به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

شبکه های عصبی و یادگیری عمیق

مینی پروژه شماره ۲

اردیبهشت ماه ۹۹

یکی از کاربرد های شبکه های عصبی Recurrent پیش بینی است. در این پروژه قصد براین است که با کمک شبکه های عصبی Recurrent آلودگی هوا را پیش بینی کنید .
دیتاست ضمیمه شده با نام pollution_dataSet.npy شامل اطلاعات آب و هوای شهر Beijing چین هست. هر ردیف شامل اطلاعات آب و هوایی یک ساعت است. اطلاعات مربوط به بازه ی زمانی 2010 تا 2015 هست و ردیف ها به ترتیب از 2010 تا 2015 برای هر ساعت ثبت شده اند.
فیچر های موجود در دیتابه ترتیب از چپ به راست :

pollution, dew, temp, pressure, wind_dir, wind_spd, snow, rain

۱. قسمت اول (طراحی شبکه های عصبی)

شبکه ی عصبی ای طراحی کنید که داده ها را به صورت سری دریافت کند(هر هشت ویژگی از هر داده) و آلودگی هوای یک ساعت آینده را پیش بینی کند. (سایز پنجره ی زمانی را ۱۱ ساعت گذشته انتخاب کنید تا بتوانید آلودگی هوا در ساعت ۱۲ رو پیش بینی کنید سپس به عنوان آلودگی روزانه ۲ مقدار را گزارش دهید یکی در ساعت ۱۲ ام و دیگری در ساعت ۲۴ ام)

۱ برای هر کدام از شبکه هایی که طراحی می کنید نمودار train, testو همچنین نمودار مقدار حقیقی و پیش بینی را رسم کنید(12000 رکورد اول را به عنوان داده train و validation و 3000داده بعدی را به عنوان داده test استفاده کنید)
۲ شبکه را با GRU،LSTM،RNN طراحی کنید و سرعت ودقت هر کدام را مقایسه کنید. (زمان آموزش برای یک تعداد epoch مشخص اندازه بگیرید) تفاوت ها را تحلیل کنید.

۳ نحوه‌ی عملکرد شبکه برای تابع های هزینه متفاوت و روش های بهینه سازی متفاوت (ADAgrad , RMSProp ,Adam)و همچنین توابع خطای متفاوت (,MSE MAE) بررسی کنید.

۴ سری های زمانی حالت های مختلفی دارند می توانند اطلاعات چند ساعت اخیر باشند و یا ساعت ثابت اما در روز های پی در پی و یا ساعت ثابت اما در هفته های پی در پی.مثلا استفاده از ساعات پشت هم برای پیش بینی (ساعات پیاپی در روز) یا مثلا یک ساعت مشخص برای هفت روز گذشته یعنی فاصله ها یک روز (یک ساعت خاص برای روز های پیاپی) و یا یک ساعت و یک روز در هفته،برای 4 هفته ی گذشته یعنی فاصله ها یک هفته(ساعت های خاص در روز های خاص برای هفته های پیاپی). عملکرد شبکه را برای سری زمانی های هفتگی(با استفاده از تابع رندم، یک ساعت رندم را انتخاب کنید و از داده ۶ روز پیاپی برای پیش بینی آلودگی در همان ساعت از روز ۷ ام استفاده کنید) و ماهانه (با استفاده از تابع رندم، یک روز رندم و یک ساعت رندم انتخاب کنید و با داده ۳ هفته پیاپی همان روز و همان ساعت، آلودگی را در همان روز و همان ساعت برای هفته ۴ ام پیش بینی کنید) نیز بررسی کنید.

۵ تاثیر لایه dropoutرا بروی یک شبکه ی طراحی شده(به دلخواه) بررسی کنید.

۶ بهترین شبکه بازگشتی در مراحل قبل را انتخاب کنید و دو شبکه ی بازگشتی دیگر با همان ساختار نیز به موازات آن بسازید .سپس سه نوع سری زمانی توضیح داده شده را برای پیش بینی مقدار آلودگی در یک ساعت مشخص به هر کدام از آن ها اعمال کنید . سپس به کمک یک لایه ی fusionخروجی سه شبکه ی recurrent را با هم ترکیب کنید. نتیجه را بررسی کنید .لزومی ندارد که سایز سری زمانی های سه شبکه ی موازی یکسان باشد.

۷ اکنون فرض کنید برای پیش بینی آلودگی، فقط می‌توانید از دو ستون دیگر (به جز آلودگی) کمک بگیرید (یعنی در مجموع ۳ ستون از ۸ ستون داده). برای اینکه میزان دقت پیش بینی شما بالاتر رود باید سعی کنید ۲ ستونی را انتخاب کنید که بیش ترین تاثیر را در پیش بینی درست آلودگی داشته باشد. روش شما برای انجام اینکار چیست؟

۸ روش خود را برای قسمت قبل پیاده سازی کرده و با استفاده از این ۳ ستون(آلودگی و ۲ ستونی که یافته اید) میزان آلودگی روزانه را برای هر سه سلول LSTMوRNNو GRU بررسی کنید(از یک تابع بهینه ساز و خطا دلخواه کمک بگیرید)

۲. قسمت دوم (نقصان دادگان)

یکی از چالش های مهم در حل مسائل یادگیری ماشین، نقصان دادگان می‌باشد. در حل مسائل واقعی گاها به علت خراب شدن ابزار های اندازه گیری یا خطای انسانی و ... اندازه گیری به خوبی انجام نمی‌شود و داده از بین می‌رود. در ادامه این پروژه می‌خواهیم تا حدی با این چالش روبرو شویم و راه حل شما برای رویارویی با آن را بررسی کنیم.

۱. برای هر ستون از قسمت دادگان آموزش، به صورت رندم ۲۰ درصد از دادگان را حذف کنید.(توجه کنید که برای هر ویژگی(ستون) باید به صورت مجزا این کار را انجام دهید)
۲. اکنون فرض کنید دادگان آموزشی که در اختیار شماست به این شکل می‌باشد و قصد داریم قسمت اول را انجام دهیم، اما پیش از حل نیاز داریم که برای حل مشکل نقصان دادگان، روشی اتخاذ کنیم.

۳. با تحقیق در منابع ۳ روش بر طرف کردن نقصان دادگان را بیابید و به صورت کامل شرح دهید.
۴. یک روش را به دلخواه انتخاب کنید و با استفاده از آن دادگان از بین رفته را پیش بینی کنید.

۵. با استفاده از روش خطای MSE، میزان خطای موارد پیش بینی شده برای دادگان از دست رفته را برای هر ستون را گزارش دهید.(۸ مقدار خطا باید گزارش دهید)(امتیازی : ۵ تیمی که بهترین روش پیش بینی دادگان از دست رفته را داشته باشند(خطای پیش بینی آنها کمتر باشد) تا ۲۵ نمره بونس خواهند داشت)

۶. اکنون با استفاده از دادگان پیش بینی شده، برای سلول های GRU، LSTM و یکی از توابع هزینه به دلخواه، میزان آلودگی روزانه را پیش بینی کنید. نتیجه و دقت را با زمانی که دادگان بی نقص در اختیار داشتید مقایسه کنید.

نکات:

- مهلت تحویل این تمرین تا ۱۳ خرداد می‌باشد.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفا تمامی نکات و فرض‌هایی که برای پیاده سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- شما می‌توانید این تمرین را در گروه های حداکثر دو نفره انجام دهید.
- در صورت مشاهده‌ی تقلب نمرات تمامی افراد شرکت کننده در آن صفر لحاظ می‌شود.
- استفاده از کدهای آماده برای تمرینها مجاز نمی‌باشد. برای مینی پروژه ها فقط برای قسمت هایی از کد برای پیاده سازی می‌توانید از کدهای آماده راهنمایی بگیرید. بنابراین کپی کردن سختار ها و کدهای آماده و حل شده از اینترنت **تقلب** محسوب می‌شود.
- نحوه‌ی محاسبه‌ی تاخیر به این شکل است : مهلت بدون کسر نمره تا تاریخ ۱۳ خرداد اعلام شده و تاخیر تا یک هفته بعد یعنی ۲۰ خرداد با ۳۰ درصد کسر نمره محاسبه خواهد شد.
- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌ی زیر با دستیار آموزشی مربوطه در تماس باشید:

هاشم پور(hamidreza.hashemp@ut.ac.ir)

حاجی قاسمی(hajighasemiamir@gmail.com)