Tutorial 4: Classification I: Logit, LDA, KNN

ECO3080: Machine Learning in Business

Instructor: Prof. Qihui Chen
Teaching Assistant: Long Ma

September 28, 2022

- Let's look at the data set first: 1250 Obs. and 9 variables

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today	Direction
1	2001	0.381	-0.192	-2.624	-1.055	5.010	1.19130	0.959	Up
2	2001	0.959	0.381	-0.192	-2.624	-1.055	1.29650	1.032	Up
3	2001	1.032	0.959	0.381	-0.192	-2.624	1.41120	-0.623	Down
4	2001	-0.623	1.032	0.959	0.381	-0.192	1.27600	0.614	Up
5	2001	0.614	-0.623	1.032	0.959	0.381	1.20570	0.213	Up
6	2001	0.213	0.614	-0.623	1.032	0.959	1.34910	1.392	Up
7	2001	1.392	0.213	0.614	-0.623	1.032	1.44500	-0.403	Down
8	2001	-0.403	1.392	0.213	0.614	-0.623	1.40780	0.027	Up
9	2001	0.027	-0.403	1.392	0.213	0.614	1.16400	1.303	Up
10	2001	1.303	0.027	-0.403	1.392	0.213	1.23260	0.287	Up
11	2001	0.287	1.303	0.027	-0.403	1.392	1.30900	-0.498	Down
12	2001	-0.498	0.287	1.303	0.027	-0.403	1.25800	-0.189	Down
13	2001	-0.189	-0.498	0.287	1.303	0.027	1.09800	0.680	Up
14	2001	0.680	-0.189	-0.498	0.287	1.303	1.05310	0.701	Up
15	2001	0.701	0.680	-0.189	-0.498	0.287	1.14980	-0.562	Down

Data Description

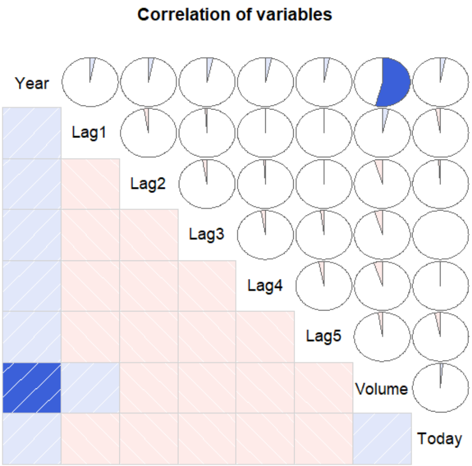
- This data set consists of percentage returns for the S&P 500 stock index over 1250 days, from the beginning of 2001 until the end of 2005.
- For each date, we have recorded the percentage returns for each of the five previous trading days, Lag1 through Lag5.
- We have also recorded Volume (the number 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).
- Our goal is to predict Direction (a qualitative response) using the other features.

Summary Statistic

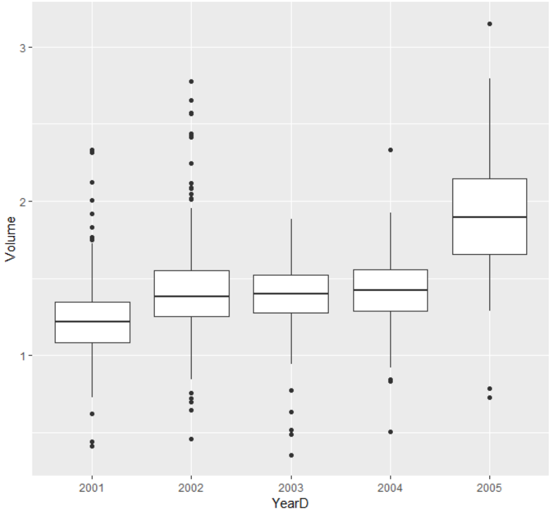
Table: Summary Statistic

Statistic	N	Mean	St. Dev.	Min	Max
Year	1,250	2,003.016	1.409	2,001	2,005
Lag1	1,250	0.004	1.136	−4.922	5.733
Lag2	1,250	0.004	1.136	−4.922	5.733
Lag3	1,250	0.002	1.139	−4.922	5.733
Lag4	1,250	0.002	1.139	−4.922	5.733
Lag5	1,250	0.006	1.148	−4.922	5.733
Volume	1,250	1.478	0.360	0.356	3.152
Today	1,250	0.003	1.136	−4.922	5.733

Summary Statistic



Summary Statistic



Sampling

- Training Set: 2001 - 2004
- Validation Set: 2005

```r
attach(Stock_Data)
train <- (Year < 2005)
Stock_Data_test <- Stock_Data[!train, ]
Stock_Data_train <- Stock_Data[train, ]
Direction_2005 <- Direction[!train]
detach(Stock_Data)
```

```
> library(kknn)
> knn <- kknn(Direction ~ Lag1 + Lag2 + Volume,
+             Stock_Data_train, Stock_Data_test, k=3)
> knn_pred <- fitted(knn)
> table(Stock_Data_test$Direction, knn_pred, dnn = c("True", "Pred."))
      Pred.
True   Down Up
  Down   64 47
  Up     79 62
```

```r
logitreg <- glm(Direction ~ Lag1 + Lag2  + Volume,
                data = Stock_Data,
                family = binomial(link = logit),
                subset = train)
summary(logitreg)
logit_probs <- predict(logitreg, Stock_Data_test, type = "response")
logit_pred <- rep("Down", 252)
logit_pred[logit_probs > 0.5] <- "Up"
table(logit_pred, Direction_2005)

probitreg <- glm(Direction ~ Lag1 + Lag2 + Volume,
                 data = Stock_Data,
                 family = binomial(link = probit),
                 subset = train)
summary(probitreg)
probit_probs <- predict(probitreg, Stock_Data_test, type = "response")
probit_pred <- rep("Down", 252)
probit_pred[probit_probs > 0.5] <- "Up"
table(probit_pred, Direction_2005)

stargazer(logitreg, probitreg, type = "html", title = "Example")
```

Logit/Probit Models

	Dependent variable:	
	Direction	
	logistic	*probit*
	(1)	(2)
Lag1	-0.054	-0.034
	(0.052)	(0.032)
Lag2	-0.046	-0.029
	(0.052)	(0.032)
Volume	-0.120	-0.075
	(0.238)	(0.149)
Constant	0.197	0.123
	(0.332)	(0.208)
Observations	998	998
Log Likelihood	-690.574	-690.575
Akaike Inf. Crit.	1,389.147	1,389.149
Note:	*p<0.1; **p<0.05; ***p<0.01	

```
           Direction_2005
logit_pred Down   Up
      Down   79  100
      Up     32   41
```

```
            Direction_2005
probit_pred Down   Up
       Down   79  100
       Up     32   41
```

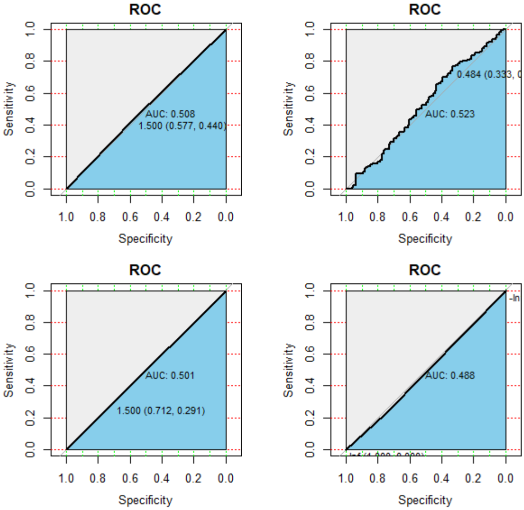
LDA/QDA

```r
library(MASS)
lda <- lda(Direction ~ Lag1 + Lag2  + Volume,
               data = Stock_Data, subset = train)
lda
plot(lda)
lda_pred <- predict(lda, Stock_Data_test)
names(lda_pred)
lda_class <- lda_pred$class
table(lda_class, Direction_2005)

qda <- qda(Direction ~ Lag1 + Lag2 + Volume,
           data = Stock_Data, subset = train)
qda
qda_pred <- predict(qda, Stock_Data_test)
table(qda_pred$class, Direction_2005)
```

```
          Direction_2005
lda_class Down  Up
     Down   79 100
     Up     32  41
```

```
      Direction_2005
       Down  Up
Down     84 110
Up       27  31
```



ROC
AUC: 0.508
1.500 (0.577, 0.440)
Sensitivity
Specificity
ROC
0.484 (0.333, 0
AUC: 0.523
Sensitivity
Specificity
ROC
AUC: 0.501
1.500 (0.712, 0.291)
Sensitivity
Specificity
ROC
AUC: 0.488
Sensitivity
Specificity

*Multi-class Logit Model

	age	blood.pressure	sex	outcome
1	69.00142	108.74175	female	fair
2	48.26654	137.89989	male	poor
3	51.02515	130.86595	female	poor
4	45.55499	126.00097	female	fair
5	56.68986	132.78573	female	fair
6	30.99371	140.14470	female	poor
7	49.40744	160.97023	female	fair
8	62.77955	83.59071	male	poor
9	67.49486	100.63270	female	fair
10	62.41583	157.18749	male	fair
11	70.06186	118.57335	female	fair
12	58.14854	171.09122	female	poor
13	50.18173	149.34822	female	good
14	60.92908	116.85020	female	good
15	59.47216	133.26287	male	good
16	59.19672	125.20865	male	fair
17	53.45896	137.21191	female	poor
18	50.49316	128.99013	male	good
19	70.19562	113.10960	female	fair
20	68.59046	102.21017	female	good

*Multi-class Logit Model

(1) Category space: $J = \{j_1, \dots, j_m\}$

(2) Probability space: $\Pi = \{\pi_1, \dots, \pi_m\}$, $\pi_1 + \pi_2 + \cdots + \pi_m = 1$

(3) Model:

$$\begin{cases} ln\left(\frac{\pi_1}{\pi_m}\right) = \beta_{1,1} + \beta_{2,1}x_2 + \beta_{3,1}x_3 + \cdots + \beta_{K,1}x_K \\ ln\left(\frac{\pi_2}{\pi_m}\right) = \beta_{1,2} + \beta_{2,2}x_2 + \beta_{3,2}x_3 + \cdots + \beta_{K,2}x_K \\ \vdots \\ ln\left(\frac{\pi_{m-1}}{\pi_m}\right) = \beta_{1,m-1} + \beta_{2,m-1}x_2 + \beta_{3,m-1}x_3 + \cdots + \beta_{K,m-1}x_K \end{cases}$$

```
> library(nnet)
> multi_logit <- multinom(outcome ~ sex + age + blood.pressure,
+                         data = data001)
# weights:  15 (8 variable)
initial  value 1098.612289
iter  10 value 1094.983885
iter  10 value 1094.983879
iter  10 value 1094.983879
final  value 1094.983879
converged
> summary(multi_logit)
Call:
multinom(formula = outcome ~ sex + age + blood.pressure, data = data001)

Coefficients:
     (Intercept)    sexmale        age blood.pressure
fair  -0.4241529 -0.1419478 0.01163071   -0.001496184
good   0.7139910 -0.2725475 0.00297585   -0.005947562

Std. Errors:
     (Intercept)   sexmale         age blood.pressure
fair   0.7921629 0.1557941 0.007830494    0.005086707
good   0.7858788 0.1560237 0.007814219    0.005084151

Residual Deviance: 2189.968
AIC: 2205.968
```

#2 Cumulative odds ratio model:

(1) Category space: $J = \{j_1, \dots, j_m\},\ j_1 < j_2 < \cdots < j_m$

(2) Probability space: $\Pi = \{\pi_1, \dots, \pi_m\},\ \pi_h \equiv \mathrm{P}(y_i \le j_h | \boldsymbol{X})$

(3) Model:

$$\begin{cases} ln\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_{1,1} + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K \\ ln\left(\frac{\pi_2}{1-\pi_2}\right) = \beta_{1,2} + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K \\ \vdots \\ ln\left(\frac{\pi_{m-1}}{1-\pi_{m-1}}\right) = \beta_{1,m-1} + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K \end{cases}$$

*Multi-class Logit Model

```
> library(MASS)
>
> # set up regression
> orderlogit <- polr(ordered(outcome) ~ sex + age + blood.pressure,
+                    data = data001)
> summary(orderlogit)

Re-fitting to get Hessian

Call:
polr(formula = ordered(outcome) ~ sex + age + blood.pressure,
    data = data001)

Coefficients:
                   Value Std. Error t value
sexmale        -0.202041   0.116722 -1.7310
age             0.002005   0.005918  0.3388
blood.pressure -0.004331   0.003777 -1.1467

Intercepts:
          Value   Std. Error t value
poor|fair -1.2312  0.5910    -2.0832
fair|good  0.1681  0.5896     0.2850

Residual Deviance: 2192.436
AIC: 2202.436
```

*Multi-class Logit Model

```
> library(brant)
> brant(orderlogit)
--------------------------------------------
Test for        X2      df      probability
--------------------------------------------
Omnibus         2.4     3       0.49
sexmale         0       1       0.96
age             2.28    1       0.13
blood.pressure  0.11    1       0.74
--------------------------------------------

H0: Parallel Regression Assumption holds
```

- Naive Bayes model: we have N observations

$$\Pr(y_i = c_m | x_1, x_2, \dots, x_k) = \frac{\Pr(y_i = c_m) \prod_{i=1}^{k} \Pr(x_i | y_i = c_m)}{\prod_{i=1}^{k} \Pr(x_i)}$$

(1) $\widehat{Pr}(y_i = c_m) = \frac{\sum I(y_i = c_m)}{N}$

(2) $\Pr(x_i | y_i = c_m) = \frac{\sum I(y_i = c_m, x_i)}{\sum I(y_i = c_m)}$

(3) $\Pr(x_i) = \frac{\sum I(x_i)}{N}$

[Question:] What is the most important assumption here? How to estimate the last two probabilities when x is continuous?

*Naive Bayes

```
install.packages("klaR")
library(klaR)
Bayes1 <- NaiveBayes(Direction ~ Lag1 + Lag2 + Volume,
                     data = Stock_Data_train)
Bayes1[1:length(Bayes1)]
par(mfrow = c(3, 1))
plot(Bayes1)

pre_Bayes1 <- predict(Bayes1, Stock_Data_test)
table(Stock_Data_test$Direction, pre_Bayes1$class)
```