

به نام خدا



دانشگاه تهران



دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین چهارم

توحید عبدی	نام دستیار طراح	پرسش ۱
Tohid.abdi@ut.ac.ir	رایانامه	
محمد ولی نژاد	نام دستیار طراح	پرسش ۲
m.valinezhad@ut.ac.ir	رایانامه	
۱۴۰۳.۰۳.۰۱	مهلت ارسال پاسخ	

فهرست

- قوانین ۱
- پرسش ۱. تحلیل احساسات متن فارسی ۱
- ۱-۱. مجموعه داده ۱
- ۲-۱. پیش پردازش داده ها ۱
- ۳-۱. نمایش ویژگی ۱
- ۴-۱. ساخت مدل ۲
- ۵-۱. ارزیابی ۳
- ۶-۱. امتیازی ۳
- پرسش ۲ - سامانه های سایبرفیزیکی: نگهداری هوشمند ۴
- ۱-۲. پیش پردازش داده ها ۴
- ۲-۲. مدل سازی و ارزیابی ۵
- ۳-۲. مقایسه با مدل های پایه ۶

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.
- **کدها حتما باید در قالب نوت‌بوک با پسوند .ipynb تهیه شوند، در پایان کار، تمامی کد اجرا شود و خروجی هر سلول حتما در این فایل ارسالی شما ذخیره شده باشد.** بنابراین برای مثال اگر خروجی سلولی یک نمودار است که در گزارش آورده‌اید، این نمودار باید هم در گزارش هم در نوت‌بوک کدها وجود داشته باشد.
- **در صورت مشاهده‌ی تقلب امتیاز تمامی افراد شرکت‌کننده در آن، 100- لحاظ می‌شود.**
- تنها زبان برنامه نویسی مجاز **Python** است.
- استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست. در صورتی که دو گروه از یک منبع مشترک استفاده کنند و کدهای مشابه تحویل دهند، تقلب محسوب می‌شود.
- نحوه محاسبه تاخیر به این شکل است: پس از پایان رسیدن مهلت ارسال گزارش، حداکثر تا یک هفته امکان ارسال با تاخیر وجود دارد، پس از این یک هفته نمره آن تکلیف برای شما صفر خواهد شد.

○ سه روز اول: بدون جریمه

○ روز چهارم: ۵ درصد

○ روز پنجم: ۱۰ درصد

○ روز ششم: ۱۵ درصد

○ روز هفتم: ۲۰ درصد

- حداکثر نمره‌ای که برای هر سوال می‌توان اخذ کرد ۱۰۰ بوده و اگر مجموع بارم یک سوال بیشتر از ۱۰۰ باشد، در صورت اخذ نمره بیشتر از ۱۰۰، اعمال نخواهد شد.

○ برای مثال: اگر نمره اخذ شده از سوال ۱ برابر ۱۰۵ و نمره سوال ۲ برابر ۹۵ باشد، نمره نهایی تمرین ۹۷.۵ خواهد بود و نه ۱۰۰.

- 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. تحلیل احساسات^۱ متن فارسی

هدف از این سوال آشنایی با وظیفه^۲ تحلیل احساسات بر روی متن فارسی استخراج شده از توییتر می باشد. تحلیل احساسات شامل شناسایی و دسته بندی احساسات یا نظرات افراد مرتبط با موضوعات مختلف است. برای آشنایی با روند سوال، [این مقاله](#) را مطالعه کنید.

۱-۱. مجموعه داده^۳

(۱۰ نمره)

برای شروع، [این مجموعه داده](#) را از Kaggle دریافت کنید. توجه داشته باشید که این مجموعه داده با داده مورد استفاده در مقاله متفاوت است. فقط ستون های tweet و emotion برای این سوال مورد نیاز هستند. در صورت عدم دسترسی، می توانید از فایل فشرده این مجموعه داده که پیوست شده است استفاده کنید. کلاس های موجود در ستون emotion و تعداد نمونه های هر کلاس را به کمک یک نمودار میله ای^۴ نمایش دهید.

۲-۱. پیش پردازش داده ها

(۲۰ نمره)

پیش پردازش متن در پردازش زبان طبیعی برای بهبود عملکرد مدل بسیار مهم است. مراحل پیش پردازش ذکر شده در قسمت ۳.۲ مقاله را اعمال کنید و برای هر یک از مراحل، یک مثال که پیش پردازش مورد نظر روی آن اعمال شده است را قبل و بعد از پیش پردازش چاپ کنید. توجه داشته باشید ممکن است برخی از این مراحل برای این مجموعه داده نیاز نباشد و تغییری در هیچ یک از سطرها ایجاد نکند. برای انجام پیش پردازش های این بخش می توانید از کتابخانه re^۵ و کتابخانه هایی که پیش پردازش زبان فارسی را پشتیبانی می کنند استفاده کنید.

۳-۱. نمایش ویژگی^۶

Bar Plot^۴
Regular Expression^۵
Feature Representation^۶

Sentiment Analysis^۱
Task^۲
Dataset^۳

(۲۰ نمره)

در وظایف پردازش زبان طبیعی^۱، داده‌هایی که به طور کلی پردازش می‌شوند، متن خام هستند. با این حال، مدل‌ها فقط می‌توانند اعداد (Id) را پردازش کنند، بنابراین باید از توکن‌ساز^۲ها برای تبدیل متن خام به اعداد استفاده کنید. داده‌های متنی پیش‌پردازش‌شده را با توکن‌ساز [ParsBERT](#) به اعداد تبدیل کنید. برای این که تمام سطرها طول یکسانی داشته باشند، از لایه‌گذاری^۳ استفاده کنید و حداکثر طول جملات را برابر با ۳۲ در نظر بگیرید.

تعبیه‌ها^۴ کلمات را به‌عنوان بردارها در فضایی با ابعاد بالا نشان می‌دهند که روابط معنایی را به تصویر می‌کشند. این تعبیه‌ها مدل‌های یادگیری ماشین را قادر می‌سازند الگوها و احساسات را از داده‌های متنی بیاموزند. در این مرحله، به کمک مدل از پیش آموزش دیده^۵ [ParsBERT](#) بردار تعبیه را برای ورودی‌ها به دست آورید. با تغییر پیکربندی^۶ مدل، ابعاد بردار تعبیه را برابر با ۱۲۰ در نظر بگیرید. توجه داشته باشید که برای مدیریت حافظه می‌توانید از تکنیک‌هایی مانند تکه‌تکه کردن و کتابخانه `gc`^۷ استفاده کنید.

ابعاد پیش‌فرض بردار تعبیه در ParsBERT چقدر است؟ تعداد ابعاد این بردار بیانگر چیست؟ مفهوم بردار تعبیه را توضیح دهید و بیان کنید به نظر شما کدام یک از کلمات موجود در مجموعه داده ممکن است تعبیه نزدیک به هم داشته باشند؟ (نیازی به کد نیست)

۴-۱. ساخت مدل

(۳۵ نمره)

داده‌ها را با نسبت ۷۰-۳۰ به دو دسته آموزش و تست تقسیم کنید و ۰.۲ از داده‌های آموزش را به عنوان اعتبارسنجی^۸ در نظر بگیرید. الگوریتم جستجوی حریصانه^۹ برای یافتن هایپرپارامترهای بهینه برای مدل CNN-LSTM را در فضای جستجو با ۸ حالت زیر اعمال کنید و در نهایت هایپرپارامترهای بهینه که منجر به کم‌ترین خطای اعتبارسنجی می‌شود را گزارش کنید. هایپرپارامترهای دیگر مدل را مطابق با جدول ۳ مقاله اعمال کنید.

```
batch_sizes = [8, 64]
learning_rates = [0.001, 0.0001]
optimizers = [Adam, SGD]
```

Configuration^۶
Garbage Collector^۷
Validation^۸
Greedy Search^۹

Natural Language Processing^{۱۱}
Tokenizer^۲
Padding^۳
Embeddings^۴
Pretrained^۵

در مرحله بعد، مدل‌های CNN و LSTM ساده را با هایپرپارامترهای بهینه به دست آمده ایجاد کرده و آموزش دهید. نیازی به اعمال الگوریتم جستجوی حریصانه برای این مدل‌ها نیست. به نظر شما، هر یک از این مدل‌ها چه نقاط ضعف و چه نقاط قوتی دارند و ادغام این دو مدل با چه هدفی انجام می‌شود؟

۵-۱. ارزیابی

(۱۵ نمره)

داده‌های تست را به کمک معیارهای ارزیابی ذکر شده در مقاله ارزیابی کنید و یک جدول مشابه جدول ۴ مقاله برای مدل‌های CNN-LSTM، CNN و LSTM چاپ کنید.

روش‌های macro averaging، micro averaging و weighted averaging برای محاسبه میانگین معیارهای ارزیابی را مقایسه کنید و توضیح دهید هر یک از این روش‌ها چه تاثیری بر مقدار عددی این معیارها در این مسئله دارد.

۶-۱. امتیازی

(۵ نمره)

از روش کیسه کلمات^۱ برای نمایش ویژگی استفاده کنید و به کمک کتابخانه sklearn روش‌های سنتی ماشین لرنینگ که در مقاله ذکر شده‌اند را آموزش داده و به کمک دادگان تست ارزیابی کنید. نتایج را به جدول نتایج بخش قبل اضافه کنید. برای کاهش استفاده از منابع، در این بخش می‌توانید از بخشی از داده‌ها نمونه^۲ گرفته و از این نمونه‌ها برای آموزش و ارزیابی مدل‌ها استفاده کنید.

^۲ Sample

^۱ Bag of Words

پرسش ۲ – سامانه‌های سایبرفیزیکی^۱: نگهداری هوشمند^۲

اصطلاح سامانه‌های سایبرفیزیکی برای اولین بار بین اعضای هیات علمی دانشگاه برکلی در سال ۲۰۰۶ معرفی شد. سامانه‌های سایبرفیزیکی را می‌توان به عنوان سامانه‌هایی تعریف کرد که اطلاعات را از محیط فیزیکی با استفاده از حسگر^۳ ها و کانال‌های ارتباطی^۴ جمع‌آوری می‌کنند، آنها را با استفاده از کنترل‌کننده^۵ هایی تجزیه و تحلیل می‌کنند، سپس محیط فیزیکی و فرآیندهای مربوطه را از طریق محرک^۶ ها برای دستیابی به یک هدف خاص، در طول عملیاتشان تحت تأثیر قرار می‌دهند.

گسترش سامانه‌های سایبرفیزیکی و پیشرفت فناوری‌های اینترنت اشیا^۷ منجر به دیجیتالی‌شدن بخش صنعتی شده است. این سامانه‌ها امکاناتی را برای فرآیند تولید با قابلیت اطمینان^۸، در دسترس بودن^۹، قابلیت نگهداری^{۱۰} و ایمنی^{۱۱} بالا فراهم می‌کنند، اما از طرف دیگر این سامانه‌ها ارزیابی سلامت را پیچیده تر و چالش برانگیز می‌کنند. هر دستگاهی به دلیل فرسودگی و بالا رفتن سن، سلامت آن به تدریج بدتر شده و در نهایت از کار می‌افتد. به این نقطه از عمر دستگاه یک نقطه خرابی یا failure می‌گویند. در پیش از این نقطه باید تعمیر و نگهداری را اجرا کنیم تا دستگاه را به وضعیت سالم برگردانیم و سامانه دچار اختلال نشود.

در این تمرین به استفاده از مدل پیشنهادی [مقاله](#) که به منظور regression و classification عمر مفید باقی مانده^{۱۲} معرفی شده است، می‌پردازیم. این مقاله یک مدل یادگیری عمیق CNN LSTM را برای این منظور معرفی کرده است و آزمایش‌هایی را بر روی مجموعه داده‌های [NASA's C-MAPSS](#) به منظور ارزیابی و مقایسه با مدل‌های پایه انجام داد.

۲-۱. پیش پردازش داده‌ها

(۳۵ نمره)

در این بخش می‌خواهیم با مجموعه داده‌های [NASA's C-MAPSS](#) آشنا شویم و پیش پردازش‌های مورد نیاز را انجام دهیم. (از طریق این [لینک](#) می‌توانید به مجموعه داده دسترسی پیدا کنید. از آنجا که

IoT^۷
Reliability^۸
Availability^۹
Maintainability^{۱۰}
Safety^{۱۱}
Remain Useful Life (RUL)^{۱۲}

Cyber-physical Systems (CPSs)^۱
Intelligent Maintenance^۲
Sensor^۳
Communication channels^۴
Controller^۵
Actuator^۶

داده ها در قالب فایل متنی می باشند، به منظور راحتی برای شما، فایل csv آن ها نیز پیوست شده است. از فایل های **train_FD001T** ، **test_FD001** و **RUL_FD001** برای این بخش استفاده کنید.)

- در ابتدا توضیح مختصری در مورد مجموعه داده ذکر شده دهید.
- سپس با توجه به عملیات های صورت گرفته در مقاله برای پیش پردازش داده ها اقدام کنید. پیش پردازش های صورت گرفته شامل **Data Selection** ، **Data Normalization** ، **Data labeling** و **Timing window** می باشد.
- این عملیات را برای مدل **regression** و **classification** انجام دهید، به صورت کامل عملیاتی که انجام داده اید را شرح داده و تحلیل های خود را با دلیل بیان کنید.

راهنمایی: برای محاسبه ستون مربوط به **Remain Useful Life** برای هر سطر داده های **test**، ابتدا بیشترین **cycle** مربوط به هر موتور در داده های **test** را پیدا کرده و با مقادیر متناظر در فایل **RUL_FD001** جمع کنید، سپس حاصل را از مقدار ستون **cycle** هر سطر تست تفریق کنید.

۲-۲. مدل سازی و ارزیابی

(۳۵ نمره)

حال در این بخش قصد داریم مدل ارائه شده توسط مقاله برای **regression** و **classification** را پیاده سازی و ارزیابی کنیم. (نکته: دستیابی به عددهای ذکر شده در مقاله ضرورتی ندارد. همچنین معیار های ارزیابی مدل ها در مقاله توضیح داده شد، باتوجه به آن توضیحات اقدام به ارزیابی مدل ها کنید.)

الف) Classification: در این قسمت مدل پیشنهادی مقاله را پیاده سازی کنید. سپس باتوجه به پارامترهای بیان شده در مقاله و داده های تهیه شده در بخش قبل اقدام به آموزش مدل کنید.

- ابتدا مدل را بدون استفاده از رویکرد **Early stopping** آموزش دهید. ($epochs=100$, $optimizer=adam$)
- در گام بعد اقدام به ارزیابی مدل با استفاده از داده های **test** تهیه شده در بخش قبل کنید و معیارهای **precision**, **recall**, **f1-score** و **accuracy** را برای هر کلاس گزارش کنید و نتایج را تحلیل کنید.
- همچنین ماتریس درهم ریختگی^۱ و نمودار منحنی **ROC**^۲ را ترسیم کنید و نمودارها را تحلیل کنید.
- حال با استفاده از رویکرد **Early stopping** این مراحل را تکرار کنید. نتایج حاصل از دو بخش را مقایسه و تحلیل کنید.

^۲ Receiver Operating Characteristic curve

^۱ Confusion matrix

ب) Regression: در این قسمت مدل پیشنهادی مقاله را پیاده سازی کنید. سپس با توجه به پارامترهای بیان شده در مقاله و داده‌های تهیه شده در بخش قبل اقدام به آموزش مدل کنید.

- ابتدا مدل را بدون استفاده از رویکرد Early stopping آموزش دهید. (epochs=100, optimizer=rmsprop)
- در گام بعد مدل را در دو حالت استفاده از تمام پنجره‌های زمانی و حالت استفاده از آخرین پنجره زمانی داده‌های test تولید شده در بخش قبل ارزیابی کنید و معیارهای RMSE، MSE، MAE و MAPE را گزارش کنید و نتایج آن‌ها را تحلیل کنید.
- همچنین نمودار مقادیر واقعی و تخمین زده شده برای RUL^1 را ترسیم کنید. نمودارها را تحلیل کنید.
- حال با استفاده از رویکرد Early stopping این مراحل را تکرار کنید. نتایج حاصل از دو بخش را مقایسه و تحلیل کنید.

۲-۳. مقایسه با مدل‌های پایه

(۳۰ نمره)

در این بخش با استفاده از معماری CNN LSTM پیشنهادی توسط مقاله، معماری یک مدل LSTM و یک مدل CNN را برای دو عملیات classification و regression طراحی کنید. حال با استفاده از داده‌های تهیه شده در بخش ۱، مدل‌های طراحی شده را با استفاده از رویکرد Early stopping آموزش داده و مانند بخش ۲ آن‌ها را ارزیابی کرده و نتایج و تحلیل‌های خود را بیان کنید. (نکته: دستیابی به عددهای ذکر شده در مقاله ضرورتی ندارد.)