Neural Networks and Deep Learning Assignment3

Instructors: **Dr. Bahrak** TA: مینا شیرازی

Deadline: 1404/02/14

پرسش 1: پیش بینی ارزش نفت

از رایج‌ترین کاربرد شبکه‌های حافظه‌دار پیش‌بینی سری‌های زمانی اشاره کرد. در این سوال با نحوه‌ی **پیش‌بینی ارزش نفت خام** با استفاده از چهار روش متفاوت آشنا خواهید شد.

1-1. مجموعه دادگان و آماده‌سازی (20 نمره)

مجموعه داده‌ی مورد استفاده در این سوال، شاخص `CL=F` را از سال ۲۰۱۰ تا کنون را از Yahoo Finance دانلود کنید. از بین ویژگی‌های داده شده، ویژگی اصلی مد نظر را `Adj Close` قرار دهید.

در برخی روزها داده‌ای ثبت نشده است که به عنوان داده‌ی `null` تلقی می‌شوند.

- علاوه بر داده‌های `null` موجود، داده‌هایی که ثبت نشده را به صورت رندم حذف کنید.
- سپس روش‌هایی برای جایگزینی داده‌ها بنویسید و داده‌ها را تکمیل کنید.
- در این پژوهش، داده‌ها به صورت سری‌های زمانی در نظر گرفته شده‌اند. برای تقسیم داده‌ها از روش تقسیم زمانی (Temporal Split) استفاده می‌کنیم.این روش تضمین می‌کند که داده‌های اعتبارسنجی و آزمون به صورت زمانی پس از داده‌های آموزشی قرار گیرند (و نه به صورت پراکنده در طول سری زمانی).

Temporal Split چیست؟

یک روش استاندارد برای تقسیم داده‌های سری زمانی به بخش‌های آموزش و آزمون است که در آن ترتیب زمانی داده‌ها حفظ می‌شود. برخلاف روش‌های تصادفی که ممکن است اطلاعات آینده را به مدل لو بدهند، در این روش ابتدا داده‌های قدیمی‌تر برای آموزش و سپس داده‌های جدیدتر برای ارزیابی مدل استفاده می‌شوند. این نوع تقسیم‌بندی از نشت اطلاعات جلوگیری می‌کند و شرایط واقعی پیش‌بینی را بهتر شبیه‌سازی می‌کند، به‌ویژه در مسائل حساس به زمان مانند پیش‌بینی مالی، ترافیک یا آب‌وهوا. برای مطالعه بیشتر به این لینک مراجعه کنید.

در مواردی که محدوده داده‌های اعتبارسنجی کوچک باشد، از تکنیک Rolling Forecast Origin استفاده می‌شود تا مدل در بازه‌های مختلف زمانی ارزیابی گردد. در نهایت، تمامی داده‌ها قبل از استفاده نرمال‌سازی می‌شوند.

Rolling Forecast Origin چیست؟

در این رویکرد، مدل به‌صورت گام‌به‌گام با داده‌های جدید به‌روزرسانی می‌شود و در هر مرحله، پیش‌بینی برای گام بعدی انجام می‌گیرد، به‌طوری‌که مبدا پیش‌بینی در هر تکرار به جلو حرکت می‌کند. این روش ارزیابی واقع‌گرایانه‌تر از تقسیم‌بندی ایستای داده‌ها (مانند **train/test** ثابت) عمل می‌کند و شباهت بیشتری به شرایط واقعی دارد که در آن مدل‌ها به‌طور مداوم با داده‌های تازه تغذیه می‌شوند. استفاده از این متد امکان تحلیل دقیق‌تر عملکرد مدل در مواجهه با داده‌های متوالی و تغییرات زمانی را فراهم می‌سازد. برای مطالعه بیشتر به این لینک یا این لینک مراجعه کنید.

2-1. پیاده‌سازی مدل‌ها (45 نمره)

در مقاله‌ی داده شده پیش‌بینی سری زمانی توسط چهار مدل LSTM, Bi-LSTM, RNN و GRU انجام شده‌است. ضمن در نظر گرفتن میانگین مربعات خطا[1] به عنوان تابع خطا، طبق هایپرپارامترهای جدول ۴ موجود در مقاله این مدل‌ها را آموزش دهید و موارد خواسته شده را گزارش کنید.

Table 4. Hyperparameters of LSTM, GRU, and Bi-LSTM modelling

Learning Rate	0.0010
Batch Size	100
Optimizer	Adam
Epochs	50
Units	512 (LSTM & GRU); 1024 (Bi-LSTM)

- برای هر سه مدل داده شده، نتایج پیش‌بینی شده را همراه مقادیر واقعی نمایش دهید. *(۱۵ نمره)*
- ابتدا به طور مختصر در مورد معیارهای خطای R-Squared، RMSE، MAE و MAPE توضیح دهید. سپس مقادیر را گزارش کرده و نتایج را تحلیل و مقایسه کنید.

3-1. ARIMA (امتیازی - 10 نمره)

در این قسمت از سوال با مدل کلاسیک ARIMA[2] و SARIMA[3] آشنا خواهید شد. در ابتدا تفاوت این دو مدل را بیان کنید.

- مدل ARIMA پارامترهایی دارد، مفهوم ریاضی این مدل را با ذکر پارامترها شرح دهید.
- پارامترهای بهینه‌ی این مدل را بدست آورده و گزارش کنید.
- ضمن ارائه‌ی جدولی مشابه جدول شماره ۵، نتایج را با نتایج داخل مقاله مقایسه کنید.

---

[1] Mean Square Error
[2] Autoregressive Integrated Moving Average
[3] Seasonal ARIMA

پرسش 2: پیش‌بینی افکار خودکشی در رسانه‌های اجتماعی

مقدمه

هدف از این تمرین تشخیص افکار خودکشی از مجموعه داده‌های توییتر است. در مقاله پیوست شده، چندین روش یادگیری ماشین و یادگیری عمیق مورد بررسی قرار گرفته است. با نگاهی اجمالی به نتایج این مقاله می‌توان دریافت که توانایی مدل‌های مبتنی بر یادگیری عمیق در تشخیص افکار خودکشی بیشتر بوده است. به بیان دیگر، نتایج این مقاله، توانایی مدل‌های مبتنی بر یادگیری عمیق در پردازش متن را نشان می‌دهد. مجموعه داده‌ای جهت انجام این تمرین پیوست شده است که شامل متن توییت و برچسب خودکشی می‌باشد.

از بین مدل‌های بررسی شده در مقاله، شما می‌بایست مدل‌های LSTM، 2-layer LSTM و CNN+2-layer LSTM را برای تشخیص افکار خودکشی بررسی کنید. از آنجایی که برخی مؤلفه‌ها مانند نرخ آموزش، تعداد نورون‌های لایه‌ها و ... در مقاله مشخص نشده است، شما می‌توانید از مقادیر معقول برای این موارد استفاده کنید.

2-1. پیش پردازش داده (30 نمره)

در ابتدا شما لازم است تمامی پیش‌پردازش‌های گفته شده در مقاله مانند ریشه‌یابی، حذف کلمات گزارشی، حذف پیوسته، حذف علائم نگارشی و ... را روی داده انجام دهید. برای مثال متن پس از پیش‌پردازش داده، مشابه یک یا دو متنی مانند شماره‌ی ۱ و ۲ خواهد شد.

متن شماره یک:

my life is meaningless i just want to end my life so badly my life is completely empty and i dont want to have to create meaning in it creating meaning is pain how long will i hold back the urge to run my car head first into the next person coming the opposite

way when will i stop feeling jealous of tragic characters like gomer pile for the swift end they were able to bring to their lives

متن شماره دو:

life meaningless want end life badly life completely empty dont want create meaning creating meaning pain long hold back urge run car head first next person coming opposite way stop feeling jealous tragic character like gomer pile swift end able bring life

2-2. ساخت ماتریس جاسازی (10 نمره)

در این بخش همان‌طور که در مقاله ذکر شده لازم است از مدل از پیش‌آموزش دیده شده word2vec ماتریس جاسازی را بسازید. ماتریس جاسازی روشی برای تبدیل داده‌های پیچیده (مثل کلمات، تصاویر یا کاربران) به بردارهای عددی فشرده است که معنا و روابط بین آنها را حفظ می‌کند و نسبت به روش‌های ساده مانند OneHot Enconding بسیار کم حجم تر و دقیق تر است. دلیل استفاده و ویژگی‌های این ماتریس را در گزارش به صورت مختصر توضیح دهید.

Word2Vec یک الگوریتم در حوزه پردازش زبان طبیعی است که کلمات را به بردارهای عددی با ابعاد ثابت تبدیل می‌کند، به‌طوری‌که شباهت معنایی بین کلمات در فضاهای عددی حفظ شود؛ برای مثال، کلماتی که از نظر معنا به هم نزدیک‌اند مانند "پزشک" و "پرستار"، بردارهایی نزدیک به هم خواهند داشت. Word2Vec با یادگیری روابط معنایی میان کلمات، پایه بسیاری از کاربردهای مدرن NLP مانند ترجمه ماشینی، تحلیل احساسات و پاسخ به پرسش‌ها را فراهم کرده است.

امتیازی(5 نمره): علاوه بر استفاده از مدل از پیش‌آموزش‌دیده‌ی Word2Vec، سایر روش‌های رایج برای تولید بردارهای معنایی کلمات (Embedding) را نیز بررسی کرده و در پروژه‌ی خود به کار بگیرید.

3-2. آموزش مدل‌های یادگیری عمیق (50 نمره)

در این بخش باید سه مدل LSTM، 2-layer LSTM و CNN + 2-layer LSTM را با استفاده از داده‌های پیش‌پردازش شده آموزش دهید.

در این بخش باید سه مدل مختلف شامل **LSTM**، **2-layer LSTM** و **ترکیب CNN + 2-layer LSTM** را با استفاده از داده‌های پیش‌پردازش‌شده طراحی و آموزش دهید. در ادامه، معرفی مختصری از هر مدل به همراه منبعی برای مطالعه بیشتر ارائه شده است:

- **(LSTM) Long Short-Term Memory**

یک نوع پیشرفته از شبکه‌های عصبی بازگشتی (RNN) است که برای یادگیری وابستگی‌های بلندمدت در داده‌های ترتیبی طراحی شده. LSTM با داشتن ساختار سلول حافظه و دروازه‌های ورودی/خروجی، می‌تواند اطلاعات مربوط به ورودی‌های قبلی را به‌خوبی حفظ و مدیریت کند.

- **layer LSTM-2**

این مدل شامل دو لایه LSTM متوالی است که خروجی لایه اول به عنوان ورودی لایه دوم استفاده می‌شود. این ساختار عمیق‌تر، توانایی مدل را در درک و یادگیری الگوهای پیچیده‌تر افزایش می‌دهد، به‌ویژه در مسائل مرتبط با متن یا توالی‌های طولانی.

- **CNN + 2-layer LSTM**

مدلی ترکیبی است که در آن ابتدا یک شبکه CNN برای استخراج ویژگی‌های محلی از داده‌های متنی به کار می‌رود و سپس خروجی آن به یک LSTM دولایه برای تحلیل وابستگی‌های زمانی داده می‌شود. این ساختار برای متونی که هم وابستگی محلی (مانند عبارات یا کلمات کلیدی) و هم ساختار ترتیبی دارند، بسیار مناسب است.

نکته: ممکن است در طراحی مدل CNN + 2-layer LSTM با خطایی مواجه شوید که ورودی لایه LSTM باید سه بعدی باشد. برای رفع این خطا می‌توانید از لایه Reshape استفاده کنید. البته راه‌حل‌های دیگری نیز وجود دارد. برای مطالعه بیشتر به منبع اول، منبع دوم یا منبع سوم مراجعه کنید.

4-2. نتایج و تحلیل آن (10 نمره)

در این بخش لازم است عملکرد سه مدل آموزش‌داده‌شده بررسی کنید ، برای این منظور، نمودارهای تغییرات دقت و خطا در طول دوره‌های آموزش (epochs) رسم شده و باید از آن‌ها برای تحلیل روند یادگیری، بررسی همگرایی مدل، و شناسایی مواردی نظیر overfitting یا underfitting استفاده شود. در گزارش خود توضیح دهید که کدام مدل به دقت بالاتری در داده‌های اعتبارسنجی دست یافته، کدام سریع‌تر همگرا شده و رفتار هر مدل در برابر داده‌های جدید چگونه بوده است.