بسم الله الرحمن الرحیم

امتحان پایان ترم درس یادگیری ماشین – بخش دوم

ریحانه قنبری ۹۶۱۰۰۹۷۴

آلودگی هوا

چگونه مدلی درست می‌کنید که بتواند میزان آلودگی هوا در تهران را پیش بینی کند؟

مقدمه:

همان طور که در بخش اول امتحان نیز توضیح دادم، آلودگی هوا به وسیله‌ی شاخص‌هایی از میزان آلاینده‌ها سنجیده می‌شود. این آلاینده‌ها عبارت اند از NO2, SO2, PM10 , PM2.5, Co ,O3 . در شهر تهران ایستگاه‌هایی در نقاط مختلف شهر جهت پایش ساعتی و روزانه این آلاینده‌ها وجود دارد و مجموعه‌ای از این داده‌ها برای چندسال اخیر به شکل آزاد در دسترس است. در نهایت شاخص AQI(Air Quality index) که میزان آلودگی و خطر ناشی از آلودگی را نتیجه می‌دهد، براساس بیشینه شاخص این آلاینده‌ها سنجیده می‌شود.

$$AQI = \max(AQI_{PM2.5}, AQI_{PM10}, AQI_{O3}, \dots)$$

همچنین باتوجه به میزان AQI گزارش شده، می‌توان میزان آلودگی را باتوجه به جدول زیر در ۶ کلاس مختلف طبقه بندی کرد.

شاخص کیفیت هوا	سطح اهمیت بهداشتی
۰-۵۰	سالم
۵۱-۱۰۰	قابل قبول
۱۰۱-۱۵۰	ناسالم برای گروه‌های حساس
۱۵۱-۲۰۰	ناسالم
۲۰۱-۳۰۰	بسیار ناسالم
۳۰۱-۵۰۰	خطرناک

جدول ۱

تعریف مسئله:

باتوجه به مقدمه‌ای که بیان شد، برای حل این مسئله به دریافت داده‌های مربوط به شاخص آلاینده‌ها نیازمندیم. همچنین به دو شکل regression و classification می‌توان این مسئله را حل کرد. در مسئله‌ی regression هدف پیش‌بینی عدد AQI براساس داده‌های موجود است و در مسئله‌ی classification باتوجه به جدول ۱ مقدار AQI به ۶ کلاس مختلف دسته بندی خواهد شد.

دریافت داده‌ها:

سایت https://aqicn.org داده‌های مربوط به شاخص آلایندگی هوا را برای کشورهای مختلف در اختیار قرار می‌دهد. این داده‌ها برای شهر تهران نیز در دسترس بود. همانطور که گفته شد ایستگاه‌های مختلفی در شهر تهران به دریافت داده‌های مربوط به آلایندگی می‌پردازند به همین جهت من داده‌های مربوط به ایستگاه‌های مختلف را از این سایت دریافت کردم.

آماده سازی داده‌ها:

داده‌های دریافت شده مربوط به ۱۵ ایستگاه مختلف در شهر تهران بود که شامل زمان پایش داده به شکل (روز/ماه/سال) و شاخص آلاینده‌هایی بود که هر ایستگاه آن را پایش می‌کند در نتیجه هدف ساخت یک داده واحد از این ۱۵ ایستگاه مختلف بود اما چالش‌هایی برای تولید این داده وجود داشت:

- **عدم تطابق کامل زمانی برای داده‌های ایستگاه‌های مختلف بود.**

از میان ۱۵ ایستگاهی که داده‌هایشان دریافت شد ۲ ایستگاه اطلاعات ماه‌های اخیر را نداشت به همین علت این دو مجموعه داده را کنار گذاشتم و با ۱۳ داده که همه تا تاریخ ۲۰۲۱/۶/۲۳ اطلاعات را در اختیار قرار می‌دادند کار کردم. نقطه زمانی شروع داده‌ها را نیز ۲۰۱۸/۹/۱ در نظر گرفتم که بین همه‌ی داده‌ها مشترک بود. زیرا برخی ایستگاه‌ها داده‌های پیش از این را نداشتند.

- **از نقطه‌ی زمانی شروع تا پایانی که در نظر گرفتم ۱۰۲۷ روز خواهیم داشت. اما از برخی از روزها داده‌ای در دسترس نبود.**

- **تمام ایستگاه‌ها اطلاعات مربوط به ۶ آلاینده‌ی مختلف را در اختیار قرار نمی‌دهند زیرا باتوجه به دستگاه‌های سنجش به کار رفته در این ایستگاه‌ها امکان پایش برخی داده‌ها نیست. درنتیجه برخی خانه‌ها نیز به همین جهت خالی بودند.**

برای حل چالش‌های گفته شده ابتدا به هر مجموعه داده ستون‌هایی از آلاینده‌هایی که دراختیار قرار نمی‌داد را اضافه کردم و مقدار صفر را به این خانه‌ها نسبت دادم.
سپس یک Data frame جدید ساختم که تمام ۱۰۲۷ روز از ۲۰۱۸/۹/۱ تا ۲۰۲۱/۶/۲۳ را پوشش می‌داد و این تاریخ‌ها را در یک لیست ذخیره کردم. از طرفی تاریخ‌های مربوط به ۱۳ مجموعه داده‌ای که در اختیار داشتم را نیز در لیست‌های جدا ذخیره کردم سپس تاریخ‌های مربوط به ۱۳ دسته را با تاریخ لیست ساخته شده مقایسه کردم و داده‌های موجود برای آلاینده‌های هر تاریخ را در Data frame جدید متناظر با تاریخ مربوطه قرار دادم. به این شکل در Data frame جدید برای هر آلاینده در هر روز مجموعی از داده‌های آلاینده‌های ایستگاه‌های مختلف وجود داشت. سپس میانگین را برای هر آلاینده در هر روز محاسبه کردم. نکته قابل توجه در این قسمت وجود مقدار صفر برای برخی آلاینده‌ها بود که همانطور که گفتم چون این داده ها در اختیار نبود به آن ها مقدار ۰ را نسبت دادم. این مقدار ۰ می‌توانست باعث شود میانگین کمتر شود. (درواقع عدم اطلاع از این داده نباید با مقدار ۰

ای که به آن نسبت داده شده یکسان تعبیر شود چون مقدار ۰ به معنای نبود آن آلاینده است و با عدم اطلاع از آن آلاینده متفاوت است.) به همین منظور تعداد داده‌های غیر صفر را نیز برای هر بار جمع کردن ذخیره کردم تا میانگین به دلیل داده‌هایی که در اختیار نبودند کمتر از میزان واقعی نباشد و تقسیم انجام شده برای میانگین گیری به تعداد داده‌های غیر صفر باشد.

در مرحله‌ی بعد خانه‌های خالی موجود را با مقدار مد برای هر ستون پر کردم.

همانطور که می‌دانیم برای اعمال مدل‌ها نیاز است که داده‌ها تماما به شکل اعداد باشند به همین جهت نیاز بود داده‌ی مربوط به تاریخ که به شکل روز/ماه/سال بود نیز به داده‌ی عددی تبدیل شود که به همین منظور ستون تاریخ به سه ستون مختلف روز، ماه و سال تبدیل شد.

به این ترتیب مجموعه X داده ها با ۹ ویژگی شامل ۶ ستون آلاینده‌های O3, NO2, SO2, PM1, PM2.5, Co و ۳ ستون مربوط به روز و ماه و سال پر شد.

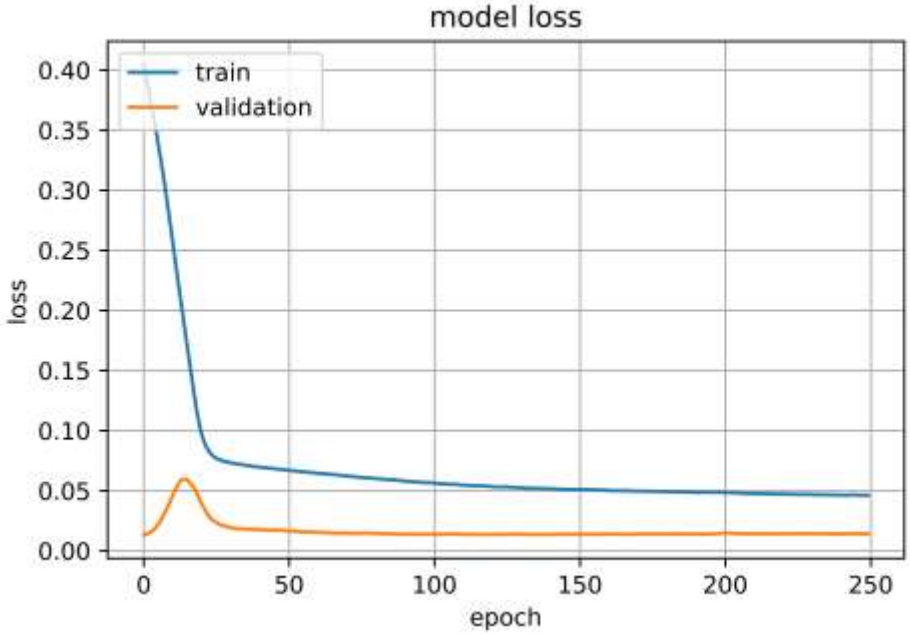
مرحله‌ی بعد ساخت داده‌های Y مورد نیاز بود که این داده‌ها بیشینه AQI آلاینده در هر ردیف بود. همچنین همانطور که گفته شد این مسئله به شکل classification نیز قابل حل است. به همین منظور داده‌های شاخص AQI به دست آمده را باتوجه به جدول ۱، به ۶ دسته طبقه بندی کردم و اعداد ۰ تا ۵ را به ان نسبت دادم.

Classification:

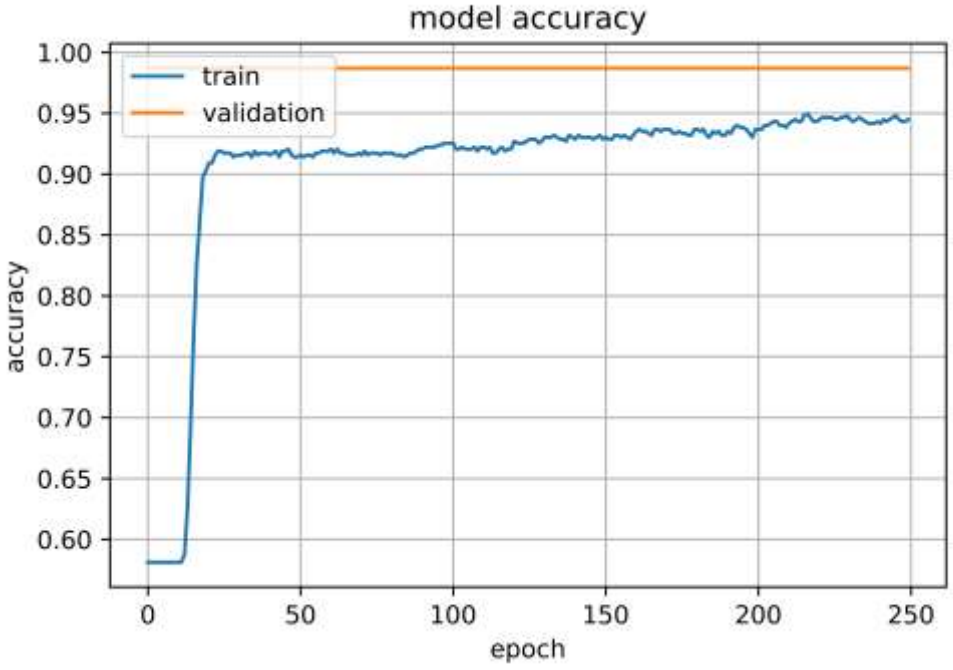
از مدل random forest برای پیش بینی کلاس‌ها استفاده کردم که score این پیش بینی ۰,۹۲ شد. (کد برنامه نویسی این مدل در فایل آمده است.)

Neural network:

همانطور که می‌دانیم شبکه‌ی عصبی برای داده‌های کم مناسب نیست. با توجه به کم بودن داده‌ها، شبکه‌ی عصبی‌ای با یک لایه ورودی، یک لایه خروجی و یک لایه میانی ساختم که تابع فعالسازی این لایه‌ها تابع relu انتخاب شد و accuracy به عنوان متریک درنظر گرفته شد. نهایتا نمودارهای تغییرات loss و metric برای داده‌های test وtrain به شکل زیر شد. که همانطور که انتظار داریم نمودار loss نزولی است و نمودار accuracy صعودی است که نشان می‌دهد مدل به خوبی کار کرده.



شکل ۱: تغییرات loss برحسب epoch های مختلف



شکل ۲: تغییرات accuracy برحسب epoch های مختلف