به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین دوم

پرسش ۱	نام دستیار طراح	امین شیخ‌زاده
	رایانامه	Aminsheykh@yahoo.com
پرسش ۲	نام دستیار طراح	محمد سپهری
	رایانامه	msepehri898@gmail.com
	مهلت ارسال پاسخ	۱۴۰۱/۱۰/۰۱

فهرست

قوانین

قبل از پاسخ دادن به پرسش‌ها، موارد زیر را با دقت مطالعه نمایید:

- از پاسخ‌های خود یک گزارش در قالبی که در صفحه‌ی درس در سامانه‌ی Elearn با نام ***REPORTS_TEMPLATE.docx*** قرار داده شده تهیه نمایید.
- پیشنهاد می‌شود تمرین‌ها را در قالب گروه‌های دو نفره انجام دهید. (بیش از دو نفر مجاز نیست و تحویل تک نفره نیز نمره‌ی اضافی ندارد) توجه نمایید الزامی در یکسان ماندن اعضای گروه تا انتهای ترم وجود ندارد. (یعنی، می‌توانید تمرین اول را با شخص A و تمرین دوم را با شخص B و ... انجام دهید)
- کیفیت گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است؛ بنابراین، لطفا تمامی نکات و فرض‌هایی را که در پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید در گزارش ذکر کنید.
- در گزارش خود مطابق با آنچه در قالب نمونه قرار داده شده، برای شکل‌ها زیرنویس و برای جدول‌ها بالانویس در نظر بگیرید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، اما باید نتایج بدست آمده از آن را گزارش و تحلیل کنید.
- تحلیل نتایج الزامی می‌باشد، حتی اگر در صورت پرسش اشاره‌ای به آن نشده باشد.
- دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند؛ بنابراین، هرگونه نتیجه و یا تحلیلی که در صورت پرسش از شما خواسته شده را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- در صورت مشاهدهٔ تقلب امتیاز تمامی افراد شرکت‌کننده در آن، ۱۰۰- لحاظ می‌شود.
- تنها زبان برنامه نویسی مجاز **Python** است.
- استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست.
- نحوه محاسبه تاخیر به این شکل است: پس از پایان رسیدن مهلت ارسال گزارش، حداکثر تا یک هفته امکان ارسال با تاخیر (به ازای هر روز ۵ درصد کسر نمره) وجود دارد، پس از این یک هفته نمره آن تکلیف برای شما صفر خواهد شد.
- لطفا گزارش، کدها و سایر ضمایم را به در یک پوشه با نام زیر قرار داده و آن را فشرده سازید، سپس در سامانه‌ی Elearn بارگذاری نمایید:

HW[Number] _[Lastname]_[StudentNumber]_[Lastname]_[StudentNumber].zip

(مثال: HW1_Ahmadi_810199101_Bagheri_810199102.zip)

- برای گروه‌های دو نفره، بارگذاری تمرین از جانب یکی از اعضا کافی است ولی پیشنهاد می‌شود هر دو نفر بارگذاری نمایند.

پرسش ۱- تخمین آلودگی هوا

آلودگی هوا یکی از معضلات بشر در قرن جدید است که عواقب جبران نشدنی برای انسان و محیط زیست به جای می‌گذارد. یکی از نخستین مراحل مقابله با آلودگی هوا، پیش‌بینی آن در سطح شهر می‌باشد.

در این سوال قصد داریم تا مقاله Air-pollution prediction in smart city, deep learning approach را شبیه سازی نماییم.

در مقاله مذکور بر اساس ۱۲ سایت اندازه‌گیری آلاینده های هوا واقع در شهر Beijing چین قصد داریم تا آلاینده‌ی $PM_{2.5}$ یکی از سایت‌ها (Aotizhongxin) را تخمین بزنیم.

۱-۱- سوالات تشریحی

متدهای زیر را مختصراً شرح دهید:

- **Linear interpolation method**
- **Pearson correlation**
- **R^2**

۱-۲- دیتاست

دیتاست مورد استفاده قرار گرفته در این مقاله در اختیار شما قرار گذاشته شده است. این دیتاست حاوی اطلاعات هر ساعت از ۱۲ سایت اندازه‌گیری آلاینده های هوا واقع در شهر Beijing چین می‌باشد. ابتدا توسط کتابخانه Pandas تمامی فایل‌های excel را فرا خوانید.

۱-۳- پیش‌پردازش

توجه داشته باشد که تنها کافیست پیش‌پردازش‌های ذکر شده را برای تمامی ستون‌های دیتاست سایت Aotizhongxin همچنین فقط برای ستون‌های $PM_{2.5}$ از باقی سایت‌ها انجام دهید.

۱-۳-۱- Missing value

همانطور که در متن مقاله آمده است یکی از روش‌های مرسوم جایگذاری مقادیر گمشده، جایگذاری آنان با مقادیر میانگین و میانه می‌باشد ولی برای داده‌های سری زمانی باید از روش دیگری همانند Linear interpolation method استفاده کنیم. در این بخش با استفاده از روش Linear interpolation method داده‌های گمشده را جایگذاری نمایید.

۱-۳-۲- Encoding Categorical Variable

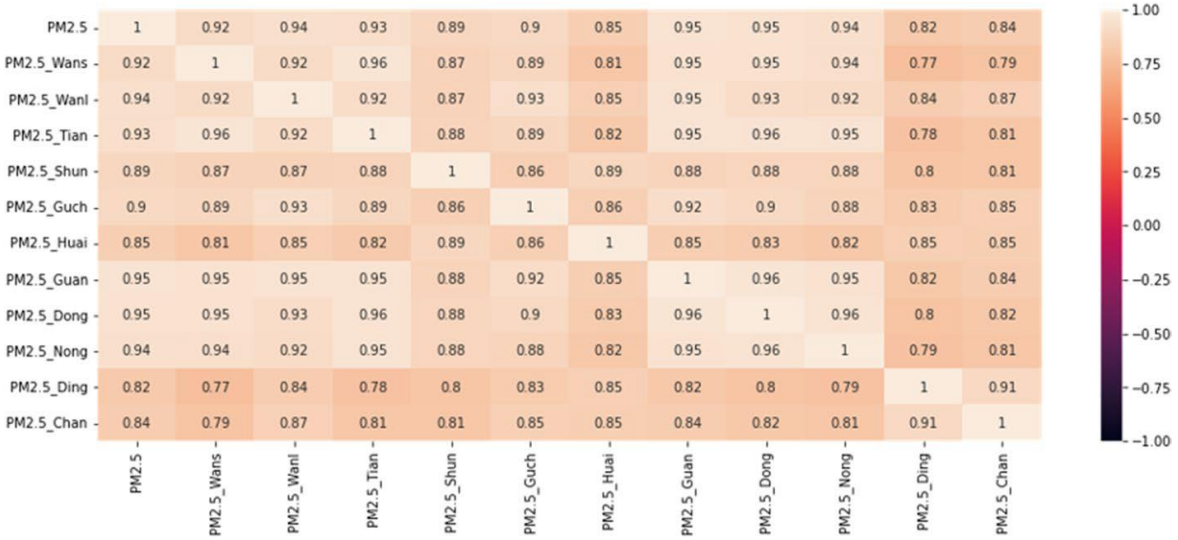
همانند مقاله ستون wd (wind direction) را به درجه تبدیل نمایید.

۱-۳-۳- Nomarlization

داده‌ها را از طریق روش Min-Max normalization نرمال کنید.

۱-۳-۴- Pearson Correlation

مقادیر $PM_{2.5}$ مربوط به ایستگاه Aotizhongxin و سایت اطراف (باقی سایت‌ها) را همانند شکل ۱ گزارش نمایید.



شکل ۱: **Pearson Correlation**

۱-۳-۵- Feature selection

یک فایل اکسل با ۲۰ ویژگی از جمله داده‌های $PM_{2.5}$ تمامی ایستگاه‌ها به همراه PM_{10}، CO، TEMP، PRES، DEWP، RAIN، wd و WSPM مربوط به ایستگاه Aotizhongx درست کنید.

توجه داشته باشید که این فایل اکسل را نیز همراه با گزارش خود ارسال نمایید.

۱-۳-۶- Supervised dataset

در نهایت داده‌ها را به فرم Supervised در بیاورید تا آماده Train شوند. همچنین ۸۰٪ دادگان یعنی ۲۸۰۵۲ ساعت ابتدایی را برای آموزش و ۲۰٪ یا ۷۰۱۲ ساعت نهایی را برای Test جدا نمایید.

۱-۴- آموزش شبکه

روش CNN-LSTM ارائه شده در مقاله را با بهترین hyperparameter های معرفی شده آموزش دهید و مقادیر MAE، RMSE و R^2 را به ازای Lag های ۱ و ۷ روز گزارش کنید.

پرسش ۲ – تشخیص اخبار جعلی

هدف این تمرین آشنایی شما با تسک fake news detection است و مقاله ای که در این بخش از آن استفاده خواهید کرد، تحت عنوان Fake news detection: A hybrid CNN-RNN based deep learning approach به پیوست برای شما وجود دارد. در این تسک شما با یک binary classification رو به رو هستید و نیاز است که متن خبر موجود را به عنوان خبر صحیح یا جعلی دسته بندی کنید. در این مقاله نتایج برای دو دیتاست گزارش شده‌اند که شما تنها باید برای دیتاست FA-KES که به پیوست برای شما ارسال شده است، نتایج را گزارش کنید. در صورتی که تمایل به استفاده از پارامترهای متفاوتی از موارد گفته شده در مقاله دارید، لطفا توجیه خود را در گزارش ذکر کنید.

۲-۱- توضیحات مدل‌ها

در ابتدا تفاوت معماری RNN و LSTM را توضیح دهید. چه توجیهی در مورد اینکه در داده‌های متنی ویژگی بازگشتی بودن موثر است، دارید. در مورد مدل Hybrid که در مقاله گفته شده نیز توضیحاتی دهید و به تفاوتی که با مدل‌های بازگشتی عادی دارد اشاره کنید.

۲-۲- ورودی مدل

با توجه به اینکه ورودی شما به عنوان اخبار در قالب متن به مدل داده خواهد شد توضیح دهید word embedding به چه منظور استفاده می‌شود؛ در مورد دلیل استفاده از آن برای ورودی‌های متنی تحقیق کنید و راه های ایجاد embedding کلمات را توضیح دهید. روشی که شما برای انجام این تمرین استفاده می‌کنید را نیز قید کنید.

۲-۳- پیاده سازی

۲-۳-۱- پیش‌پردازش

جهت پیاده سازی پیش پردازش های لازم را انجام دهید و موارد استفاده شده را ذکر کنید.

۲-۳-۲- آموزش مدل‌ها

پس از انجام پیش پردازش شما به ایجاد مدل برای انجام تسک نیاز دارید در این بخش شما یک مدل RNNو یک مدل hybrid(CNN-RNN) ایجاد کنید و به آموزش مدل ها بپردازید و دو نمودار loss و Accuracy را در طول زمان یادگیری رسم کنید. معیار های Accuracy, Precision, recall, F1-score در دو مدل را بدست آورید و به مقایسه دو مدل بپردازید.

۲-۴- تحلیل نتایج

تحلیل خود را در مورد نتایج به دست آمده بیان کنید نظر شما در مورد علت دقت‌های بدست آمده چیست، چگونه می‌توان بهبودی در دقت داشت؟ آیا به نظر شما ضعفی در مدل ها وجود دارد که بتوان با برطرف کردن آن ها به بهبود دقت کمک کرد؟