پروژه شماره ۴ : تحلیل ممیزی و دسته بندی با داده واقعی

پاییز ۱۴۰۰

مقدمه

سرمایه‌گذاری و انباشت سرمایه در تحول اقتصادی کشور نقش بسزایی داشته است .اهمیت این عامل و نقش مؤثر آن را میتوان به وضوح درسیستم کشورهایی با نظام سرمایه‌داری مشاهده کرد. بدون شک بورس یکی از مناسب ترین جایگاه‌ها جهت جذب سرمایه‌های کوچک و استفاده از آنها در جهت رشد یک شرکت، در سطح کلان و نیز رشد شخصی فرد سرمایه‌گذار است . از آنجایی که هدف و تعریف سرمایه‌گذاری، به تعویق انداختن مصرف جهت مصرف بیشتر و بهتر در آینده است؛ افراد با سرمایه گذاری انتظار دستیابی به سود مورد انتظار خود را دارند. بنابراین مهمترین امر در این زمینه، خرید یک سهم به قیمت پایین و فروش آن به قیمت بالاتر است که این موضوع؛ به معنی پیش بینی قیمت سهام است.(عباس طلوعی و شادی)

پیش بینی قیمت بازارهای مالی مانند بازار سهام همواره یک موضوع جالب برای سرمایه گذاران و محققان بوده است. پیچیدگی بازارهای مالی مانند بی ثباتی ، بی نظمی ، نوسانات و روندهای زنجیره ای منجر به روشهای مختلفی برای پیش بینی آن می شود.

پیش بینی حرکت بازار سهام یکی از مشکلات مورد علاقه در حوزه مالی بوده است (پاگولو و همکاران ، ۲۰۱۶).

پیش بینی بازار سهام عبارت است از نتیجه گیری برای قیمت آتی سهام شرکت یا سایر ابزارهای مالی معامله شده در بورس.

سالهاست که معامله گران و تحلیلگران مالی مشغول پیش بینی رفتار بازار سهام هستند. این امر برای تحلیلگران مالی و معامله گران ضروری است زیرا به آنها اجازه می دهد استراتژی های تجاری بهتری تدوین کنند.

دو روش متداول برای پیش بینی قیمت بازار سهام وجود دارد. یکی از آن نظریه های چارتیستی یا فنی است(تحلیل تکنیکال) و دومی تحلیل ارزش اساسی یا ذاتی (تحلیل فاندامنتال)است.

در تحلیل بنیادی (فاندامنتال) از اطلاعات گسترده ای همچون: اطلاعات اقتصاد جهانی، طرح‌های توسعه، اقتصاد کلان داخلی، اطلاعات درون شرکت، تحلیل صنایع، ، و... استفاده می‌شود.

تحلیل تکنیکال در بازار های مالی روشی می باشد که کارشناسان بازار سرمایه برای پیش بینی رفتار احتمالی نمودار، از طریق داده های گذشته همچون حجم معاملات، قیمت و تغییرات آن، و ... است.

اکثر مطالعات سنتی در مورد پیش بینی بازار سهام از عوامل بنیادی کلان اقتصادی مانند تولید ناخالص داخلی و **CPI** برای پیش بینی جهت بازار استفاده می کنند (وانگ ، ۲۰۱۴ ؛ راجپوت و بابده ، ۲۰۱۶).

با پیشرفت علم کامپیوتر و هوش مصنوعی ، تنوع و پیچیدگی روش های پیش بینی بسیار بیشتر شده است.

ما در این پروژه به پیش بینی بورس با استفاده از روشهای **LDA , QDA** می‌پردازیم.

Linear Discriminant Analysis and Quadratic Discriminant Analysis

آنالیز تشخیصی خطی (به انگلیسی: **Linear Discriminant Analysis**) و تشخیص خطی فیشر روش‌های آماری هستند که از جمله در یادگیری ماشین و بازشناخت الگو برای پیدا کردن ترکیب خطی خصوصیاتی که به بهترین صورت دو یا چند کلاس از اشیا را از هم جدا می‌کند، استفاده می‌شوند.

آنالیز تشخیصی خطی بسیار به تحلیل واریانس و تحلیل رگرسیونی نزدیک است؛ در هر سه این روش‌های آماری متغیر وابسته به صورت یک ترکیب خطی از متغیرهای دیگر مدل‌سازی می‌شود. با این حال دو روش آخر متغیر وابسته را از نوع فاصله‌ای در نظر می‌گیرند در حالی که آنالیز افتراقی خطی برای متغیرهای وابسته اسمی یا رتبه‌ای به کار می‌رود. از این رو آنالیز افتراقی خطی به رگرسیون لجستیک شباهت بیشتری دارد.

آنالیز تشخیصی خطی همچنین با تحلیل مؤلفه‌های اصلی و تحلیل عاملی هم شباهت دارد؛ هر دوی این روش‌های آماری برای ترکیب خطی متغیرها به شکلی که داده را به بهترین نحو توضیح بدهد به کار می‌روند یک کاربرد عمده هر دوی این روش‌ها، کاستن تعداد بعدهای داده است. با این حال این روش‌ها تفاوت عمده‌ای با هم دارند: در آنالیز افتراقی خطی، تفاوت کلاس‌ها مدل‌سازی می‌شود در حالی که در تحلیل مؤلفه‌های اصلی تفاوت کلاس‌ها نادیده گرفته می‌شود.

LDA ارتباط نزدیکی با تحلیل واریانس و تحلیل رگرسیون دارد که سعی دارند یک متغیر مستقل را به عنوان ترکیبی خطی از ویژگی‌های دیگر بیان کنند. این متغیر مستقل در **LDA** به شکل برچسب یک کلاس است. همچنین **LDA** ارتباطی تناتنگ با تحلیل مؤلفه‌های اصلی **PCA** دارد. چرا که هر دو متد به دنبال ترکیبی خطی از متغیرهایی هستند که به بهترین نحو داده‌ها را توصیف می‌کنند. **LDA** همچنین سعی در مدلسازی تفاوت بین کلاس‌های مختلف داده‌ها دارد. از **LDA** زمانی استفاده می‌شود که اندازه‌های مشاهدات، مقادیر پیوسته باشند.

مجموعه ای از مشاهدات را به نام $\overrightarrow{x}$ برای هر نمونه از یک شی یا پدیده با کلاس شناخته شده **y** در نظر بگیرید. این مجموعه از نمونه‌ها مجموعه آموزش نامیده می‌شود

مسئله دسته‌بندی پیدا کردن یک پیش‌بینی‌کننده **(predictor)** برای هر کلاس از همان توزیع (نه لزوماً از مجموعه آموزش) داده شده از مجموعه مشاهده **x** است.

با این فرض که تابع چگالی احتمال شرطی $p(\overrightarrow{x}|y=1), p(\overrightarrow{x}|y=0)$ هر دو، توزیع نرمال با پارامترهای میانگین و کواریانس $(\mu_0, \sum_0), (\mu_1, \sum_1)$ هستند.

بدون هیچ فرض اضافه‌ای دسته بندی‌کننده حاصل به عنوان **QDA (Quadratic discriminant analysis)** شناخته می‌شود **.LDA** .علاوه براین‌ها فرض ساده‌کننده هموارایانسی یعنی برابری کواریانس کلاسها، و کواریانس‌ها رتبه کامل هستند.

مجموعه داده

داده‌های ما در بازه زمانی ۲۰۱۷.۱۰.۰۶ تا ۲۰۲۱.۱۱.۰۵ جمع آوری شده است.(۱۰۵۸ روزکاری)

متغیرهای ما خروجی ۱۸۵ اندیکاتور و شاخص کل و... میباشند که در تایم فریم روزانه جمع آوری شده‌اند.

متغیر **MOVEMENT** جهت سهام در هر روز را نمایش میدهد، اگر سهام در روز مثبت باشد متغیر **MOVEMENT** مقدار یک را اختیار میکند و اگر سهم منفی باشد متغیر **MOVEMENT** مقدار ۰ را اختیار میکند.

ما در این پروژه میخواهیم به کمک **LDA** و **QDA** به پیش بینی جهت جفت ارز **UREUSD** بپردازیم.

تجزیه و تحلیل داده

```
# فراخوانی دیتا
> bors<-read.csv("~/Processed_DJI.csv")
# حذف ستون تاریخ( چون به آن نیاز نداریم)
> bors<-bors[,-1]
# فاکتور سازی متغیر
> bors$MOVEMENT<-as.factor(bors$MOVEMENT)
# ابعاد دیتا
> dim(bors)
[1] 1059  186
# تعداد روزهای مثبت
> sum(bors$MOVEMENT==1)
[1] 533
# تعداد روزهای منفی

> sum(bors$MOVEMENT==0)
[1] 526
```

LDA

```
# پکیج مورد نیاز
> library(MASS)
# اجرای مدل LDA
> bors.lda= lda(MOVEMENT ~ .,data=bors )
# پیش بینی
> pred = predict(bors.lda,bors)
# جدول توافقی ( ماتریس درهمریختگی )
> table.lda<-table(bors$MOVEMENT,pred$class,dnn = c('Actual Group','Predicted Group')
)
> table.lda
            Predicted Group
Actual Group   0   1
           0 445  81
```

```
           1  81 452
```

برآورد دقت

```
> Accuracy.lda<-sum(diag(table.lda))/sum(table.lda)
> Accuracy.lda
[1] 0.8470255
```

نرخ خطا

```
> error.rate.lda<-1-Accuracy.lda
> error.rate.lda
[1] 0.1529745
```

طبق خروجی بالا میبینیم که توانستیم ۸۴٪ جهت سهام را درست پیش بینی کنیم.یعنی خطای ما ۱۵٪ است.

طبق جدول درهمریختگی ما مثبت بودن سهام را بهتر از منفی بودن آن پیش بینی میکردیم.

در ماتریس درهمریختگی میبینم که ما در ۱۰۵۸ روز کاری قبل ، ۱۶۲ روز را اشتباه پیش بینی کردیم.

$$\boldsymbol{Accuracy} = \frac{\mathbf{445 + 452}}{\mathbf{445 + 81 + 81 + 452}} = \frac{\mathbf{445 + 452}}{\mathbf{1058}}$$

خیلی نمیتوان به نتایج بالا اطمینان داشت چون برای آموزش مدل و پیش بینی مدل از یک مجموعه داده استفاده کردیم که این کار باعث ایراداتی میشود که باعث میشود نتایج ما قابل اطمینان نباشد.

برای رفع این موضوع ما داده‌های خورد را به دو بخش تقسیم میکنیم:

داده‌های آموزشی : طبق مطالعاتی که قبلا شده ۸۰٪ از داده‌ها را به این مجموعه اختصاص میدهیم. به کمک این داده‌ها ما مدل را اجرا میکنیم.

داده‌های تست : ۲۰٪ از دادهای باقی مانده را به این داده‌ها اختصاص میدهیم . به کمک این داده‌ها ما دقت مدل را تست میکنیم.

در بورس هر چه به تاریخ روز نزدیکتر میشویم اهمیت اطلاعات بالاتر میرود مثلا مقادیر متغیرهای ما در روز قبل از اهمیت بیشتری نسبت به مقادیر متغیرهای ما در سال قبل برخوردارند. پس ما هرچقدر که بهتر بتوانیم داده‌های جدیدتر را پیش بینی کنیم نتیجه کار ما موثرتر خواهد بود.

پس در اینجا ۸۰٪ ابتدایی داده‌ها را یعنی ۸۴۷ روزکاری را به داده‌ها آموزشی اختصاص میدهیم و بقیه در به عنوان داده تست در نظر میگیریم.

داده‌های آموزشی

```
> train.df <- bors[1:847,]
```

داده‌های تست

```
> valid.df <- bors[848:1059,]
```

پکیج مورد نیاز

```
> library(MASS)
```

```
# اجرای مدل با داده‌های آموزش
> bors.lda= lda(MOVEMENT ~ .,data=train.df )
# پیش بینی داده‌های تست
> pred = predict(bors.lda,valid.df)
#ماتریس درهمریختگی
> table.lda<-table(valid.df$MOVEMENT,pred$class,dnn = c('Actual Group','Predicted Gro
up'))
> table.lda
            Predicted Group
Actual Group  0  1
           0 86 23
           1 53 50
# دقت پیش بینی
> Accuracy.lda<-sum(diag(table.lda))/sum(table.lda)
> Accuracy.lda
[1] 0.6415094
# نرخ خطا
> error.rate.lda<-1-Accuracy.lda
> error.rate.lda
[1] 0.3584906
# دقت پیش بینی روزهای منفی
> (Accuracy.1<-86/(86+23))
[1] 0.7889908
# دقت پیش بینی روزهای مثبت

> (Accuracy.0<-50/(53+50))
[1] 0.4854369
```

میبینم که دقت مدل ما در اینجا نسبت به قبل کاهش یافت ولی این دقت بیشتر از قبل قابل اعتماد است.

با استفاده از ***LDA*** ما منفی بودن سهام را خیلی بهتر از مثبت بودن آن پیش بینی کردیم، به گونه‌ای که از ۱۱۰ روز منفی ما ۸۶ روز را درست پیش بینی کردیم ولی از ۱۰۳ روز مثبت ما فقط ۵۰ روز را درست پیش بینی کردیم ،

بطور کلی دقت مدل ما ۶۴٪ است ولی منفی بودن سهام را در روز جاری با دقتی ۷۸درصدی و مثبت بودن سهام را با دقتی ۴۸٪ پیش بینی میکند .

پس اگر بخواهیم از ***LDA*** برای طراحی استراتژی معاملاتی استفاده کنیم بهتر است که برای معامله فقط در یک جهت یعنی معامله ***SHORT*** استفاده کنیم . یا میتوان از ***LDA*** برای سیگنال خروج در معاملات ***LONG*** استفاده کرد .