hw4

Tinglei Wu

2/21/2022

Exercise 4:

Part a:

- Since we know that if we want to predict the response for a test observation with X=0.6, we will use observations in the range [0.55,0.65], in this case, if x is between 0 and 1, then the observation we want to use are in the interval [x–0.05,x+0.05] which represents a length of 0.1 and a fraction of 10%. However, we would also consider a situation that if x is less than 0.05, which the observation interval becomes [0, x+0.05], because the interval cannot be negative. Another situation is that when x is greater than 0.95, so the interval would become [x-0.05, 1]. Therefore, the average fraction we will use to make the prediction is:

$$\int_{0.05}^{0.95} 10\,dx + \int_{0}^{0.05} 100x + 5\,dx + \int_{0.95}^{1} 105 - 100x\,dx = 9 + 0.375 + 0.375 = 9.75$$

Therefore, the average fraction of observations we would use for prediction is 9.75%

Part b:

- When it becomes 2 features with p = 2, we can simply calculate the fraction of observations that we would use for prediction by using

$$9.75\%^2 = 0.950625$$

.

Part c:

- When the features become 100 with p = 100, it is the same thing for us to calculate the fraction of observations that we would use for prediction except the power would become 100:

$$9.75\%^{100} \approx 0$$

.

Part d:

- As we can see from the previous questions, as the number of features increases, the fraction of observations that we would use for prediction decreases. When p becomes infinity, the fraction of observations that we would use for prediction becomes 0.

Part e:

- Since it contains 10% of the trainning observations, when p = 1, length of each side of the hypercube is 0.1. When p = 2, the length of the each side of the hypercube is

$$0.1^{1/2}$$

, when p = 100, the length of the each side of the hypercube is

$$0.1^{1/100}$$

.

Exercise 10:

Part a:

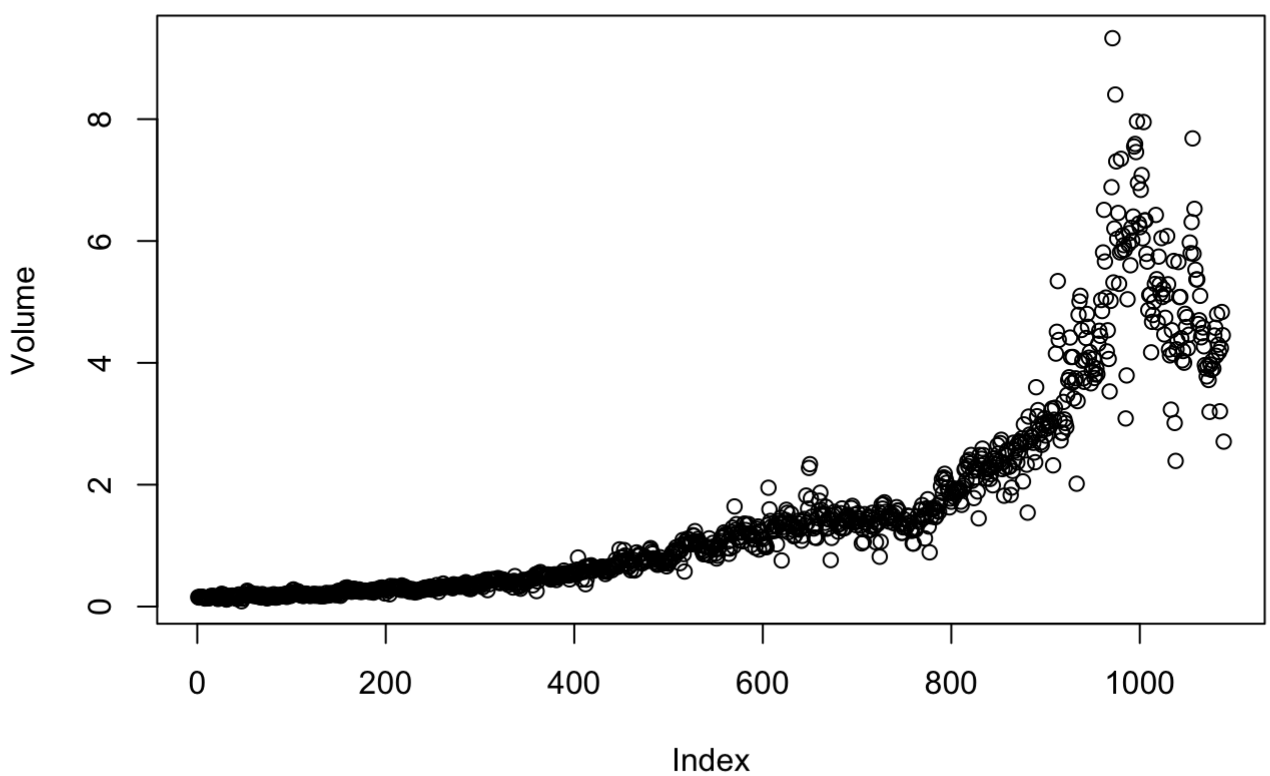
```
library(ISLR)
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
##       Lag4               Lag5              Volume             Today
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260
##  Direction
##  Down:484
##  Up  :605
##
##
##
##
```

```
cor(Weekly[, -9])
```

```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

```
attach(Weekly)
plot(Volume)
```



- The Year and Volume variables seem to have very high positive correlation between each other,0.84194162, and the graph of Volume is also increasing over time.

Part b:

```
head(Weekly)
```

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	1990	0.816	1.572	-3.936	-0.229	-3.484	0.1549760	-0.270	Down
2	1990	-0.270	0.816	1.572	-3.936	-0.229	0.1485740	-2.576	Down
3	1990	-2.576	-0.270	0.816	1.572	-3.936	0.1598375	3.514	Up
4	1990	3.514	-2.576	-0.270	0.816	1.572	0.1616300	0.712	Up
5	1990	0.712	3.514	-2.576	-0.270	0.816	0.1537280	1.178	Up
6	1990	1.178	0.712	3.514	-2.576	-0.270	0.1544440	-1.372	Down

6 rows

```
fit.glm <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Weekly, fam
ily = binomial)
summary(fit.glm)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

- From the results above, we can see that Lag2 is the only predictor that has a p-value lower than 0.05, so Lag2 is statistically significant.

Part c:

```
probs <- predict(fit.glm, type = "response")
pred.glm <- rep("Down", length(probs))
pred.glm[probs > 0.5] <- "Up"
table(pred.glm, Direction)
```

```
##         Direction
## pred.glm Down  Up
##     Down   54  48
##     Up    430 557
```

- Overall, the accuracy of the prediction is about (54+557)/1089 = 56.1%, thus the error rate of the prediction is about 43.9%.

Part d:

```
train <- (Year < 2009)
Weekly.20092010 <- Weekly[!train, ]
Direction.20092010 <- Direction[!train]
fit.glm2 <- glm(Direction ~ Lag2, data = Weekly, family = binomial, subset = train)
summary(fit.glm2)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
##     subset = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

```
probs2 <- predict(fit.glm2, Weekly.20092010, type = "response")
pred.glm2 <- rep("Down", length(probs2))
pred.glm2[probs2 > 0.5] <- "Up"
table(pred.glm2, Direction.20092010)
```

```
##          Direction.20092010
## pred.glm2 Down Up
##      Down    9  5
##      Up     34 56
```

- In this case, we only use Lag2 as the predictor to predict the Direction, and the accuracy of the prediction is (9+56)/104 = 62.5%, thus the error rate of the prediction is 37.5%.

Part e:

```
library(MASS)
fit.lda <- lda(Direction ~ Lag2, data = Weekly, subset = train)
fit.lda
```

```
## Call:
## lda(Direction ~ Lag2, data = Weekly, subset = train)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4477157 0.5522843
##
## Group means:
##             Lag2
## Down -0.03568254
## Up    0.26036581
##
## Coefficients of linear discriminants:
##            LD1
## Lag2 0.4414162
```

```
pred.lda <- predict(fit.lda, Weekly.20092010)
table(pred.lda$class, Direction.20092010)
```

```
##       Direction.20092010
##        Down Up
##   Down    9  5
##   Up     34 56
```

- Using the LDA actually gives us the same result as glm, the accuracy of the prediction is (9+56)/104 = 62.5%, thus the error rate of the prediction is 37.5%.

Part f:

```
fit.qda <- qda(Direction ~ Lag2, data = Weekly, subset = train)
fit.qda
```

```
## Call:
## qda(Direction ~ Lag2, data = Weekly, subset = train)
##
## Prior probabilities of groups:
##      Down        Up
## 0.4477157 0.5522843
##
## Group means:
##             Lag2
## Down -0.03568254
## Up    0.26036581
```

```
pred.qda <- predict(fit.qda, Weekly.20092010)
table(pred.qda$class, Direction.20092010)
```

```
##       Direction.20092010
##        Down Up
##   Down    0  0
##   Up     43 61
```

- Using the QDA gives us the accuracy of the prediction to be 61/104 = 58.65%, and the error rate of prediction is 41.35%. However, we can see that the model is only choosing Up as the answer and not even have one Down answer.

Part g:

```
library(class)
train.X <- as.matrix(Lag2[train])
test.X <- as.matrix(Lag2[!train])
train.Direction <- Direction[train]
set.seed(1)
pred.knn <- knn(train.X, test.X, train.Direction, k = 1)
table(pred.knn, Direction.20092010)
```

```
##         Direction.20092010
## pred.knn Down Up
##     Down   21 30
##     Up     22 31
```

- The accuracy of prediction using KNN with k = 1 is (21+31)/104 = 50%, and thus the error rate of the prediction is also 50%.

Part h:

- From the previous results, we can see that the logistic regression and LDA have the best performances in terms of accuracy of the prediction.

Part i:

```
# Logistic regression with Lag2:Lag4
fit.glm3 <- glm(Direction ~ Lag2:Lag4, data = Weekly, family = binomial, subset = train)
probs3 <- predict(fit.glm3, Weekly.20092010, type = "response")
pred.glm3 <- rep("Down", length(probs3))
pred.glm3[probs3 > 0.5] = "Up"
table(pred.glm3, Direction.20092010)
```

```
##          Direction.20092010
## pred.glm3 Down Up
##      Down    1  4
##      Up     42 57
```

```
mean(pred.glm3 == Direction.20092010)
```

```
## [1] 0.5576923
```

```
# LDA with Lag2 interaction with Lag3
fit.lda2 <- lda(Direction ~ Lag3:Lag1, data = Weekly, subset = train)
pred.lda2 <- predict(fit.lda2, Weekly.20092010)
mean(pred.lda2$class == Direction.20092010)
```

```
## [1] 0.5961538
```

```
# QDA with Volume
fit.qda2 <- qda(Direction ~ Lag2 + Volume, data = Weekly, subset = train)
pred.qda2 <- predict(fit.qda2, Weekly.20092010)
table(pred.qda2$class, Direction.20092010)
```

```
##       Direction.20092010
##        Down Up
##   Down   32 44
##   Up     11 17
```

```
mean(pred.qda2$class == Direction.20092010)
```

```
## [1] 0.4711538
```

```
# KNN k = 19
pred.knn2 <- knn(train.X, test.X, train.Direction, k = 19)
table(pred.knn2, Direction.20092010)
```

```
##          Direction.20092010
## pred.knn2 Down Up
##      Down   19 22
##      Up     24 39
```

```
mean(pred.knn2 == Direction.20092010)
```

```
## [1] 0.5576923
```

- After examine the combinations of predictors, the original logistic regression and LDA still have the best performaces in terms of accuracy of the prediction overall.