



## پروژه پنجم

**هدف:** آشنایی با سری های زمانی و مهندسی ویژگی<sup>۱</sup> در بازارهای مالی.

**کد:** پیاده سازی این پروژه را به زبان پایتون انجام دهید؛ در این فعالیت مجاز به استفاده از tensorflow یا pytorch می باشید.

**گزارش:** ملاک اصلی انجام فعالیت، گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این فعالیت یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر سوال، تصاویر ورودی، تصاویر خروجی و توضیحات کامل و جامعی تهیه کنید. **تذکر:** مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیداً برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری بلامانع است، اما کپی کردن غیرمجاز است.

**راهنمایی:** در صورت نیاز میتوانید سوالات خود را در خصوص پروژه از تدریس یارهای درس، از طریق ایمیل زیر یا در گروه تلگرامی بپرسید. (لینک گروه تلگرامی در سایت کورسز در دسترس بوده و قبلاً به همه ی دانشجویان ایمیل شده است)

Email: ann.ceit.aut@gmail.com

**توجه:** برای آموزش شبکه های عمیق می توانید از منابع و بسترهای سخت افزاری برخط رایگان نظیر Google Colab یا Kaggle استفاده نمایید.

**تاخیر مجاز:** در طول ترم، مجموعاً مجاز به حداکثر ده روز تاخیر برای ارسال تمرینات هستید (بدون کسر نمره). این تاخیر را می توانید بر حسب نیاز بین تمرینات مختلف تقسیم کنید؛ اما مجموع تاخیرات تمام تمرینات شما نباید بیشتر از ده روز شود. پس از استفاده از این تاخیر مجاز، هر روز تاخیر باعث کسر ۱۰٪ نمره ی آن تمرین خواهد شد.

**ارسال:** فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID\_HW05.zip تا تاریخ ۱۴۰۲/۰۳/۰۱ ارسال نمایید.

**توجه مهم:** لطفاً برای تمام مدل های آموزش داده شده و گزارشی، نمودار خطا و دقت بر حسب epoch را رسم نموده و ماتریس درهم ریختگی و معیار F1 و accuracy را به ازای هر سه دسته مجموعه محاسبه نمایید و در گزارش خود بیاورید.

در این تمرین کار با سری های زمانی هدف بوده و فعالیت مد نظر تشخیص جهت قیمت بیت کوین<sup>۲</sup> در candle بعدی معاملاتی می باشد به این معنا که آیا قیمت close در candle بعدی نسبت به close در candle ماقبل بالاتر است یا پایین تر- مساوی (یک

<sup>۱</sup> Feature Engineering

<sup>۲</sup> BitCoin

مسئله‌ی دسته‌بندی دو کلاسه). مجموعه داده‌ای که در اختیار شما قرار داده شده است، candle های معاملاتی بیت‌کوین از تاریخ ۱۵ ماه May سال ۲۰۱۸ میلادی الی ۲۸ ماه February سال ۲۰۲۲ میلادی بصورت ساعتی می‌باشد.



برای آشنایی با ویژگی‌های open/close/high/low در مجموعه داده، تصویر مقابل آورده شده است؛ همچنین منظور از volume حجم

معاملات انجام شده است که در دو واحد دلاری و بیت‌کوینی آورده شده است. مجموعه داده را بارگذاری کنید. آنچه مسلم است صرفاً با استناد به یک candle نمی‌توان جهت candle بعدی را پیش‌بینی نمود (چرا؟) لذا بایستی یک سری از candle های قبلی را در نظر گرفت. به تعداد نمونه‌ی قبلی که برای پیش‌بینی استفاده می‌شود، پنجره لغزان<sup>۳</sup> گفته می‌شود.

۱. در مسئله‌ی سری‌های زمانی، طول پنجره لغزان چگونه تعیین می‌شود و به چه عواملی وابسته است و چه تاثیری در نتیجه دارد؟ کوتاه یا بلند بودن آن را تحلیل کنید. برای مسئله‌ی داده شده، شما طول پنجره لغزان را چگونه و چند تعیین می‌کنید؟ (۷ امتیاز)

۲. معمولاً در انتهای لایه‌های بازگشتی (RNNs) یک لایه‌ی convolution از نوع یک بعدی اعمال می‌شود. به نظر شما مقصود از این عمل چیست و چه تاثیری به همراه دارد؟ (۱۱ امتیاز)

مهندسی ویژگی در سری‌های زمانی اهمیت بسزایی دارد و میتواند تاثیر بالایی در عملکرد مدل داشته باشد. از این رو بایستی تا حد امکان تمامی جوانب مسئله را در نظر گیرید.

۳. آیا ویژگی 'date' که در مجموعه داده در اختیارتان قرار دارد، ویژگی مهمی است یا میتوان از آن صرف نظر کرد؟ پاسخ خود را با دلایل و استدلال‌های کافی تقویت کنید. در صورتی استفاده چگونگی آن را توضیح دهید. (۴ امتیاز)

در بازارهای مالی، استفاده از ابزارهای تحلیل تکنیکال که پایه‌ی آماری دارند، مهم تلقی می‌شوند، چرا که از اطلاعات خام گذشته میتوان ویژگی مفید استخراج نمود و صرفاً متکی بر ۵ ویژگی پایه نبود.

۴. یکی از ویژگی‌هایی که از آن در تحلیل تکنیکال استفاده می‌شود، میانگین متحرک یا moving average نام دارد که برای هر candle، میانگین قیمت تعدادی از candle های قبلی را به عنوان ویژگی در نظر می‌گیرد. شما با پیاده سازی مقتضی، این ویژگی را به مجموعه داده اضافه نمایید. پیشنهاد می‌شود بطور مشابه، تعداد ویژگی آماری بیشتری وارد مسئله کنید؛ مانند کمینه قیمت، بیشینه قیمت، انحراف معیار قیمت‌ها، میانگین/انحراف معیار حجم، ضرب حجم در قیمت و انواع اندیکاتورهای مبتنی بر Trending, Momentum, Volatility و Volume. در این قسمت (مهندسی ویژگی) از چه تعداد از candle های قبلی استفاده می‌کنید؟ (۲۵ امتیاز)

۵. پس از اعمال مهندسی ویژگی، مجموعه داده بارگذاری شده را به سه دسته آموزش، اعتبار سنجی و آزمون به گونه‌ای تقسیم کنید که ۶ ماه انتهایی مجموعه داده آزمون، ۴ ماه قبل از آن مجموعه داده‌ی اعتبار سنجی و مابقی مجموعه داده‌ی آموزشی باشد. سپس بر اساس پاسخ‌های خود در قسمت‌های فوق، نمونه‌های آموزشی و برچسب‌های متناظر را بدست آورید؛ تعداد ویژگی نهایی و عناوین

<sup>3</sup> sliding window

هر یک را در گزارش بیاورید. به نظر شما میتوان مقادیر را normal نمود؟ اگر پاسخ تان مثبت است، این مورد را نیز انجام دهید.  
آیا میتوان نمونه‌های داخل هر مجموعه را shuffle نمود؟ چرا؟ (۱۶ امتیاز)

۶. با سعی و خطا و بدست آوردن ابرپارامتر<sup>۴</sup> های مورد نیاز و معماری مطلوب، یک شبکه‌ی بازگشتی (نظیر GRU, LSTM, Elman یا ترکیبی از آنان) را بروی مجموعه داده آموزش دهید و گزارش های لازم را ارائه نمایید. (۳۷ امتیاز)

برای بدست آوردن بهترین مدل، میتوانید اعمال convolution 1D بین/آخر لایه های بازگشتی یا افزودن یک لایه‌ی Feed-Forward به انتهای شبکه برای دسته‌بندی را آزمایش کنید. همچنین در قسمت فوق، در صورتی که معیار F1 بالاتر از ۷۰٪ برای داده‌های آزمون کسب کنید و تعداد پارامتر های قابل آموزش شبکه کمتر از ۵۰۰ هزار باشد، به ازای هر ۱ درصد، یک امتیاز بیشتری به شما تعلق خواهد گرفت که بایستی در روز تحویل امکان آزمایش مجدد وجود داشته باشد، لذا مدل را ذخیره نمایید.

برخی اندیکاتور های تحلیلی پیچیده‌تری نیز وجود دارد که کمک شایانی در تحلیل و در نتیجه استخراج ویژگی می‌کند؛ سه مورد از موارد مذکور (MACD) Moving Average Convergence/Divergence، Bollinger Bands و Relative Strength Index (RSI) نام دارد. به عنوان فعالیت اضافی میتوانید ضمن توضیح مختصری از کارکرد سه اندیکاتور فوق در بازارهای مالی، پیاده‌سازی های مورد نیاز در بخش مهندسی ویژگی را انجام دهید و نتیجه را با حالت قبل مقایسه و توجیه کنید. (هر یک حداکثر ۱۰ الی ۱۵ امتیاز اضافی)

موفق باشید

---

<sup>4</sup> Hyper-Parameters