حمید پورمحمد
۹۸۲۰۲۹۳۸

امتحان پایان ترم درس یادگیری ماشینی در فیزیک
(قسمت دوم)

تابستان ۱۴۰۰

آلودگی هوا

داده

وقتی در مورد آلوده بودن یا نبودن هوا صحبت می‌شود، معمولاً منظور شاخص‌هایی همچون $pm2.5$, $pm10$, O_3, NO_2, SO_2, CO_2 مدنظر خواهند بود که در این میان، $pm2.5$ از بقیه مهم‌تر است. از این رو، ابتدا با مراجعه به سایت ((پلتفرم داده‌ی تاریخی کیفیت هوا[1]))، داده‌های روزانه‌ی مربوط به این شاخص‌های آلودگی را دریافت می‌کنیم. بنده با مراجعه به این سایت، اطلاعات تاریخی مربوط به پنج ایستگاه[2] کنترل کیفیت هوا را دریافت نمودم؛ این پنج داده، از جهات مختلفی همچون تعداد شاخص‌ها، روزهایی که اندازه‌گیری انجام شده، تاریخ شروع فعالیت ایستگاه، و... با یکدیگر تفاوت داشتند. بنابراین با تعریف چند تابع، توانستم این پنج فایل را با یکدیگر متحد سازم (مجموعاً شش ویژگی مختلف[3]). همچنین یک ستون جدید به داده‌ها اضافه نمودم که شماره‌ی روز مدنظر $(6-0)$ در هفته را نشان می‌دهد؛ مثلاً شنبه، 0 است؛ یکشنبه، 1 است، و الی آخر. زیرا من معتقدم که روزِ مدنظر، در پیش‌بینی آلودگی تاثیر خواهد گذاشت (قطعاً شرایط ابتدای هفته، با انتهای هفته متفاوت خواهد بود). علاوه بر این اطلاعات، داده‌های دیگری را نیز به در نظر گرفتم، که در زیر به آن‌ها اشاره می‌کنم.

با مراجعه به سایت $meteostat$[4]، داده‌های هواشناسی روزانه‌ی مربوط به فرودگاه بین‌المللی تهران (امام) را دریافت نمودم؛ این اطلاعات شامل سرعت متوسط باد (m/s)، جهت وزش باد (بر حسب درجه؛ $360-0$)، فشار هوا (hPa)، دمای متوسط روزانه (درجه‌ی سانتی گراد)، دمای کمینه‌ی روزانه، و دمای بیشینه‌ی روزانه بودند (شش ویژگی مختلف). پس در مجموع توانستم 13 ویژگی مختلف را به ازای هر روز جمع آوری کنم[5]؛ از سال 2015 تاکنون، که شامل داده‌های 2323 روز می‌شدند. این داده‌ها در شش فایل (پنج فایل برای آلودگی هوا، و یک فایل برای اطلاعات هواشناسی جوی) قرار دارند. پیش از شروع فرآیند آموزش، این شش فایل را متحد خواهم ساخت.

سپس باید داده‌های خود را مرتب نماییم. ابتدا همه‌ی داده‌ها را مطابق $MinMaxScaler$ بهنجار می‌کنیم. سپس ستون Y را می‌سازیم؛ این ستون، مقدار شاخص $pm2.5$ در روز n اُم را نشان می‌دهد. حال، می‌خواهیم به ازای $n_{ts}=6$ و $n_f=13$ ، داده‌ای ورودی را با ساختار (n_s, n_{ts}, n_f) تهیه نماییم. توجه گردد که ما قصد داریم با در دست داشتن داده‌های مربوط به شش روز قبلی (یعنی $n-1, n-2, ..., n-6$)، مقدار شاخص $pm2.5$ در روز n اُم را پیش‌بینی کنیم.

مدل

با توجه به این که صحبتی در مورد $transfer\ learning$ نشده است، فرض خود را بر این می‌گذارم که اجازه استفاده از $transfer\ learning$ را نداریم. از سویی دیگر، یک راهکار بسیار مناسب برای انجام پیش‌بینی آلودگی هوا، استفاده از لایه‌ی $LSTM$ خواهد بود. دو دلیل برای این انتخاب وجود دارد؛ اول اینکه $LSTM$ می‌تواند هم تاریخچه‌ی کوتاه مدت و هم تاریخچه‌ی بلند مدتِ مدل را برای پیش‌بینی‌هایش مدنظر قرار دهد؛ دوم اینکه $LSTM$

[1] https://aqicn.org/data-platform/register/

[2] sharif_university, tehran_university, razi, golbarg, and shad_abad.

[3] Feature

[4] https://meteostat.net/en/station/40730?t=2019-01-01/2019-12-31

[5] به ازای داده‌های ثبت نشده، از میانگین ستون مدنظر استفاده نمودم.

علاوه بر بررسی سری‌های زمانی، می‌تواند به عوامل دیگرِ دخیل در این پیش‌بینی‌ها نیز بپردازد (مثل دما و...). به همین جهت، انتخاب اول من استفاده از شبکه‌ای با لایه‌ی $LSTM$ است. بنابراین به تعریف یک شبکه عصبی با مشخصات $n_{ts} = 6$ و $n_f = 13$، و یک خروجی، با ساختاری ($Sequential$) به صورت زیر خواهد بود:

ورودی: که 13 نوع ویژگی مختلف در 6 زمان متوالی را به ازای هر داده، دریافت می‌کند؛

لایه $LSTM$: دارای 128 واحد[6] و $input\_shape = (6, 13)$؛

لایه دراپ‌اوت: دراپ‌اوت با احتمال 0.1؛

لایه $Dense$: با تابع فعال سازی $relu$، و دارای 32 نود، به همراه رگولاریزیشن $l2$ با ضریب تاثیر 0.01؛

لایه خروجی: از نوع $Dense$ تابع فعالسازی $linear$، و دارای یک نود.

همچنین تابع خطا را $l2$، روش بهینه سازی را SGD (با $epochs = 20$ و $batch\_size = 5$) و نرخ یادگیری را 0.05 قلمداد می‌کنیم (اگرچه می‌توان از $adam$ هم استفاده کرد)؛ در مجموع، 76,865 پارامتر خواهیم داشت که باید آن‌ها را آموزش دهیم. همچنین 100 داده‌ی آخر (مربوط به زمان اخیر) را به منظور بررسی کارایی مدل نگه خواهیم داشت.

من با بررسی حالت‌های مختلف توانستم مقادیر مناسبی را برای هایپر پارامترها بیابم. به طوری که هم زمان نسبتاً کوتاهی برای آموزش نیاز باشد، و هم دقت قابل توجهی به دست آید. همچنین یک تابع به نام $score$ تعریف نمودم تا به پیش‌بینی‌های مدل مورد نظر (خواه برای داده‌های آموزشی، خواه برای داده‌های مربوط به آینده) یک امتیاز نسبت دهد. در نهایت مدلی به دست آمد (مدلِ ذکر شده در بالا) که دقت بسیار بالایی را از خود نشان می‌دهد؛ معمولاً بالای 98 %. برای مشاهده‌ی کارایی این مدل، کافی است که به نمودار پیش‌بینی 100 روزه نگاه کنید؛ وقتی خودم این نتیجه‌ها را دیدم، لذت بردم!

[6] Unit