« پروژه اول - فاز ۲ »

بوت‌کمپ علم داده کوئرا

تابستان و پاییز ۱۴۰۲

مهلت ارسال پاسخ: تا ساعت ۲۳:۵۹ روز چهارشنبه ۲۶ مهر

زمان ارائه‌ی گروهی: شنبه ۲۹ مهر و یکشنبه ۳۰ مهر

آخرین ویرایش: ساعت ۱۲:۴۵ روز ۲۳ مهر

مسئله‌ی ۱: خوشه‌بندی

جهت دریافت مجموعه‌داده‌ی این بخش کلیک کنید.

در فایل مجموعه‌داده‌ی بخش نخست، اطلاعات تاریخچه‌ی ۴ کوین برتر بازار در بازه‌ی زمانی یک‌ساله در اختیارتان قرار گرفته است. اسکتر پلات این داده‌ها را با محورهای market cap و volume رسم نمایید (تقریباً ۳۶۵ نقطه به ازای هر کوین خواهید داشت).

بخش ۱)

حال الگوریتم خوشه‌بندی K-means را تنها بر حسب دو ویژگی market cap و volume با **۵ خوشه** برای این مجموعه‌داده اجرا کنید. سپس بر روی اسکتر پلات رسم‌شده مشخص کنید کدام نقاط مربوط به کدام خوشه هستند و مرکز هر خوشه را نیز رسم کنید. به انتخاب رنگ، مارکر، نام‌گذاری محورها و به‌طور کلی قابل درک بودن تصویر دقت داشته باشید.

بخش ۲)

پس از آن الگوریتم K-means را برای k هایی از ۱ تا ۱۰ اجرا کرده و با محاسبه‌ی *مجموع مجذورات درون خوشه‌ای (Within-Cluster Sum of Square)*، مقدار مناسبی برای هایپرپارامتر k انتخاب کنید. توجه کنید که بخش زیادی از نمره‌ی این بخش مربوط به نحوه‌ی انتخاب مقدار k است و چنان‌چه روش‌های تدریس‌شده و معمول پاسخگوی حل مسئله نبود، از شما به عنوان دیتا سانتیست‌های آینده انتظار می‌رود با جستجو و مطالعه‌ی بیشتر، روشی مناسب برای رفع چالش‌های احتمالی پیشنهاد دهید.

بخش ۳)

در آخرین گام از این سوال از شما می‌خواهیم با استفاده از روش DBScan داده‌ها را بر حسب دو ویژگی market cap و volume خوشه‌بندی کنید و هایپرپارامترها را به‌نحوی تغییر دهید که ۵ کلاستر بامعنا در خروجی تولید شود. اسکتر پلات داده‌ها و نحوه‌ی خوشه‌بندی آن‌ها را رسم کنید. نحوه‌ی اثرگذاری هر یک از هایپرپارامترها بر خروجی را توضیح دهید.

مسئله‌ی ۲: خوشه‌بندی سلسله‌مراتبی

جهت دریافت مجموعه‌داده‌ی این بخش کلیک کنید.

در مجموعه‌داده‌ی این سوال اطلاعات کلی‌ای از ۲۰ رمزارز با بیش‌ترین market cap در اختیارتان قرار گرفته است. ستون‌های market cap و volume بیانگر مقدار میانگین آن رمزارز در این سال هستند.

بخش ۱)

ابتدا تنها با در نظر گرفتن دو ویژگی market cap و volume، الگوریتم خوشه‌بندی سلسله‌مراتبی را اجرا کرده و دندوگرام (Dendrogram) به‌دست‌آمده را نمایش دهید. با توجه به نتیجه‌ی حاصل‌شده، چنان‌چه بخواهیم این ۲۰ رمزارز را به ۲ خوشه‌ی مجزا تقسیم کنیم، این دو خوشه را مشخص کنید و سعی کنید برای نتایج حاصل را تحلیل و تفسیر کنید.

بخش ۲)

اکنون ویژگی ProofType را به دو ویژگی قبلی اضافه کرده و این‌بار ارزها را با سه ویژگی خوشه‌بندی کنید. تقسیم‌بندی رمزارزها به ۲ خوشه مطابق این خوشه‌بندی به چه صورت خواهد بود؟ نتیجه را با حالت قبل مقایسه کرده و تفسیر کنید.

بخش ۳)

در انتها یک یا چند ویژگی که فکر می‌کنید منجر به یک خوشه‌بندی بامعنا و تفسیرپذیرتر خواهد شد را نیز در نظر گرفته و آزمایش را مجدد تکرار کنید. در این بخش می‌توانید هر نوع ویژگی مرتبطی که قابل استدلال باشد را اضافه کنید (مثل ستون Network در مجموعه‌داده‌ی فعلی).

راهنمایی: برای خوشه‌بندی و رسم دندوگرام می‌توانید از توابع dendrogram و linkage در کتابخانه‌ی scipy بهره ببرید.

مسئله‌ی ۳: پیش‌بینی

در بازار سرمایه، پیش‌بینی دقیق قیمت، به‌ویژه برای بازه‌های زمانی کوتاه، به‌دلیل رفتار تصادفی آن (شبیه به Random Walk) امری ناممکن است. حال اگر از الگوریتم‌های ویژه‌ی سری زمانی نیز بهره‌ای نبریم، حتی تخمین آن نیز بسیار دشوار خواهد بود. بنابراین هدف خود را در این بخش به‌روی **پیش‌بینی افزایش یا کاهش قیمت** برای **روز بعد** خواهیم گذاشت.

در این بخش از شما می‌خواهیم مدلی آموزش دهید که با توجه به اطلاعات دریافتی از امروز (که می‌تواند شامل اطلاعاتی از روزهای پیشین نیز باشد)، پیش‌بینی کند آیا قیمت رمزارز **مونرو (Monero یا XMR)** در روز بعد نسبت به امروز افزایش خواهد داشت یا کاهش؟ منظور ما از قیمت نیز **قیمت پایانی (Close price)** است. بنابراین مدل شما باید پیش‌بینی کند آیا قیمت پایانی روز بعد بیشتر از امروز خواهد بود یا خیر؟

جهت استخراج اطلاعات بازار مالی نظیر قیمت پایانی، بیشینه، کمینه و غیره می‌توانید از کتابخانه‌ی yfinance کمک بگیرید. به‌عنوان مثال جهت استخراج اطلاعات رمزارز مونرو با بازه‌های روزانه برای تمام روزهای موجود می‌توانید از قطعه‌کد زیر استفاده کنید:

```
import yfinance as yf
xmr = yf.Ticker("XMR-USD")
df_xmr = yf.download(tickers = "XMR-USD",
                     period = "max",
                     interval = "1d")
```

شما مجاز هستید از هر کدام از الگوریتم‌های یادگیری ماشین که تاکنون در کلاس‌های بوت‌کمپ آموخته‌اید برای مدل‌سازی استفاده کنید. توجه داشته باشید که متغیر هدف یا همان برچسب شما از مقایسه‌ی قیمت

نهایی امروز و روز بعد به دست می‌آید. به‌عنوان مثال در شکل زیر قیمت نهایی ردیف دوم (۱۰۵.۵۸۵۹۹۹) کمتر از ردیف اول (۱۲۰.۷۷۹۹۹۹) است، بنابراین برچسب شما باید نشانگر کاهش قیمت باشد (به‌عنوان مثال عدد ۱ برای افزایش و عدد ۰ برای کاهش).

Date	Open_xmr	High_xmr	Low_xmr	Close_xmr	Adj Close_xmr	Volume_xmr
2017-11-09	112.531998	123.404999	112.219002	120.779999	120.779999	86864600
2017-11-10	121.344002	121.665001	101.757004	105.585999	105.585999	84614000
2017-11-11	105.750000	127.106003	103.877998	119.615997	119.615997	107708000
2017-11-12	119.597000	133.675003	110.617996	123.856003	123.856003	144948000
2017-11-13	128.960007	136.528000	120.921997	123.402000	123.402000	116200000
...	...	...	...	...	...	...
2023-10-04	147.168442	150.702347	145.940781	150.469055	150.469055	59400400
2023-10-05	150.474197	151.328369	148.565491	149.623718	149.623718	55704972
2023-10-06	149.623337	152.669296	148.641647	151.992264	151.992264	49535004
2023-10-07	151.988235	155.247528	151.100983	155.212143	155.212143	61159796
2023-10-08	155.193466	155.708115	153.763336	155.280838	155.280838	65680976

توجه: استفاده از الگوریتمی غیر از الگوریتم‌های اصلی‌ای که در کلاس‌ها آموزش داده شده‌اند در بخش اصلی مجاز نیست. در صورت علاقه و تسلط می‌توانید از آن‌ها برای بخش امتیازی استفاده کنید. البته توجه داشته باشید که نیاز است تمام اعضای گروه نسبت به نحوه‌ی کار آن الگوریتم دانش کافی داشته باشند.

در صورت نیاز می‌توانید هر ویژگی دلخواهی را به مجموعه‌داده اضافه کنید یا آن‌ها را مهندسی کنید. البته دقت کنید که ویژگی‌های ورودی مدل منجر به نشت متغیر هدف نشود. برخی از ویژگی‌های دیگری که ممکن است در این پیش‌بینی مفید واقع شود عبارتند از:

- اطلاعات مالی رمزارزهای دیگر همچون بیت‌کوین
- اطلاعات مالی طلا، نقره، مس و غیره
- اطلاعات پول‌های رایج مثل تبدیل دلار به یورو
- شاخص‌های سهام مثل S&P 500
- اطلاعات شبکه‌ی رمزارز مثل نرخ هش، سختی، اندازه‌ی بلوک و غیره
- سیگنال‌ها و شاخص‌های تکنیکال مثل RSI، EMA، SMA و غیره
- و هر اطلاعات دیگری که به پیش‌بینی مدل شما کمک می‌کند.

به‌منظور ارزیابی مدل نهایی خود از داده‌های مربوط به تاریخ 2023-09-08 تا 2023-10-07 به‌عنوان مجموعه‌ی آزمون استفاده کنید. یعنی نمونه‌ی آزمون آخر شما شامل اطلاعات روز 2023-10-07 است و برچسب متغیر هدف طبق مقایسه با قیمت پایانی روز 2023-10-08 تعیین می‌شود.

راهنمایی: از آن‌جا که هدف مسئله، پیش‌بینی برای روز بعد است می‌توانید مدل‌سازی خود را با این فرض انجام دهید که اطلاعات روزهای پیشین می‌تواند در اختیار مدل قرار گیرد. پس به‌عنوان مثال ممکن است بخواهید به‌ازای هر کدام از نمونه‌های آزمون، یک مدل جدید بسازید که براساس اطلاعات روزهای پیش از آن تاریخ آموزش دیده و سعی در پیش‌بینی برای آن نمونه دارد. در نهایت طبق پیش‌بینی‌های انجام‌شده توسط هر مدل، معیارهای ارزیابی را محاسبه کنید.

با مقایسه‌ی پیش‌بینی مدل خود با برچسب‌های حقیقی برای این ۳۰ روز معیارهای Accuracy، Precision، Recall، F1 score و AUC را گزارش دهید. همچنین ماتریس درهم‌ریختگی (Confusion Matrix) نتایج به‌دست‌آمده را رسم کنید. نیاز است در زمان ارائه تحلیل مناسبی از نتایج به‌دست‌آمده ارائه دهید و از آن‌جا که مدل شما باید بتواند هم افزایش و هم کاهش قیمت را تا حد مناسبی پیش‌بینی کند، تمرکز اصلی شما باید در بهبود معیار F1 score باشد.

توجه: در آزمایش‌های خود و انتخاب مدل و هایپرپارامترهای آن نباید از داده‌های آزمون (Test) استفاده کنید، بلکه این کار باید با داده‌های اعتبارسنجی (Validation) انجام گیرد. تنها پس از دست‌یابی به مدل نهایی خود از مجموعه‌ی آزمون بهره ببرید.

نکته‌های کلی

- کدهای خود را خوانا و تمیز بنویسید.
- مهم‌ترین بخش این پروژه، تحلیل و تفسیر شما از شرایط مسئله و نتایج آن است. باید بتوانید برای هر کدام از انتخاب‌های خود در طول مسیر، دلیلی موجه و علمی داشته باشید. ارائه‌ی شما نیز باید بر همین محور باشد، یعنی روند حل مسئله، نتایج و تحلیل و تفسیر را ارائه دهید، نه توضیح کد.
- به نکات ذکر شده در ارتباط با نحوه‌ی ارسال فایل در صفحه‌ی پروژه در کلاس توجه فرمایید.

بخش امتیازی (بیشینه: ۲۰ نمره)

- مستندسازی غنی و مناسب در نت‌بوک‌ها (۲ نمره)
- استفاده از گیت و مشارکت فعال در آن (۲ نمره)
- استخراج و اضافه کردن ویژگی‌های مناسب و بامعنا در بخش ۳ از سوال ۲ (۲ نمره)
- استفاده از مدل‌های حرفه‌ای‌تر و دست‌یابی به نتایج بهتر با تسلط کامل اعضای گروه به الگوریتم (۷ نمره)
- ساخت داشبورد به‌کمک Power BI یا Streamlit به‌صورت تعاملی (۴ نمره)
- دست‌یابی به بهترین نتیجه روی مجموعه‌ی آزمون از نظر F1 score برای سه تیم نخست (به‌ترتیب ۵، ۴ و ۳ نمره)
- طرح مسئله‌ای جدید با توجه به داده‌های موجود و مرتبط (با تایید منتور) و دست‌یابی به نتایج قابل قبول و تفسیرپذیر (۱۰ نمره)
- ارائه‌ای جذاب با بهره‌گیری از خط داستانی و استفاده از ابزارهای مناسب ارائه همچون اسلاید (۲ نمره)

موفق باشید 🤩