به نام خدا

دانشگاه صنعتی شریف

تابستان ۱۴۰۲

گزارش پروژه سوم درس برنامه نویسی

پیشرفته – بخش اول

آرش مقدم نژاد – ۴۰۰۱۷۰۲۹۹

هدف بخش اول پروژه انجام برخی از آنالیزهای آماری و تصویری روی داده‌های مختلفی است. داده‌های انتخابی برای این بخش سهام‌های شرکت‌های Intel و AMD از ابتدای سال ۲۰۱۰ تا اواسط ۲۰۲۳ است. این داده‌های از طریق API سایت Yahoo Finance بدست آمده است.

در ادامه به بررسی عملیات‌های انجام شده روی این داده‌ها و بررسی سوالات مختلف پروژه می‌پردازیم.

- **سوال ۱:**

پس از فراخوانی کتابخانه‌های مورد نیاز برای انجام این بخش (که به مرور اضافه می‌شوند)، به سراغ سوال اول می‌رویم. در اینجا نیاز است تا داده‌های این دو سهام را از طریق ماژول yfinance دانلود کرده و به صورت فایل csv ذخیره کنیم. این کار با انتخاب بازه مورد نظر برای بررسی و وارد کردن نام سهام مورد نظر قابل انجام است. سپس برای حصول اطمینان می‌توان داده‌های دانلود شده را پرینت و مشاهده کرد.

- **سوال ۲:**

برای نمونه برداری و تطبیق داده‌ها و مدیریت داده‌های null که بر اثر تعطیلی بازار در بعضی روزها و... به وجود آمده‌اند، از resample (برای متوالی و پیوسته کردن تاریخ داده‌ها به صورت روزانه D) و interpolate (برای پر کردن مقادیر ستون‌های باقی مانده از سری داده به صورت خطی یا میانگین با استفاده از متد linear) در این دو سری داده استفاده شده است.

برای حاصل شدن اطمینان از درستی این عملیات، در مرحله بعدی تعداد سلول‌های خالی ستون‌ها را می‌شماریم.

سپس برای نرمال‌سازی و پایدارسازی داده‌ها، از روش‌های آماری موجود که در این درس و سایر دروس آموخته‌ایم استفاده می‌کنیم.

بدین منظور تابعی برای نرمال سازی سری داده‌ها می‌نویسیم و مشاهده می‌کنیم که قیمت های واحد هر سهم و حجم معاملات دیگر بر اساس دلار نیستند و نرمال‌سازی شده‌اند.

سپس برای پایدار سازی نیز از عملگر diff. استفاده می‌کنیم و سپس برای ستون close این داده‌های پایدار سازی شده، نمودار خطی قیمت–تاریخ رسم می‌کنیم.

❖ نکته مهم: از این مرحله به بعد به جز برای قسمت آماره‌های توصیفی که از داده‌های معمولی استفاده می‌کنیم، داده‌های مورد استفاده از نوع **داده‌های نرمال‌سازی** شده هستند.

• سوال ۳:

برای همه ستون‌های موجود در سری داده نرمال‌سازی شده این دو سهام، نمودار خطی قیمت–تاریخ را رسم می‌کنیم. توجه داشته باشید که مولفه عمودی این نمودار یعنی Price، قیمت هر سهم به دلار نیست چون این عدد نرمال شده و یکی عدد نسبتی و مقیاسی است. رسم این نمودار هم با ایجاد یک حلقه for برای همه ستون‌های سری داده قابل انجام است.(همچنین مشاهده می‌کنیم هیچگونه ناهنجاری و پرتی وجود ندارد)

✓ برای زیباتر شدن نمودارها هم از موارد مختلفی مانند تنظیم اندازه تصویر نمودار، تنظیم dpi برای بهتر شدن تصاویر و ... استفاده شده است.

برای بدست آوردن آماره‌های توصیفی برای هر ستون از سری داده(برای جلوگیری از شلوغی و طولانی شدن جدول، فقط سه ستون close، adj close و volume انتخاب شدند)، ابتدا آن‌ها را به صورت یک pivot table درآوردیم تا ویژه بصری آن بهتر باشد. سپس در قسمت توابع عملیاتی از آرگومان‌های این جدول، گزینه‌های میانگین، میانه، مجموع، ماکزیمم، مینیمم، انحراف معیار و واریانس را انتخاب کردیم تا به نمایش گذاشته شوند. در آخر هم برای زیبایی و وضوح بیشتر در مقایسه آمار برای داده‌ها در هر سال، در یکی از جداول از حالت پس‌زمینه gradient و در مورد بعدی از bar chart برای سلول‌های موجود جدول استفاده کردیم.

در قسمت بعدی سوال از ما خواسته شده تا روی ستون‌های این سری‌های داده تحلیل همبستگی داشته باشیم. بدین منظور با برقراری دو حلقه‌ی for پی در پی، در نظر گرفتن دو شرط با if (برای برقرار نشدن این بررسی برای یک ستون با خودش یا بین دو ستون تکراری که قبلا در حلقه بررسی شده) و در نهایت استفاده از عملگر corr. برای سری داده‌ها و پرینت گرفتن نتایج، این خواسته نیز برآورده شد.

- **سوال ٤:**

طبق خواسته‌ی قسمت اول سوال، تحلیل روند این دو سری داده با استفاده از روش moving averages و rolling().mean(). در بازه‌های ۷ و ۹۰ روزه (هفتگی و فصلی) محاسبه و برای بازه ۳۰ روزه (ماهانه) به وسیله نمودار خطی محاسبه شده است.

برای انجام بخش دوم سوال و تحلیل فصلی بودن، ایندکس‌های سال و ماه سری‌های داده را برای هر سری داده مشخص می‌کنیم تا به نوعی داده های هر ستون و ردیف را با این دو تگ‌دار کنیم.(دو ستون سال و ماه به سری‌های داده اضافه می‌شوند) سپس برای ستون‌های این دو سری داده(به جز دو ستون سال و ماه)، عملیات دسته بندی(groupby.) برای ماه و سال انجام داده و سپس برای باز شدن آن‌ها و ترتیب گرفتن از unstack. استفاده کرده و در آخر نمودار را برای ماه‌های مختلف بازه سال‌های ستون‌های مختلف رسم کردیم.

- **سوال ٥:**

برای قسمت اول این سوال ابتدا تابع نرمال‌سازی که پیش‌تر هم استفاده کرده بودیم را تعریف می‌کنیم. سپس با تعریف تابعی به نام get داده‌های مورد نظر را دانلود و پرینت می‌کنیم. مشاهده می‌کنیم که ستون volume برای سه سری زمانی اول خواسته شده پس از نرمال‌سازی به ما مقادیر ۰ می‌دهند و این میتواند باعث ایجاد مشکل موقع تشکیل PCها بشود، بنابرین این ستون را برای این سه سری حذف می‌کنیم.

در قسمت بعد با توجه به آموخته‌های درس، با ایجاد حلقه for که برای تابع get استفاده کردیم، تحلیل ابعاد برای دو بعد را برای سری داده‌های مورد نظر با به دست آوردن PC1 و PC2 شروع کرده و این مقادیر را برای هر کدام پرینت می‌کنیم.

سپس با انجام عملیات کاهش ابعاد مقدار توزیع‌های واریانس و روند کاهش آن‌ها تا ۰ را مشاهده می‌کنیم. در قدم بعد و با توجه به فایل آموزشی، نمودار Scree را برای توزیع‌های این سری‌ها و در قدم بعدی نمودار Biplot را برای آن‌ها رسم می‌کنیم. در نمودار Scree مجددا روند کاهشی این توزیع‌ها و در نمودار Biplot توزیع داده‌ها در طول PC های بدست آمده را مشاهده می‌کنیم.

توضیحاتی در مورد دو نمودار مورد استفاده که از پس از جست و جو در اینترنت بدست آمد:

- **Scree Plots:**

نمودارهای Scree برای تجسم مقدار واریانس توضیح داده شده توسط هر جز اصلی به دست آمده از PCA استفاده می شود. در یک نمودار اسکری، محور x تعداد مؤلفه‌های اصلی را نشان می‌دهد (معمولاً از بالاترین به پایین‌ترین واریانس مرتب می‌شوند)، و محور y نشان‌دهنده نسبت واریانس توضیح داده شده توسط هر مؤلفه است. نمودار معمولاً یک منحنی کاهشی را نشان می‌دهد، که در آن مؤلفه‌های اولیه مقدار قابل‌توجهی از واریانس را توضیح می‌دهند، و مؤلفه‌های بعدی به تدریج واریانس کمتری را توضیح می‌دهند. "زانو" نمودار، جایی که منحنی سطح آن پایین است، می‌تواند برای تعیین تعداد بهینه اجزای اصلی برای حفظ، متعادل کردن مقدار واریانس توضیح داده شده با پیچیدگی مدل استفاده شود.

- **Biplots:**

Biplot ها تجسم نقاط داده و ویژگی‌های اصلی را در فضایی با بعد کمتر ایجاد شده توسط PCA ترکیب می‌کنند. در یک Biplot ، نقاط داده به صورت نقاط پراکنده و ویژگی‌ها به صورت بردارهایی که از مبدا سرچشمه می‌گیرند نشان داده می‌شوند. جهت و طول بردارهای ویژگی نشان دهنده سهم و جهت تأثیر هر ویژگی بر اجزای اصلی است. موقعیت نقاط داده در رابطه با بردارهای ویژگی بینش هایی را در مورد روابط بین نقاط داده و ویژگی ها ارائه می دهد. Biplot ها امکان درک جامع از ساختار داده‌ها، خوشه‌ها و روابط را فراهم می‌کند و به تفسیر و کاوش مجموعه داده‌های پیچیده کمک می‌کند.

✓ موارد استفاده:

نمودارهای Scree برای تعیین تعداد مناسب اجزای اصلی برای حفظ در PCA استفاده می‌شود. آن‌ها به انتخاب یک زیرمجموعه کوچکتر از مؤلفه‌ها کمک می‌کنند که مقدار قابل توجهی از واریانس را به تصویر می‌کشند و ابعاد داده‌ها را کاهش می‌دهند و در عین حال اطلاعات مهم را حفظ می‌کنند.

Biplot ها نمایشی بصری از روابط بین نقاط داده و ویژگی‌ها در فضایی با ابعاد کاهش یافته ارائه می دهند. آنها به شناسایی الگوها، خوشه‌ها و نقاط پرت و همچنین درک مشارکت ویژگی‌های فردی در اجزای اصلی کمک می‌کنند.
Biplot ها اغلب در زمینه‌هایی مانند ژنتیک، علوم اجتماعی، امور مالی و تحقیقات بازار برای کاوش داده‌ها، انتخاب ویژگی و تشخیص ناهنجاری استفاده می‌شوند.