.922000   Min.   :-4.92200   Min.   :0.3561   Min.   :-4.922
000
 1st Qu.:-0.640000   1st Qu.:-0.64000   1st Qu.:1.2574   1st Qu.:-0.639
500
 Median : 0.038500   Median : 0.03850   Median :1.4229   Median : 0.038
500
 Mean   : 0.001636   Mean   : 0.00561   Mean   :1.4783   Mean   : 0.003
138
 3rd Qu.: 0.596750   3rd Qu.: 0.59700   3rd Qu.:1.6417   3rd Qu.: 0.596
750
 Max.   : 5.733000   Max.   : 5.73300   Max.   :3.1525   Max.   : 5.733
000
 Direction
 Down:602
 Up  :648
```

Display the number of data points:

In [88]:
```
dim(Smarket)[1]
```

```
1250
```

Display the data in a table (subset of rows is sufficient):

In [89]:
```
1 Smarket
```

A data.frame: 1250 × 9

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	2001	0.381	-0.192	-2.624	-1.055	5.010	1.1913	0.959	Up
2	2001	0.959	0.381	-0.192	-2.624	-1.055	1.2965	1.032	Up
3	2001	1.032	0.959	0.381	-0.192	-2.624	1.4112	-0.623	Down
4	2001	-0.623	1.032	0.959	0.381	-0.192	1.2760	0.614	Up
5	2001	0.614	-0.623	1.032	0.959	0.381	1.2057	0.213	Up
6	2001	0.213	0.614	-0.623	1.032	0.959	1.3491	1.392	Up
7	2001	1.392	0.213	0.614	-0.623	1.032	1.4450	-0.403	Down
8	2001	-0.403	1.392	0.213	0.614	-0.623	1.4078	0.027	Up
9	2001	0.027	-0.403	1.392	0.213	0.614	1.1640	1.303	Up
10	2001	1.303	0.027	-0.403	1.392	0.213	1.2326	0.287	Up
11	2001	0.287	1.303	0.027	-0.403	1.392	1.3090	-0.498	Down
12	2001	-0.498	0.287	1.303	0.027	-0.403	1.2580	-0.189	Down
13	2001	-0.189	-0.498	0.287	1.303	0.027	1.0980	0.680	Up
14	2001	0.680	-0.189	-0.498	0.287	1.303	1.0531	0.701	Up
15	2001	0.701	0.680	-0.189	-0.498	0.287	1.1498	-0.562	Down
16	2001	-0.562	0.701	0.680	-0.189	-0.498	1.2953	0.546	Up
17	2001	0.546	-0.562	0.701	0.680	-0.189	1.1188	-1.747	Down
18	2001	-1.747	0.546	-0.562	0.701	0.680	1.0484	0.359	Up
19	2001	0.359	-1.747	0.546	-0.562	0.701	1.0130	-0.151	Down
20	2001	-0.151	0.359	-1.747	0.546	-0.562	1.0596	-0.841	Down
21	2001	-0.841	-0.151	0.359	-1.747	0.546	1.1583	-0.623	Down
22	2001	-0.623	-0.841	-0.151	0.359	-1.747	1.1072	-1.334	Down
23	2001	-1.334	-0.623	-0.841	-0.151	0.359	1.0755	1.183	Up
24	2001	1.183	-1.334	-0.623	-0.841	-0.151	1.0391	-0.865	Down
25	2001	-0.865	1.183	-1.334	-0.623	-0.841	1.0752	-0.218	Down
26	2001	-0.218	-0.865	1.183	-1.334	-0.623	1.1503	0.812	Up
27	2001	0.812	-0.218	-0.865	1.183	-1.334	1.1537	-1.891	Down
28	2001	-1.891	0.812	-0.218	-0.865	1.183	1.2572	-1.736	Down
29	2001	-1.736	-1.891	0.812	-0.218	-0.865	1.1122	-1.851	Down
30	2001	-1.851	-1.736	-1.891	0.812	-0.218	1.2085	-0.195	Down
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1221	2005	0.179	-0.385	-0.078	0.305	0.845	2.12158	0.941	Up

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1222	2005	0.941	0.179	-0.385	-0.078	0.305	2.29804	0.440	Up
1223	2005	0.440	0.941	0.179	-0.385	-0.078	2.45329	0.527	Up
1224	2005	0.527	0.440	0.941	0.179	-0.385	2.11735	0.508	Up
1225	2005	0.508	0.527	0.440	0.941	0.179	2.29142	0.347	Up
1226	2005	0.347	0.508	0.527	0.440	0.941	1.98540	0.209	Up
1227	2005	0.209	0.347	0.508	0.527	0.440	0.72494	-0.851	Down
1228	2005	-0.851	0.209	0.347	0.508	0.527	2.01690	0.002	Up
1229	2005	0.002	-0.851	0.209	0.347	0.508	2.26834	-0.636	Down
1230	2005	-0.636	0.002	-0.851	0.209	0.347	2.37469	1.216	Up
1231	2005	1.216	-0.636	0.002	-0.851	0.209	2.61483	0.032	Up
1232	2005	0.032	1.216	-0.636	0.002	-0.851	2.12558	-0.236	Down
1233	2005	-0.236	0.032	1.216	-0.636	0.002	2.32584	0.128	Up
1234	2005	0.128	-0.236	0.032	1.216	-0.636	2.11074	-0.501	Down
1235	2005	-0.501	0.128	-0.236	0.032	1.216	2.09383	-0.122	Down
1236	2005	-0.122	-0.501	0.128	-0.236	0.032	2.17830	0.281	Up
1237	2005	0.281	-0.122	-0.501	0.128	-0.236	1.89629	0.084	Up
1238	2005	0.084	0.281	-0.122	-0.501	0.128	1.87655	0.555	Up
1239	2005	0.555	0.084	0.281	-0.122	-0.501	2.39002	0.419	Up
1240	2005	0.419	0.555	0.084	0.281	-0.122	2.14552	-0.141	Down
1241	2005	-0.141	0.419	0.555	0.084	0.281	2.18059	-0.285	Down
1242	2005	-0.285	-0.141	0.419	0.555	0.084	2.58419	-0.584	Down
1243	2005	-0.584	-0.285	-0.141	0.419	0.555	2.20881	-0.024	Down
1244	2005	-0.024	-0.584	-0.285	-0.141	0.419	1.99669	0.252	Up
1245	2005	0.252	-0.024	-0.584	-0.285	-0.141	2.06517	0.422	Up
1246	2005	0.422	0.252	-0.024	-0.584	-0.285	1.88850	0.043	Up
1247	2005	0.043	0.422	0.252	-0.024	-0.584	1.28581	-0.955	Down
1248	2005	-0.955	0.043	0.422	0.252	-0.024	1.54047	0.130	Up
1249	2005	0.130	-0.955	0.043	0.422	0.252	1.42236	-0.298	Down
1250	2005	-0.298	0.130	-0.955	0.043	0.422	1.38254	-0.489	Down

Compute the pairwise correlation of the predictors in the data set.

In R , we need to download and install a library first.

In [90]:
```
1 install.packages("corrplot")
2 source("http://www.sthda.com/upload/rquery_cormat.r")
```

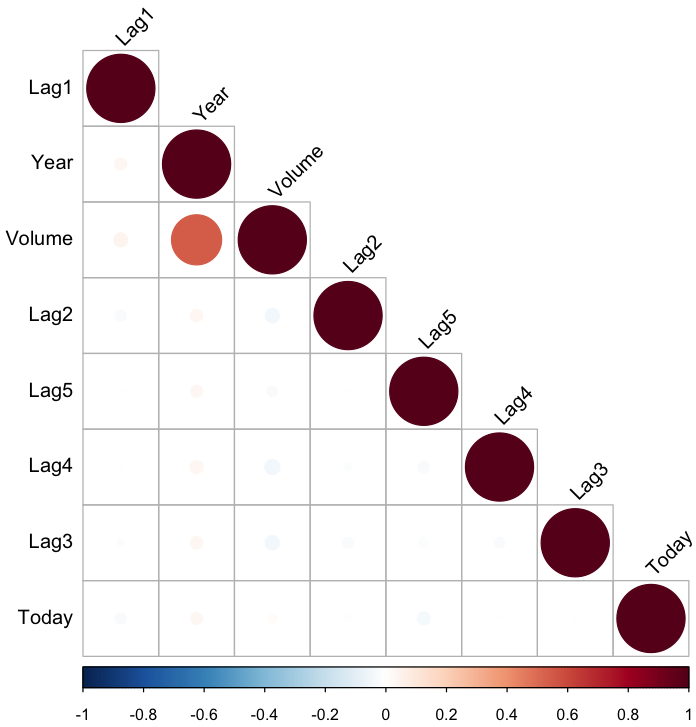
```
The downloaded binary packages are in
        /var/folders/ct/4pcck8t94sdfc73rhymq4t140000gp/T//Rtmpt4aO3y/do
wnloaded_packages
```

In [91]:
```
1 rquery.cormat(Smarket[,-9])
```

```
$r
          Lag1  Year Volume    Lag2   Lag5    Lag4    Lag3 Today
Lag1         1
Year      0.03     1
Volume   0.041  0.54      1
Lag2    -0.026 0.031 -0.043       1
Lag5   -0.0057  0.03 -0.022 -0.0036      1
Lag4    -0.003 0.036 -0.048  -0.011 -0.027       1
Lag3    -0.011 0.033 -0.042  -0.026 -0.019  -0.024       1
Today   -0.026  0.03  0.015   -0.01 -0.035 -0.0069 -0.0024     1

$p
       Lag1  Year Volume Lag2 Lag5 Lag4 Lag3 Today
Lag1      0
Year   0.29     0
Volume 0.15 4e-95      0
Lag2   0.35  0.28   0.13    0
Lag5   0.84  0.29   0.44  0.9    0
Lag4   0.92  0.21  0.087  0.7 0.34    0
Lag3    0.7  0.24   0.14 0.36 0.51  0.4    0
Today  0.36  0.29   0.61 0.72 0.22 0.81 0.93     0

$sym
       Lag1 Year Volume Lag2 Lag5 Lag4 Lag3 Today
Lag1   1
Year        1
Volume      .    1
Lag2                    1
Lag5                         1
Lag4                              1
Lag3                                   1
Today                                       1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```



Interprete the results. *Your interpretation of the results goes here!*

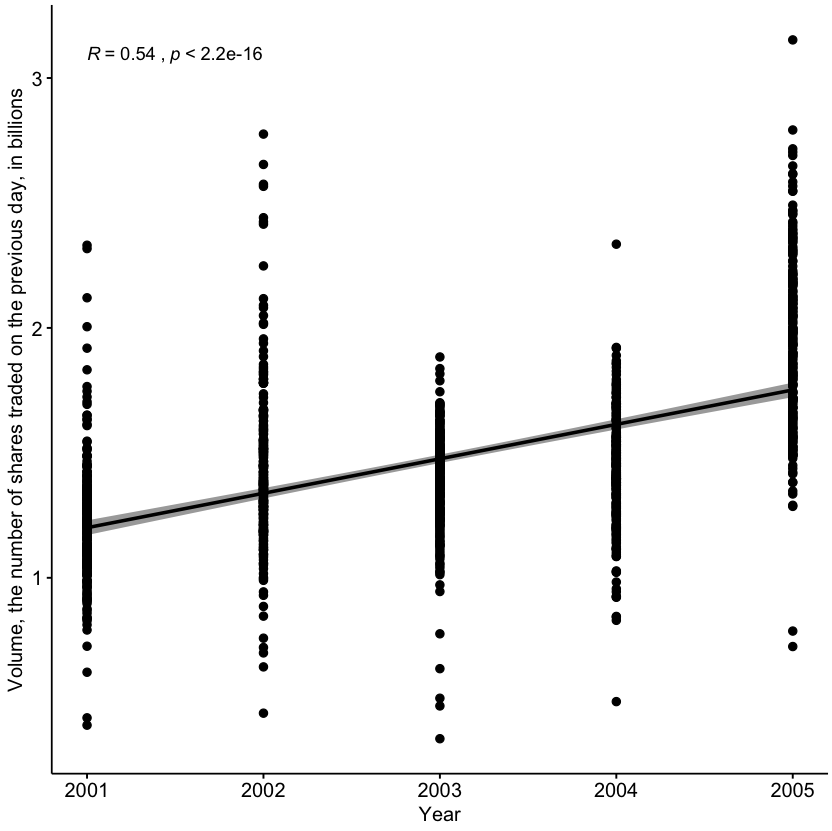
Plot the correlated predictors `Volume` and `Year`.

In R, we need to download and install a library first.

In [92]:
```
install.packages("ggpubr")
library("ggpubr")
```

```
The downloaded binary packages are in
	/var/folders/ct/4pcck8t94sdfc73rhymq4t140000gp/T//Rtmpt4aO3y/do
wnloaded_packages
```

In [93]:
```
ggscatter(Smarket, x = "Year", y = "Volume",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Year",
          ylab = "Volume, the number of shares traded on the previou
```



Interprete the results. *Your interpretation of the results goes here!*

Perform logistic regressions

Fit a logistic regression model in order to predict `Direction` using `Lag1` through `Lag5` and `Volume` .

In `R` , the `glm()` function fits generalized linear models, a class of models that includes logistic regression. The syntax of the `glm()` function is similar to that of `lm()` , except that we must pass in the argument `family=binomial` in order to run a logistic regression rather than some other type of generalized linear model.

In [94]:
```
glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Smarket,
summary(glm.fit)$coef
```

A matrix: 7 × 4 of type dbl

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.126000257	0.24073574	-0.5233966	0.6006983
Lag1	-0.073073746	0.05016739	-1.4565986	0.1452272
Lag2	-0.042301344	0.05008605	-0.8445733	0.3983491
Lag3	0.011085108	0.04993854	0.2219750	0.8243333
Lag4	0.009358938	0.04997413	0.1872757	0.8514445
Lag5	0.010313068	0.04951146	0.2082966	0.8349974
Volume	0.135440659	0.15835970	0.8552723	0.3924004

Interprete the results. *Your interpretation of the results goes here!*

Use the logistic regression models

Predict the probability that the market will go up, given values of the predictors.

In R , the `type="response"` is used to output probabilities of the form $P(Y = 1|X)$, as opposed to other information such as the logit.

In [95]:
```
glm.probs=predict(glm.fit,type="response")
glm.probs[1:10]
```

1	0.507084133395401
2	0.481467878454591
3	0.481138835214201
4	0.515222355813022
5	0.510781162691538
6	0.506956460534911
7	0.492650874187038
8	0.509229158207377
9	0.517613526170958
10	0.488837779771376

These values correspond to the probability of the market going up rather than down.

In R we see this by in voking the `contrasts()` function indicating that a dummy variable has been created with a 1 for `Up` . The probabilities must be converted to prediction labels.

In [96]:
```
contrasts(Smarket$Direction)
glm.pred=rep("Down",1250)
glm.pred[glm.probs >.5]="Up"
```

A matrix: 2 ×
1 of type dbl

	Up
Down	0
Up	1

Compute and a confusion maytrix in order to determine how many observations were correctly or incorrectly classified.

In [97]:
```
table(glm.pred,Smarket$Direction)
mean(glm.pred==Smarket$Direction)
```

```
glm.pred Down  Up
    Down  145 141
    Up    457 507
```

0.5216

Interprete the results. *Your interpretation of the results goes here!*

Recall the low p values of the predictors. Check if a subset of predictors gives better results

In [98]:
```
glm.fit=glm(Direction~Lag1+Lag2,data=Smarket ,family=binomial)
glm.probs=predict(glm.fit,type="response")
glm.pred=rep("Down",1250)
glm.pred[glm.probs >.5]="Up"
table(glm.pred,Smarket$Direction)
mean(glm.pred==Smarket$Direction)
```

```
glm.pred Down  Up
    Down  114 102
    Up    488 546
```

0.528

Interprete the results. *Your interpretation of the results goes here!*

Perform an LDA

Now perform an LDA on the `Smarket` data and analyze the result.

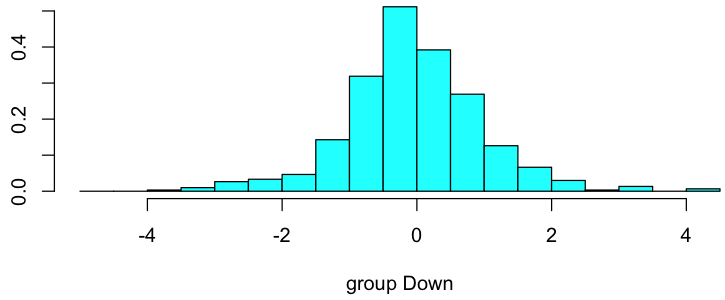
In [103]:
```
library(MASS)
lda.fit=lda(Direction~Lag1+Lag2,data=Smarket)
lda.fit
plot(lda.fit)
```

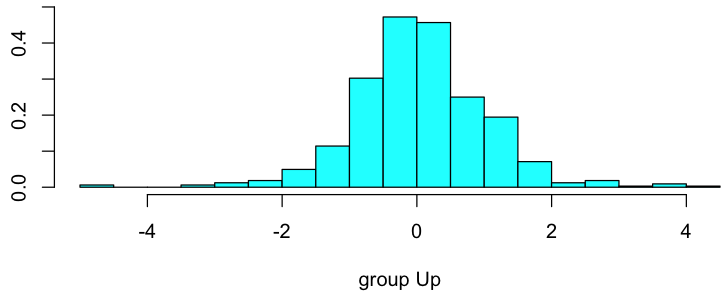
```
Call:
lda(Direction ~ Lag1 + Lag2, data = Smarket)

Prior probabilities of groups:
  Down     Up
0.4816 0.5184

Group means:
            Lag1        Lag2
Down  0.05068605  0.03229734
Up   -0.03969136 -0.02244444

Coefficients of linear discriminants:
            LD1
Lag1 -0.7567605
Lag2 -0.4707872
```





Interprete the results. *Your interpretation of the results goes here!*

Use the LDA model

Predict the `Direction` as a response for the selected predictor values using the trained LDA model.

In [110]:
```
lda.pred=predict(lda.fit)
```

Compute a confusion matrix.

In [113]:
```
table(lda.pred$class,Smarket$Direction)
mean(lda.pred$class==Smarket$Direction)
```

```
      Down  Up
 Down  114 102
 Up    488 546
```

0.528

Interprete the results. *Your interpretation of the results goes here!*

Perform a QDA

Now perform a QDA on the `Smarket` data and analyze the result.

In [116]:
```
qda.fit=qda(Direction~Lag1+Lag2,data=Smarket)
qda.fit
```

```
Call:
qda(Direction ~ Lag1 + Lag2, data = Smarket)

Prior probabilities of groups:
  Down     Up
0.4816 0.5184

Group means:
            Lag1        Lag2
Down  0.05068605  0.03229734
Up   -0.03969136 -0.02244444
```

Interprete the results. *Your interpretation of the results goes here!*

Use the QDA model

Predict the `Direction` as a response for the selected predictor values using the trained QDA model. Compute and analyze a confusion matrix.

In [118]:
```
qda.pred=predict(qda.fit)
table(qda.pred$class,Smarket$Direction)
mean(qda.pred$class==Smarket$Direction)
```

```
      Down  Up
Down   109  94
Up     493 554
```

0.5304

Interprete the results. *Your interpretation of the results goes here!*

Use K-Nearest Neighbors Clustering

Create a training data set used to fined the k nearest neighbors of a data point and their actual classes.

In [121]:
```
train=(Smarket$Year <2005)
Smarket.2005= Smarket [! train ,]
dim(Smarket.2005)
```

252 9

In [123]:
```
library(class)
train.X=cbind(Smarket$Lag1, Smarket$Lag2)[train ,]
test.X=cbind(Smarket$Lag1, Smarket$Lag2)[!train,]
train.Direction =Smarket$Direction[train]
Direction.2005=Smarket$Direction[!train]
```

Use and analyze KNN for $k = 1$.

In [125]:
```
set.seed (1)
knn.pred=knn(train.X,test.X,train.Direction ,k=1)
table(knn.pred,Direction.2005)
mean(knn.pred==Direction.2005)
```

```
        Direction.2005
knn.pred Down Up
    Down   43 58
    Up     68 83
```

0.5

Use and analyze KNN for $k = 3$.

In [126]:
```
knn.pred=knn(train.X,test.X,train.Direction ,k=3)
table(knn.pred,Direction.2005)
mean(knn.pred==Direction.2005)
```

```
        Direction.2005
knn.pred Down Up
    Down   48 54
    Up     63 87

0.535714285714286
```

Interprete the results. *Your interpretation of the results goes here!*