

DBMS Team Project



노스트라다무스

2012011636 김원형
2012011692 민병찬
2012011829 신기한
2012011863 유누리
2013039970 이수환
2014017529 김은서
2014012015 김현우



Problem formulation



Objective

Nasdaq.csv 데이터 최종 1개월에서 1일전 close기준으로
최종 1개월 내에 high가 10% 이상이 된 적이 있는 종목을 예측

예측 모델 찾기

Training set

Dates : 2015-11-17 ~ Dates : 2016 -02 -24 : 58개 DATA

검증하기

Testing set

Dates : 2016-02-25 ~ Dates : 2016-03-18 : 30개 DATA

오늘 사용할 Method를 소개합니다

예측 Count Method

Nasdaq.csv 데이터 최종 1개월에서 1일전 close기준으로
최종 1개월 내에 high가 10%이상인 적이 있는 종목을 예측

Max margin

MAX(기준일 대비 30일의 주가 상승률)

Confidence Interval

Nasdaq.csv 데이터 최종 1개월에서 1일전 close기준으로
최종 1개월 내에 high가 10%이상인 적이 있는 종목을 예측

오늘 사용할 Attribute를 소개합니다

m <- Max margin

TABLE

STOCK [2441669,9]

	symbol	type	dates	open	high	low	close	volume	date-seq	m	sym-seq
1	AAAP	D	2015-11-17	24.46	25.51	24.38	24.62	25900	1	1.337124	1
2	AAAP	D	2015-11-18	24.62	26.31	24.06	25	111420	2	1.3168	1
3	AAAP	D	2015-11-19	24.85	26	24.71	25.9	113100	3	1.271042	1
4	AAAP	D	2015-11-20	26	27.01	25.1	25.2	60300	4	1.306349	1
5	AAAP	D	2015-11-23	25.6	25.6	25	25.15	59700	5	1.308946	1
6	AAAP	D	2015-11-24	25.21	27	25.09	25.62	56473	6	1.284934	1
7	AAAP	D	2015-11-25	26.49	26.49	25.19	25.8	11800	7	1.275969	1
8	AAAP	D	2015-11-26	25.8	25.8	25.8	25.8	0	8	1.275969	1
9	AAAP	D	2015-11-27	25.37	26.5	25.01	26.5	15400	9	1.242264	1
10	AAAP	D	2015-11-30	27.67	28.08	25.21	26.44	33100	10	1.245083	1
11	AAAP	D	2015-12-01	26.5	27.5	26.11	26.58	96200	11	1.238525	1
12	AAAP	D	2015-12-02	26.2	27.94	26	26.44	53500	12	1.245083	1

오늘 사용할 Attribute를 소개합니다

TABLE

SYMBOL	TYPE	DATES	OPEN	HIGH	LOW	CLOSE	VOLUME
AAAP	D	20151117	24.46	25.51	24.38	24.62	25900

SYMBOL	DATES	HIGH
AAAP	20151118	24.62
AAAP	20151119	24.72
AAAP	20151120	22.83
AAAP	20151121	21.50
.....		
AAAP	20151221	32.92
.....		
AAAP	20151229	30

30개 DATA

MARGIN = HIGH / CLOSE

SELECT MAX(MARGIN)



symbol	...	dates	...	m	...
1	AAAP	...	20151117	...	1.337124

KEY CODE 설명

(1) 주식종목 선택하기

```
stock <- sqldf("SELECT * FROM stock  
                WHERE SYMBOL IN  
                (SELECT SYMBOL FROM stock GROUP BY SYMBOL  
                HAVING MAX(dates)=20160318 AND COUNT(dates)>=50)")
```

조건1 : 마지막 dates의 데이터가 있어야 한다
Why? 중간에 사라진 종목을 제거 하기 위함

조건2 : 데이터 수가 50개 이상은 있어야 한다.
Why? 중간에 생긴 종목도 고려하기 위함

KEY CODE 설명

(2) Stock table, dateList만들기

```
stock$dates <-  
paste(substr(stock$dates,1,4),substr(stock$dates,5,6),substr(stock$  
dates,7,8),sep="-")  
dateList <- sqldf("select distinct dates from stock")  
dateList$date_seq <- 1:dim(dateList)[1]
```

날짜데이터에 "표시" 넣기

STOCK [2441669,9]

X	symbol	type	dates	open	high	low	close	volume
1	AAAP	D	2015-11-17	24.46	25.51	24.38	24.62	25900

dates에 sequence 부여하기

dateList [88,2]

dates	date_seq
2015-11-17	1
2015-11-18	2
2015-11-19	3
2015-11-20	4
2015-11-23	5

KEY CODE 설명

(3) Table Stock1 , dateList 만들기

```
stock1 <- sqldf ("select a.*, b.date_seq from stock a, dateList b  
                where a.dates=b.dates")  
head(stock1)  
stock2 <- stock1[,-1]  
stock2 <- sqldf("select a.*, 0 as m from stock2 a")
```

STOCK1 [2441669,10]

	X	symbol	type	dates	open	high	low	close	volume	Date_seq
1	1	AAAP	D	2015-11-17	24.46	25.51	24.38	24.62	25900	1

‘m’이라는 속성값을 만들어서
0으로저장

STOCK2 [2441669,10]

	X	symbol	type	dates	open	high	low	close	volume	Date_seq	m
1	1	AAAP	D	2015-11-17	24.46	25.51	24.38	24.62	25900	1	0

KEY CODE 설명

(4) Table dl 만들기, Stock2 의 종목 정리하기

마지막데이터가있고, 총데이터의수가50개이상있는주식의
종목만을선택한다

```
sqldf("select distinct(symbol) from stock2")  
  
dl <- sqldf("select symbol from stock2  
            group by symbol  
            having max(date_seq)-min(date_seq)+1 !=count(date_seq)")  
stock2 <- stock2[!(stock2$symbol %in% dl$symbol),]
```

dl [417,1]

symbol

AAME

AAPC

ABAC

....

자! Stock2에서dl의종목과겹치는종목을제거한다

Dim(stock2)

STOCK2 [2441669,10] ➡ STOCK2 [209557,10]

Table dl은전체date개수만큼의데이터수가없는주식종목
을저장한다

KEY CODE 설명

(5) STOCK2에 sym_seq 속성 집어넣기

Dates갯수가88개없는주식을종목과날짜를보여준다

```
sqldf("select symbol, count(dates) from stock2  
      group by symbol  
      having count(dates)!=88")
```

```
symbolList <- sqldf("select distinct symbol from stock2 order by symbol")
```

```
dim(symbolList) symbolList [2385,1]
```

```
symbolList$sym_seq <- 1:dim(symbolList)[1]
```

```
stock2 <- sqldf("select a.*, b.sym_seq from stock2 a, symbolList b  
                where a.symbol=b.symbol  
                order by symbol, dates")
```

Stock2에 sym_seq 속성을추가한다

	symbol	Count(Dates)
1	AMTD	69
2	AXSM	86
3	CCRC	87
...
26	XRDC	76

symbolList [2385,2]

symbol	Sym_seq
AAAP	1
AAL	2
AAOI	3
...	...
ZYNE	2385

KEY CODE 설명

```
> head(stock2)
```

	symbol	type	dates	open	high	low	close	volume	date_seq	m	sym_seq
1	AAAP	D	2015-11-17	24.46	25.51	24.38	24.62	25900	1	0	1
2	AAAP	D	2015-11-18	24.62	26.31	24.06	25.00	111420	2	0	1
3	AAAP	D	2015-11-19	24.85	26.00	24.71	25.90	113100	3	0	1
4	AAAP	D	2015-11-20	26.00	27.01	25.10	25.20	60300	4	0	1
5	AAAP	D	2015-11-23	25.60	25.60	25.00	25.15	59700	5	0	1
6	AAAP	D	2015-11-24	25.21	27.00	25.09	25.62	56473	6	0	1

(6) count_date 로 종목별로 데이터 개수 알아보기

```
count_date = sqldf("select symbol, count(dates) cnt from stock2  
group by symbol")  
head(count_date)
```

Count_date[2385,2]

	symbol	cnt
1	AAAP	88
2	AAL	88
3	AAOI	88
...
26	ZYNE	88

주식종목과 Dates갯수(cnt)를보여주는TABLE COUNT_date만들기

STRATEGY CODE

(7) 각 종목의 해당 date까지의 max(margin)값을 m으로 설정

```
cum_seq = 0
for (i in 1:2385){
  for (j in (cum_seq+1):(cum_seq+count_date$cnt[i]-30)){
    max_margin = 0
    for (k in (j+1):(j+30)){
      margin = stock2$high[k]/stock2$close[j]
      if (margin > max_margin){ max_margin =margin }
    }
    stock2$m[j] = max_margin
    cum_seq <- cum_seq + count_date$cnt[i] }
write.csv(stock2, file="KIHAN2.csv")
```

각 종목별로 가지고 있는 cnt 기준으로 첫날 다음날부터 마지막날에 30일을 뺀만큼의 개수만큼의 반복문이 돌아간다

j = 1일 때, k는 2:31이 되며, k in 2:31 'margin' = high[k]/close[1] 중에 max값이 m[1]에 들어간다.

STRATEGY CODE

How to make margin?

J=1(FIX)

SYMBOL	TYPE	DATES	OPEN	HIGH	LOW	CLOSE	VOLUME
AAAP	D	20151117	24.46	25.51	24.38	24.62	25900

SYMBOL	DATES	HIGH
AAAP	20151118	24.62
AAAP	20151119	24.72
AAAP	20151120	22.83
AAAP	20151121	21.50
.....		
AAAP	20151221	32.92
.....		
AAAP	20151229	30

K=2:31

MARGIN = HIGH[K] /CLOSE [1]

SELECT MAX(MARGIN)

30개 DATA



symbol	...	dates	...	m	...
1	AAAP	...	20151117	...	1.337124

STRATEGY CODE

```
stock2 <- read.csv("KIHAN2.csv")
```

```
Stock2 [209557:12]
```

X	symbol	type	dates	open	high	low	close	volume	date-seq	m	sym-seq
1	AAAP	D	2015-11-17	24.46	25.51	24.38	24.62	25900	1	1.337124	1
2	AAAP	D	2015-11-18	24.62	26.31	24.06	25	111420	2	1.3168	1
3	AAAP	D	2015-11-19	24.85	26	24.71	25.9	113100	3	1.271042	1
4	AAAP	D	2015-11-20	26	27.01	25.1	25.2	60300	4	1.306349	1
5	AAAP	D	2015-11-23	25.6	25.6	25	25.15	59700	5	1.308946	1
6	AAAP	D	2015-11-24	25.21	27	25.09	25.62	56473	6	1.284934	1
7	AAAP	D	2015-11-25	26.49	26.49	25.19	25.8	11800	7	1.275969	1
8	AAAP	D	2015-11-26	25.8	25.8	25.8	25.8	0	8	1.275969	1
9	AAAP	D	2015-11-27	25.37	26.5	25.01	26.5	15400	9	1.242264	1
10	AAAP	D	2015-11-30	27.67	28.08	25.21	26.44	33100	10	1.245083	1
11	AAAP	D	2015-12-01	26.5	27.5	26.11	26.58	96200	11	1.238525	1
12	AAAP	D	2015-12-02	26.2	27.94	26	26.44	53500	12	1.245083	1

STRATEGY CODE – Count method

Symbol별로 m 값 ≥ 1.1 의 개수를
symbol과 num 속성으로 table m1을 만든다.

```
m1 <- sqldf("select symbol, count(symbol) as num from stock2  
            where m>=1.1 group by symbol order by symbol")
```

m1[2137,2]

	symbol	num
1	AAAP	43
2	AAL	11
3	AAOI	23
...
2137	ZYNE	29

||

Symbol별로 m 값 ≥ 1.1 이 있는 것의
symbol만을 table name으로 만든다.

```
name <- sqldf("select symbol from stock2 where m>=1.1  
              group by symbol order by symbol")  
mean(m1$count(symbol))  
summary(m1$count(symbol))
```

name[2137,1]

	symbol
1	AAAP
...	...
2137	ZYNE

STRATEGY CODE – Count method

stock2.symbol= name.symbol ?

m 값 ≥ 1.10 이 있는 것의 symbol값만을 찾기 위함!

```
m2 <- sqldf("select b.symbol, count(date_seq)-30 as datenum  
            from stock2 a, name b  
            where a.symbol=b.symbol  
            group by a.symbol order by a.symbol")
```

그 symbol의 전체 date_seq에서 30을 뺀 값을 datenum으로 저장

m2[2137,2]		
	symbol	datenum
1	AAAP	58
2	AAL	58
3	AAOI	58
...
26	ZYNE	58

STRATEGY CODE – Count method

stock2.symbol= name.symbol ?

m 값 ≥ 1.10 이 있는 것의 symbol값만을 찾기 위함!

```
m2 <- sqldf("select b.symbol, count(date_seq)-30 as datenum  
            from stock2 a, name b  
            where a.symbol=b.symbol  
            group by a.symbol order by a.symbol")
```

그 symbol의 전체 date_seq에서 30을 뺀 값을 datenum으로 저장

m2[2137,2]		
	symbol	datenum
1	AAAP	58
2	AAL	58
3	AAOI	58
...
26	ZYNE	58

STRATEGY CODE – Count method

```
m3<-sqldf("select a.*, b.datenum, 0 ratio
          from m1 a, m2 b
          where a.symbol=b.symbol")
```

```
for(i in 1:2137)
  m3$ratio[i] = m3$num[i] / m3$datenum[i]
```

```
m3 <- sqldf("select a.* from m3 a order by ratio desc")
```

M3 table에 m1 table의 모든 속성과 m1.symbol에 일치하는 m2의 symbol의 Datenum 속성을 추가 그리고 ratio 속성을 모두 '0' default 값으로 추가한다

m3[2137,4]

	symbol	num	Datenum	ratio
1	AAAP	43	58	0
2	AAL	11	58	0
3	AAOI	23	58	0
...
26	ZYNE	29	58	0

Ratio[i] = num[i] /date num[i]

단순히 $m \geq 1.1$ 이상의 개수를 세는 것이 아니라, date의 개수를 고려하여 상대도수로 나타내어 정확성을 높이하고자 하였다.
(중간 시점부터 시작하는 주식이기 때문)

STRATEGY CODE—Confidence interval method

```
t1 <- sqldf("select distinct symbol, avg(m) ave, stdev(m) std, 0 YN
            from stock2
            where m > 0 group by symbol")
```

```
t1 <- sqldf("select a.*, count(date_seq)-30 cnt from t1 a, stock2 b
            where a.symbol = b.symbol
            group by a.symbol")
```

• MAX margin 에 대한
평균하고 표준편차를 구해
준다.

• 몇 개의 date 를 가지
고 있는지에 대한 정보
를 'cnt'속성을 추가시
켜 보여준다

t1 [2385 , 4]

symbol	ave	std	YN
AAAP	1.2198	0.14840	0

• 최종적으로 우리 팀에서 답이라고 정한 속
성으로, '0'이면 답이다.

t1 [2385 , 5]

symbol	ave	std	YN	cnt
AAAP	1.2198	0.14840	0	58

STRATEGY CODE –Confidence interval method

신뢰도 95%

```
z = 1.96
for (i in 1:2385){
  if (((t1$ave[i] - z*t1$std[i]/sqrt(t1$cnt[i]))>= 1.25)
      t1$YN[i] = 'O'
  else
      t1$YN[i] = 'X'
}
```

MAX margin $\sim N(m, \text{std}^2)$ 일때,
해당 MAX margin의 값이 1.25 이상이라면
YN = 'o'로 분류한다.

• 휴리스틱 enumeration으로 얻은 결과
1.25의 정확도가 가장 높아서 1.25 선택

T1

symbol	YN
AAAP	O
AAL	O
AAOI	O
AAON	O
AAPL	O



symbol	YN
AAAP	O
AAL	X
AAOI	O
AAON	O
AAPL	X

Testing 결과 – what is result of our predict?

```
answer <-  
read.csv("C:\\Users\\USER\\Documents\\answer_march.csv")  
our_answer <- sqldf("select symbol from m3 where ratio < 0.05")  
dim(our_answer)  
y_answer <- sqldf("select symbol from answer where YN == 'O'")  
compare <- sqldf("select a.symbol from y_answer a, our_answer b  
where a.symbol=b.symbol")  
dim(compare)  
compare  
1069/1420  
794/1042
```