چالش پیش‌بینی میزان مصرف اینترنت مشترکین همراه اول

Order of the Phoenix

(دانشگاه صنعتی خواجه نصیرالدین طوسی)

سالار حسینی شمعچی

احمد احمدی

محمد حسین‌زاده

تعریف مسئله:

در این مسابقه به پیش‌بینی مصرف اینترنت گروه خاصی از مشترکین منتخب همراه اول خواهیم پرداخت. در واقع نیاز است تا مدلی هوشمند و منعطف بر پایه‌ی علم هوش مصنوعی روز طراحی و توسعه یابد. در این مسابقه داده‌های اینترنت مصرفی روزانه‌ی کاربران به مدت سه ماه متوالی از سیم کارت‌های مختلف گمنام شده در اختیار شرکت کنندگان قرار می‌گیرد. شرکت کنندگان می‌بایست مدلی طراحی نمایند تا با استفاده از ۶۹ روز (دو ماه و یک هفته) داده‌ی دریافتی، به بررسی رفتار داده‌ها پرداخته و در قالب سه مرحله، به پیش‌بینی میزان مصرف اینترنت مشترکین در سه هفته‌ی انتهایی اقدام نمایند. هفته‌ی اول به عنوان مرحله‌ی آزمایشی در نظر گرفته می‌شود و نتایج آن در ارزیابی تاثیر نخواهد داشت. در انتهای هر مرحله، مقدار واقعی مصرف اینترنت آن مرحله‌ی مشترکین در اختیار شرکت کنندگان قرار می‌گیرد تا در مراحل بعدی از آن استفاده شود.

روش حلّ مسئله:

شبکه‌های عصبی حافظه‌ی کوتاه-بلند مدت[1] در سال‌های اخیر به دلیل عملکرد پرقدرت در تحلیل داده‌های ترتیب‌دار، به خصوص سری‌های زمانی که در آن‌ها اثر طولانی مدت نیز حائز اهمیت است، از جایگاه ویژه‌ای برخودار گشته‌اند. در این چالش نیز، از آنجا که نوع داده‌های مورد استفاده بر اساس سری زمانی در نظر گرفته شده است، این نوع شبکه به عنوان ساختار مطلوب برای حل مسئله انتخاب شده است.

برای رسیدن به بهترین نتایج ممکن، ابتدا روش مناسب پیش پردازش داده‌ها بررسی و انتخاب شده، سپس با تحلیل و آزمون و خطای مکرر، ساختار و پارامترهای مناسب برای شبکه‌ی LSTM برگزیده شده‌اند. سپس نتایج شبکه با روش‌های مختلف سنجیده شده و با استفاده از بهترین شبکه‌ی آموزش دیده، خروجی‌های لازم پیش‌بینی شده‌اند. در ادامه جزئیات این روند مورد بحث قرار می‌گیرد.

---

[1] Long-short Term Memory (LSTM)

پیش پردازش داده‌ها:

ابتدا درون یک حلقه داده‌های مربوط به هر شخص از فایل آموزش و فایل هفته‌ی اول جدا شده و جای خالی داده‌های گم شده با استفاده از میانگین گیری از ویژگی و شخص مربوط به آن داده پر می‌شود. سپس داده‌های ورودی مربوط به هر شخص به صورت سری‌هایی از داده‌های چند روز متوالی انتخاب می‌شود. این داده‌ها شامل تمام ویژگی‌های داده شده مربوط به هر روز بوده، اما حجم اینترنت مصرفی مربوط به روز قبل استفاده می‌شود. دلیل این کار این است که حجم اینترنت مصرفی روز آخر هر سری به عنوان خروجی آن سری در نظر گرفته شده، لذا این ویژگی برای روزهای قبل نیز لازم است یک روز به جلو شیفت پیدا کند. تعداد روزهای هر سری نیز از ۵ الی ۷۵ روز می‌تواند متغیر باشد. به این معنی که اگر طول یک سری n باشد، داده‌های مربوط به یک شخص از روز اول تا روز nام را در بر می‌گیرد. نهایتا ویژگی‌های مربوط به روزهای پیشین در سری‌های کوتاه‌تر از ۷۵ روز همگی به عنوان صفر در نظر گرفته شده‌اند، تا تمامی سری‌های آموزشی طول یکسان داشته باشند.

ابتدا داده‌های مربوط به فایل آموزش و فایل هفته‌ی اول پس از گذراندن مراحل پیش پردازش و آماده سازی سری‌ها به دو دسته‌ی آموزش و صحه گذاری تقسیم شده‌اند تا معیار خوبی برای سنجش عملکرد شبکه داشته باشیم. داده‌های مربوط به فایل هفته‌ی دوم نیز با اضافه شدن به داده‌های آموزش و داده‌های هفته‌ی اول، به عنوان داده‌های تست مورد استفاده قرار می‌گیرند. این کار به ما اجازه می‌دهد تا مقاومت و عمومیت پذیری شبکه را به درستی بسنجیم. همچنین لازم به ذکر است، همه‌ی این داده‌ها بر اساس مقادیر کمینه و بیشینه‌ی داده‌های آموزش نرمالیزه شده و این مقادیر اکسترمم نیز برای استفاده‌های آتی ذخیره می‌شوند.

پس از آزمون و خطای بسیار، و انتخاب ساختار و پارامترهای مناسب بر اساس نتایج داده‌های تست، برای افزایش هر چه بیشتر دقت مدل، و استفاده‌ی حداکثری از داده‌های موجود، داده‌های تست نیز به داده‌های آموزش اضافه شده و شبکه‌ی آموزش دیده این بار با داده‌های جدید آموزش را ادامه می‌دهد، تا به بهترین حالت ممکن برسد.

پارامترها و ساختار شبکه:

به عنوان ساختار از یک شبکه‌ی نسبتا ساده‌ی LSTM با دو لایه‌ی میانی که ۱۶ واحد دارند استفاده شده است. دلیل استفاده از این ساختار ساده نیز جلوگیری از مشکل overfit شدن می‌باشد. از آنجا که توابع فعال‌ساز لایه‌ی LSTM مخصوص سری‌های مرتب بهینه سازی شده‌اند، دخالتی در تغییر این توابع صورت نگرفته است. در لایه‌ی آخر نیز یک ساختار ساده‌ی dense با یک نورون و یک تابع فعال‌ساز خطی قرار گرفته تا حجم اینترنت مصرفی مربوط به روز آخر سری را پیش‌بینی کند.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, None, 32)]        0

 lstm (LSTM)                 (None, None, 16)          3136

 lstm_1 (LSTM)               (None, 16)                2112

 dense (Dense)               (None, 1)                 17

=================================================================
Total params: 5,265
Trainable params: 5,265
Non-trainable params: 0
_________________________________________________________________
```

شکل ۱- ساختار شبکه

بُعد دوم لایه‌ی اول شبکه به صورت None انتخاب شده، تا سری‌های مختلف با هر طولی را بتوان به عنوان ورودی به مدل آموزش دیده داد. تفاوت ابعاد لایه‌ی دوم و سوم نیز به این دلیل است که ورودی لایه‌ی LSTM لازم است سه بُعد داشته باشد، لذا آرگمان return_sequence در لایه‌ی دوم فعال شده تا تمامی داده‌های یک سری به لایه‌ی بعدی منتقل شود.

برای آموزش شبکه، تابع هزینه‌ی انتخاب شده میانگین مربع خطاها[۲] است که با استفاده از تابع بهینه‌ساز Adam به مقدار کمینه می‌رسد. سپس داده‌ها به صورت دسته‌های[۳] ۵۱۲ تایی با تکرار ۲۰۰۰ مرتبه به شبکه داده شده و با تنظیم وزن‌ها روند آموزش پیش می‌رود. در طول آموزش نیز هزینه‌ی ناشی از داده‌های صحه گذاری مورد بررسی قرار می‌گیرد. تمامی این پارامترها نیز با آزمون و خطای بسیار بهینه سازی شده‌اند.

برای جلوگیری بیشتر از مشکل overfit شدن و هدایت روند آموزش در راستای مطلوب، ابتدا از تنظیم کننده‌ی وزن‌های شبکه[۴] استفاده شد، که منجر به مشکل underfit شدن شبکه شد. سپس از ترفندهایی همچون کاهش نرخ یادگیری کسینوسی[۵] و ذخیره‌ی بهترین وزن‌ها در طول آموزش استفاده شد، که هر یک تاثیر مثبت خود را در مراحل آزمون و خطا نشان دادند.

همان طور که در بخش قبل گفته شد، پس از اتمام آموزش شبکه، بار دیگر روند آموزش را با داده‌های اضافه شده‌ی هفته‌ی دوم ادامه می‌دهیم تا به بهترین نتایج ممکن برسیم.

نتایج:

پس از اتمام یادگیری ابتدا برای بررسی عملکرد شبکه، آن را با استفاده از داده‌های صحه گذاری و آموزشی ارزیابی کرده، امتیاز R2 آن را محاسبه می‌کنیم، که برای داده‌های آموزش و صحه گذاری به ترتیب ۰.۷۴۱۶ و ۰.۲۴۷۹ به دست می‌آید. کم شدن امتیاز داده‌های صحه گذاری به دلیل داده‌های موجود در این مجموعه می‌باشد، که یک داده‌ی بسیار بزرگ باعث آن شده است. سپس برای بررسی مقاومت و عمومیت پذیری مدل، داده‌های هفته‌ی دوم به عنوان داده‌های تست بررسی می‌شوند، که امتیاز R2 آن نیز ۰.۵۷۱۲ به دست می‌آید. علاوه بر آن نمودار تغییرات تابع هزینه برای داده‌های آموزش و صحه گذاری را نیز برای ارزیابی روند آموزش رسم می‌کنیم که در شکل ۲ قابل مشاهده است. نهایتا مقادیر پیش‌بینی شده توسط شبکه برای داده‌های آموزش، صحه گذاری و تست را در برابر مقادیر واقعی رسم می‌کنیم تا عملکرد شبکه به صورت بصری مشخص شود. این نمودارها نیز در شکل‌های ۳، ۴ و ۵ قابل مشاهده هستند.

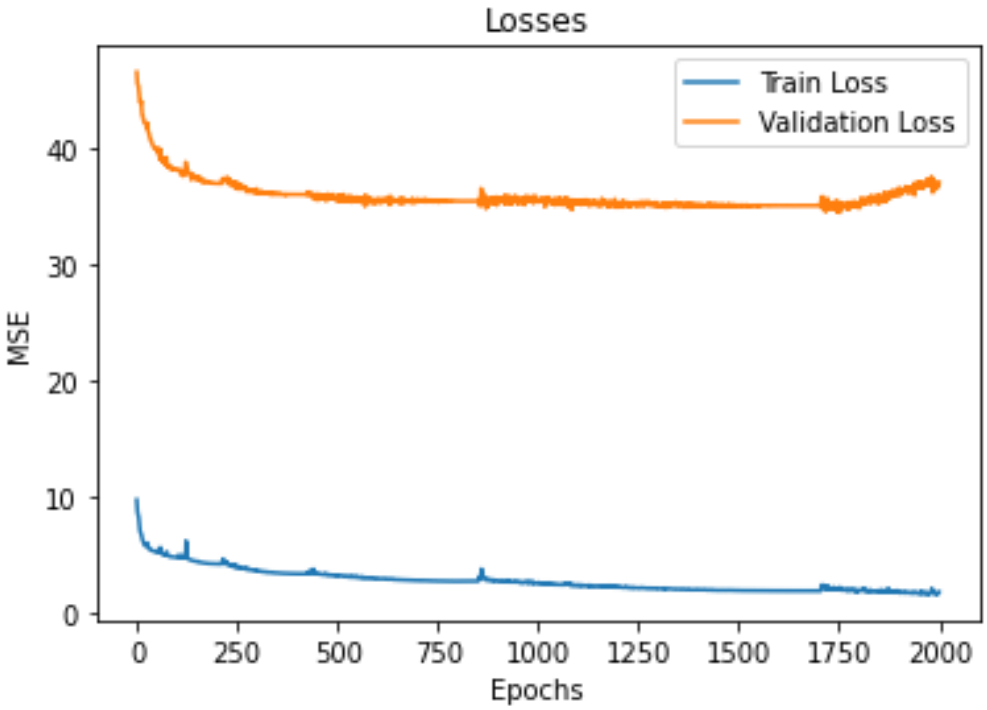
---

[۲] Mean Squared Error

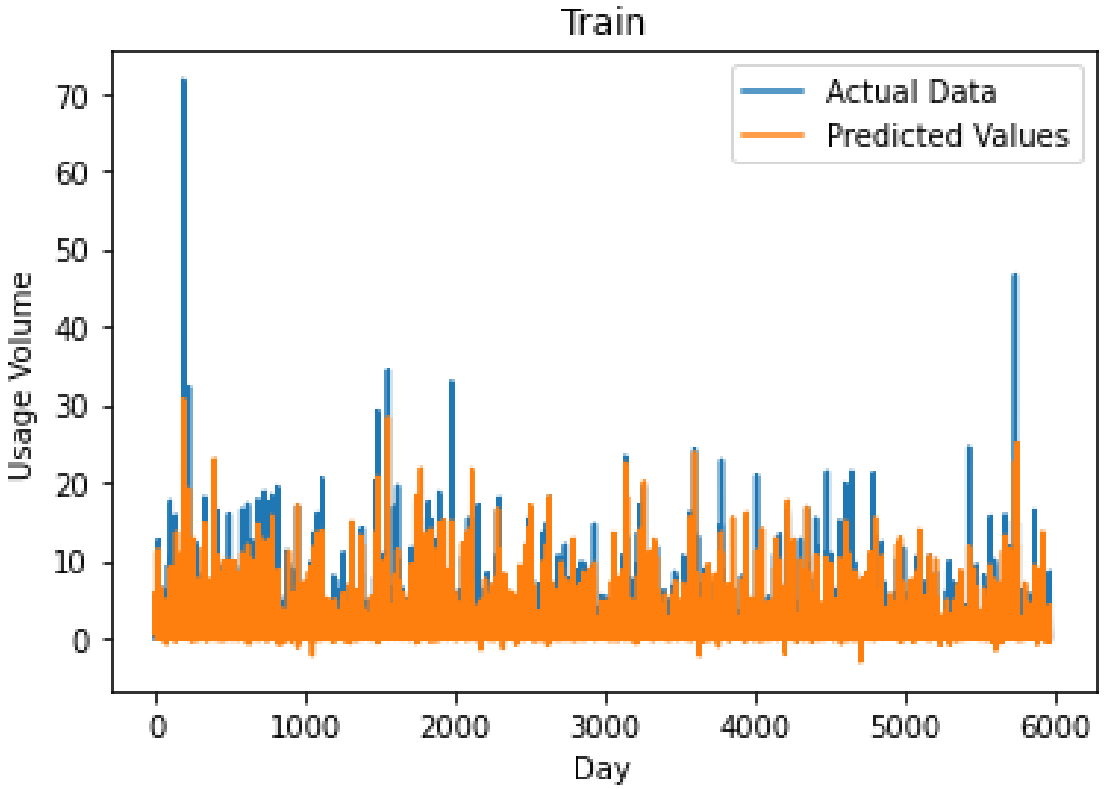
[۳] Batches

[۴] Kernel Regularizer

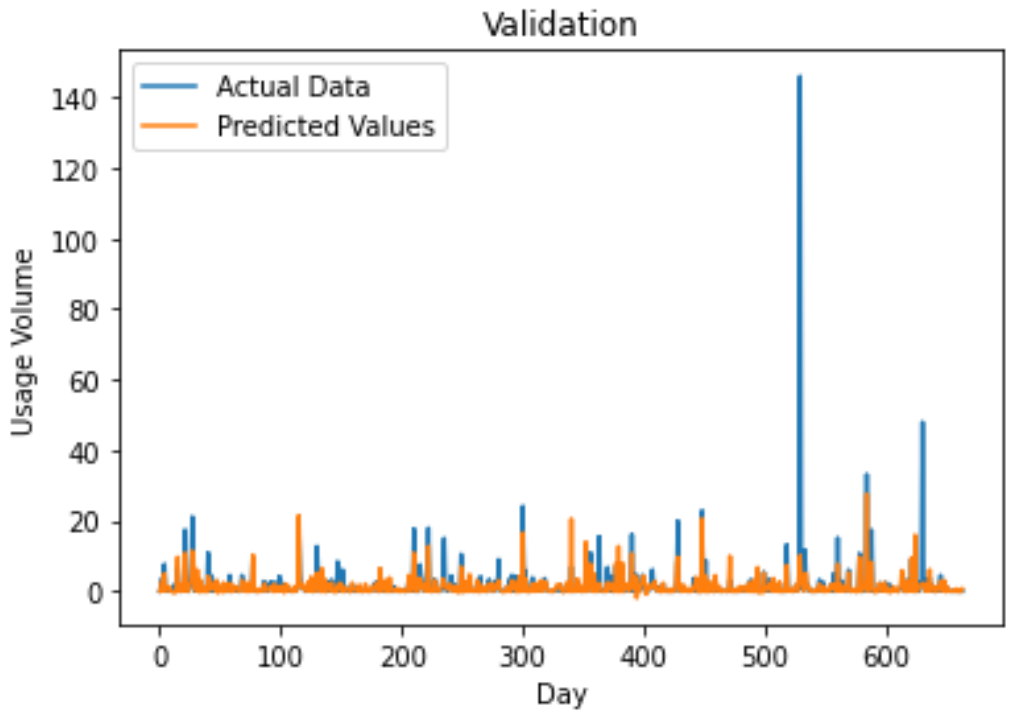
[۵] Cosine Learning Rate Decay



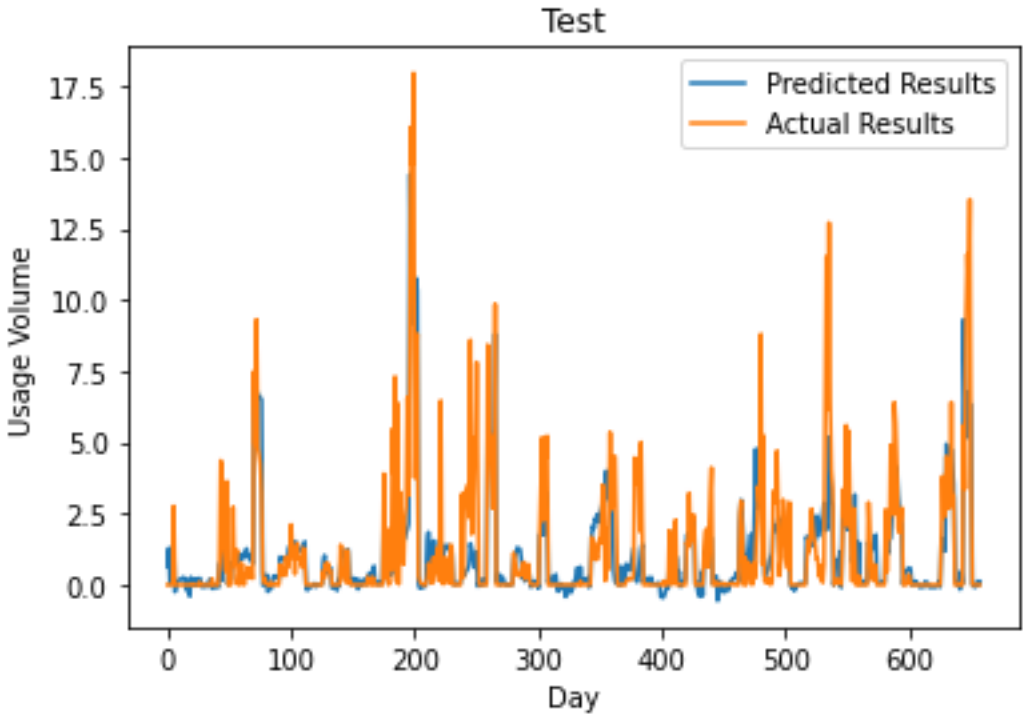
شکل ۲- روند تابع هزینه برای داده‌های آموزش و صحه گذاری



شکل ۳- مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های آموزش



شکل ۴- مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های صحه گذاری



شکل ۵ مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های صحه گذاری

پس از ارزیابی عملکرد شبکه با روش‌های ذکر شده، داده‌های هفته‌ی سوم به عنوان بخشی از ورودی به شبکه داده شده و حجم اینترنت مصرفی کاربران به عنوان خروجی شبکه گرفته شده و ذخیره می‌شود. این نتایج در فایل week3_results.csv قابل دسترسی هستند.

لازم به ذکر است برای ساخت داده‌های خروجی باید از بخش آخر فایل نوتبوک main.ipynb استفاده شود.

پیشنهاد: برای بهبود نتایج می‌توان شماره‌ی روز هفته را نیز به عنوان یکی از ویژگی‌ها در نظر گرفت.