به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین چهارم

پرسش ۱	نام دستیار طراح	حمید نعمتی
	رایانامه	Hamid.nemati@ut.ac.ir
پرسش ۲	نام دستیار طراح	علی یعقوبیان
	رایانامه	Aliyaghoubian@ut.ac.ir
	مهلت ارسال پاسخ	۱۴۰۲.۱۰.۰۱

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 ... هم‌گروه باشید)
- **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.**

- نحوه محاسبه تاخیر به این شکل است: پس از پایان رسیدن مهلت ارسال گزارش، حداکثر تا یک هفته امکان ارسال با تاخیر (مطابق با سیاست جریمه‌ای که در اعلان‌های ایلرن گفته شده) 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. پیشبینی سری زمانی

پیش‌بینی را می‌توان به دو دسته regression و classification تقسیم کرد. در regression مقدار عددی روزهای بعدی برای یک سری زمانی پیش‌بینی می‌شود. در classification صعودی یا نزولی بودن روند سری زمانی برای چند روز آینده پیش‌بینی می‌شود. در این سوال شما با حالت regression آشنا خواهید شد و بخشی از این مقاله را پیاده سازی می‌کنید.

۱-۱. دانلود داده ها

(۱۰ نمره)

در این قسمت داده ها را به کمک کتابخانه yahoo finance دانلود می‌کنیم.

برای اینکار ابتدا لیست اسامی تمام سهم‌های موجود در SP500 را دریافت کنید.

```
table = pd.read_html('https://en.wikipedia.org/wiki/List_of_S%26P_500_companies')
table[0]
```

	Symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Headquarters Location	Date first added	CIK	Founded
0	MMM	3M	reports	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1976-08-09	66740	1902
1	AOS	A. O. Smith	reports	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	91142	1916
2	ABT	Abbott	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	1800	1888
3	ABBV	AbbVie	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	1551152	2013 (1888)
4	ABMD	Abiomed	reports	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	815094	1981
...	...	...	...	...	...	...	...	...	...
498	YUM	Yum! Brands	reports	Consumer Discretionary	Restaurants	Louisville, Kentucky	1997-10-06	1041061	1997
499	ZBRA	Zebra Technologies	reports	Information Technology	Electronic Equipment & Instruments	Lincolnshire, Illinois	2019-12-23	877212	1969
500	ZBH	Zimmer Biomet	reports	Health Care	Health Care Equipment	Warsaw, Indiana	2001-08-07	1136869	1927
501	ZION	Zions Bancorporation	reports	Financials	Regional Banks	Salt Lake City, Utah	2001-06-22	109380	1873
502	ZTS	Zoetis	reports	Health Care	Pharmaceuticals	Parsippany, New Jersey	2013-06-21	1555280	1952

503 rows × 9 columns

سپس سهامی که از سال ۲۰۱۰ به صورت کامل رکورد شده‌اند را به صورت زیر جدا کنید:

```
# Stocks that Have Data Available from 2010
New_table_2 = New_table[ [ datetime.strptime(dt, '%Y-%m-%d') < datetime(2010, 1, 1) for dt in New_table['Date first added'] ] ]
New_table_2
```

	Symbol	Security	SEC filings	GICS Sector	GICS Sub-Industry	Headquarters Location	Date first added	CIK	Founded
0	MMM	3M	reports	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1976-08-09	66740	1902
2	ABT	Abbott	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	1800	1888
7	ADM	ADM	reports	Consumer Staples	Agricultural Products	Chicago, Illinois	1981-07-29	7084	1902
8	ADBE	Adobe Inc.	reports	Information Technology	Application Software	San Jose, California	1997-05-05	796343	1982
9	ADP	ADP	reports	Information Technology	Data Processing & Outsourced Services	Roseland, New Jersey	1981-03-31	8670	1949
...	...	...	...	...	...	...	...	...	...
495	WYNN	Wynn Resorts	reports	Consumer Discretionary	Casinos & Gaming	Paradise, Nevada	2008-11-14	1174922	2002
496	XEL	Xcel Energy	reports	Utilities	Multi-Utilities	Minneapolis, Minnesota	1957-03-04	72903	1909
498	YUM	Yum! Brands	reports	Consumer Discretionary	Restaurants	Louisville, Kentucky	1997-10-06	1041061	1997
500	ZBH	Zimmer Biomet	reports	Health Care	Health Care Equipment	Warsaw, Indiana	2001-08-07	1136869	1927
501	ZION	Zions Bancorporation	reports	Financials	Regional Banks	Salt Lake City, Utah	2001-06-22	109380	1873

254 rows × 9 columns

حال داده ها را به کمک yahoo finance دانلود کنید. (توجه کنید که yahoo finance تنها منبع موجود برای داده نیست و در دانلود داده برای تایم فریم‌های پایین تر مثل دقیقه ای محدودیت زمانی ایجاد می‌کند. اگر مایل هستید می‌توانید از کتابخانه historic crypto استفاده کنید و روی بازار رمز ارز مابقی مراحل را انجام دهید.)

```
Raw_Price_For_My_Tickers = yf.download(tickers=My_Tickers, start="2010-01-04", interval="1d", group_by = 'ticker')
Raw_Price_For_My_Tickers
```

```
[**********100%**********]  254 of 254 completed

1 Failed download:
- BF.B: No data found for this date range, symbol may be delisted
```

	EMR						HIG				...	MDT				BALL					
	Open	High	Low	Close	Adj Close	Volume	Open	High	Low	Close	...	Low	Close	Adj Close	Volume	Open	High	Low	Close	Adj Close	Volume
Date																					
2010-01-04	43.150002	43.400002	42.730000	43.349998	29.884455	3781000	23.709999	23.959999	23.510000	23.860001	...	43.799999	43.900002	33.006214	6109400	12.965000	13.137500	12.955000	13.122500	11.762474	2468400
2010-01-05	43.380001	43.430000	43.070000	43.419998	29.932705	2707500	23.730000	25.750000	23.670000	25.690001	...	43.750000	44.799999	33.682880	6568600	13.082500	13.095000	12.937500	12.975000	11.630261	3053600
2010-01-06	43.250000	43.830002	43.160000	43.810001	30.201557	4314700	25.510000	26.379999	25.090000	26.120001	...	44.419998	45.330002	34.238037	7470700	12.980000	13.030000	12.910000	13.000000	11.652667	1879200
2010-01-07	43.599998	43.840000	43.290001	43.810001	30.201557	3085700	26.000000	26.719999	25.820000	26.520000	...	45.020000	45.750000	34.555271	6895200	12.935000	12.987500	12.872500	12.970000	11.625777	2105600
2010-01-08	43.730000	44.090000	43.419998	44.060001	30.373894	3477500	26.389999	26.570000	25.770000	26.219999	...	45.450001	45.990002	34.736538	4947200	12.877500	13.060000	12.832500	13.017500	11.668359	1227600
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2022-09-23	74.489998	74.620003	72.870003	73.839996	73.839996	3258200	61.750000	62.090000	60.889999	61.919998	...	81.620003	82.660004	82.660004	5960400	48.040001	49.160000	47.209999	49.139999	49.139999	5646300
2022-09-26	73.529999	74.559998	72.919998	73.139999	73.139999	2668300	61.509998	62.160000	60.169998	60.630001	...	80.690002	81.330002	81.330002	7099800	48.430000	49.900002	47.700001	48.310001	48.310001	3555600
2022-09-27	74.120003	74.400002	72.410004	73.059998	73.059998	2658800	61.029999	62.070000	60.779999	61.970001	...	80.389999	80.889999	80.889999	5223000	48.880001	49.380001	47.270000	47.490002	47.490002	3411500
2022-09-28	74.019997	75.519997	73.330002	75.019997	75.019997	3851100	62.130001	63.520000	61.930000	63.060001	...	80.809998	82.589996	82.589996	6099900	48.009998	49.480000	47.529999	49.230000	49.230000	1696700
2022-09-29	74.230003	74.339996	73.050003	74.160004	74.160004	811702	62.570000	62.849998	61.900002	62.730000	...	81.449997	82.800003	82.800003	1961284	48.709999	48.720001	47.004002	47.562000	47.562000	624406

3208 rows × 1524 columns

۱-۲. کاوش در داده های سری زمانی و آشنایی با تئوری ها و کتابخانه های معروف

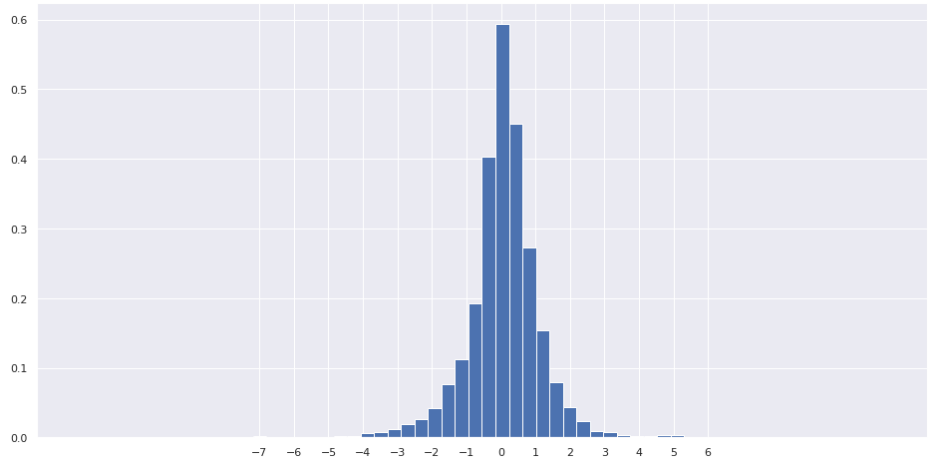
(۱۰ نمره)

در داده‌هایی که دانلود کرده‌اید ممکن است تعدادی null value موجود باشد. به دو سوال زیر پاسخ دهید و مطابق با پاسخ خود داده ها را برای مراحل بعدی پاکسازی کنید.

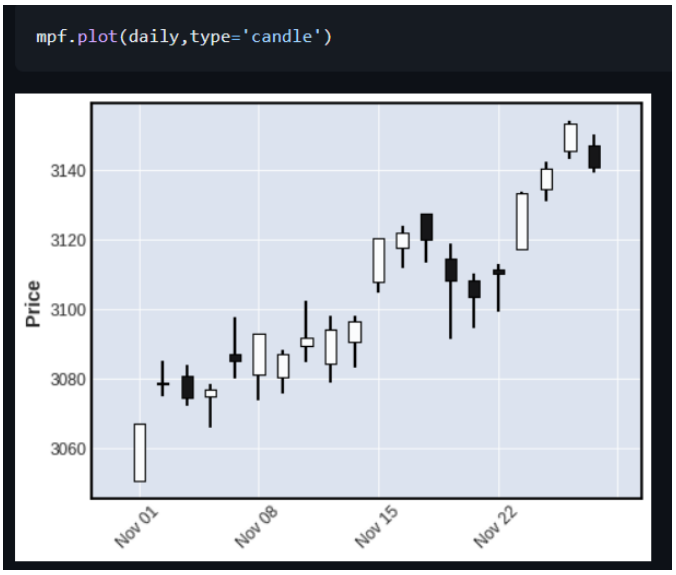
- جستوجو کنید که اگر داده های null در ابتدا قرار داشتند با چه روش های می‌توان این مشکل را حل کرد؟

- اگر در وسط یا انتهای سری زمانی داده‌های null value داشتیم به چه روشی آن را حل می‌کنند؟

حال یک سهم را به صورت تصادفی انتخاب کنید و نمودار هیستوگرام close price return را برای آن رسم کنید. نمودار شما مشابه شکل زیر خواهد بود. تفسیر کنید که این منحنی شبه نرمال price return به چه معناست؟ سعی کنید آن را به random walk theory ربط دهید (تئوری راه رفتن تصادفی ادعا می‌کند که قیمت‌ها در بازار سهام هیچ الگوی خاصی ندارند و بهترین پیش‌بینی برای قیمت روز بعد قیمت امروز است)



در نهایت به کمک کتابخانه MPL Finance داده‌ها را به صورت CandleStick نشان دهید.



۱-۳. TimeSeriesSplit

(۱۰ نمره)

نحوه‌ی انجام cross validation برای داده‌های سری زمانی تفاوت بنیادین با مسائل رایج یادگیری ماشین که تا به حال دیده‌اید دارد. در مورد چرایی این موضوع جستوجو کنید (راهنمایی: look ahead bias) و همچنین به کمک تابع زیر داده‌های خود را برای cross validation شدن آماده کنید.

(از این مرحله به بعد فقط با این سهم‌ها کار کنید و مابقی داده را دور بریزید: ['AAPL', 'MSFT', 'AMZN', 'GOOGL', 'META'])

```
cv_n_splits = 5

tss = sklearn.model_selection.TimeSeriesSplit( n_splits=cv_n_splits )
```

۱-۴. آماده سازی ورودی و خروجی مدل

(۱۰ نمره)

ورودی مدل داده‌های روزهای پیشین است. این داده می‌تواند شامل فقط close price چند روز گذشته باشد. یا می‌تواند داده‌های open و high و low و volume را هم شامل شود. برای کاهش dimensionality معمولا فقط از close price برای آموزش مدل استفاده می‌شود.

خروجی مدل قیمت یک روز خاص در آینده (Horizon) است. برای این تمرین شما داده روز آینده را فقط پیش‌بینی می‌کنید.

در تصویر زیر مشاهده می‌کنید که X و y مدل شما چگونه آماده می‌شود.

```
y_tickers_sp500 = []
x_tickers_sp500 = []

for ticker_index, ticker_name in enumerate(My_Tickers_SP500):
    X_time_series = []
    Y_time_series = []
    for i in range(window_size, len(data_full_scaled_sp500[ticker_index])):
        X_time_series.append( data_full_scaled_sp500[ticker_index][i-window_size:i] )
        Y_time_series.append( data_full_scaled_sp500[ticker_index][i] )

    X_time_series = np.array(X_time_series)
    Y_time_series = np.array(Y_time_series)

    x_tickers_sp500.append( X_time_series[(31-window_size):] )
    y_tickers_sp500.append( Y_time_series[(31-window_size):] )

print(len(y_tickers_sp500[0]))
print(len(x_tickers_sp500[0]))
```

توجه کنید که ورودی‌های شما باید normalize شوند و به بازه 1- تا 1 یا ۰ تا ۱ بیایند. دلیل این کار وجود Activation function های مانند Tanh و sigmoid است. (همچنین توجه کنید که شما فقط اجازه دارید از داده‌های Train برای ساختن scaler خود استفاده کنید.)

در قسمت بعدی یک بار مدل را با داده‌های مقیاس نشده آموزش دهید و گزارش خود را از نحوه آموزش دیدن مدل بنویسید.

۱-۵. مدل های شبکه عصبی حافظه دار

(۵۰ نمره)

در این قسمت با LSTM و GRU و Bi-LSTM آشنا می‌شوید. در مورد هر کدام تحقیق کنید که چگونه آموزش می‌بینند و علت وجود هر Gate چیست. تفاوت و شباهت ها را هم بیان کنید.

سپس مدل‌ها را آموزش دهید و با هم مقایسه کنید و نتایج را گزارش کنید.

برای MLP و CNN و Conv-LSTM هم همین کار را تکرار کنید ولی اینبار نیاز به توضیح شیوه کارکرد مدل نمی‌باشد.

برای آموزش مدل می‌توانید از loss='mean_square_error' استفاده کنید.

برای مقایسه مدل‌ها از MAE و MSE و MAPE استفاده کنید. توضیح خلاصه‌ای از شیوه ارزیابی هر کدام از این معیارها ارائه دهید.

نتایج را تحلیل کنید و نظر خود را درباره چرایی عملکرد خوب یا ضعیف هر کدام از مدل‌ها ارائه دهید.

تصویر زیر برای راهنمای شما برای چگونگی آموزش مدل‌ها و ارائه نتایج رو ۵ سهم مشخص شده قرار داده شده است. قسمت set کردن seed برای قابل تکرار بودن نتایج کد شماست زیرا که مدل‌های شبکه عصبی پارامترهای زیادی دارند که که به صورت تصادفی initialize می‌شوند و با هر بار اجرا نتایج متفاوتی خواهند داد.

```
for ticker_index, ticker_name in enumerate(My_Tickers_SP500):

    full_data = data_full_scaled_sp500 [ticker_index]

    print(ticker_name, len(full_data))

    # Time Series Split
    for j, (train_index, test_index) in enumerate(tss.split(full_data)):

        # Different Seeds
        for new_seed in range(5):

            seed_value = new_seed
            random.seed(seed_value)
            np.random.seed(seed_value)
            tf.random.set_seed(seed_value)
            keras.utils.set_random_seed(seed_value)

            number_of_runs += 1

            train_data = full_data[train_index]
            test_data  = full_data[test_index ]
```

۱-۶. Naïve Forecast

(۱۰ نمره)

در قسمت دوم این سوال به random walk اشاره شد که بهترین پیش‌بینی را برای روز بعد قیمت امروز می‌داند. به این روش naïve forecast می‌گویند. سعی کنید آن را پیاده سازی کنید (استفاده از کتابخانه‌ها مجاز است) و معیارهایی که در قسمت قبلی گزارش کردید برای این روش هم محاسبه کنید. سپس نتایج را تحلیل کنید.

پرسش ۲. پیش‌بینی افکار خودکشی در رسانه‌های اجتماعی

هدف از این تمرین تشخیص افکار خودکشی از مجموعه داده‌های توییتر[۱] است. در مقاله پیوست شده، چندین روش یادگیری ماشین و یادگیری عمیق مورد بررسی قرار گرفته است. با نگاه 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 مولفه‌ها مانند نرخ آموزش[۲]، تعداد نورون‌های لایه‌ها و ... در مقاله مشخص نشده است، شما می‌توانید از مقادیر معقول برای این موارد استفاده کنید.

۲-۱. پیش‌پردازش داده

(۳۰ نمره)

در ابتدا شما لازم است تمامی پیش‌پردازش‌های گفته شده در مقاله مانند ریشه‌یابی[۳]، حذف کلمات کم‌ارزش، حذف پسوند، حذف علائم نگارشی و ... را روی داده انجام دهید. برای مثال متن پس از پیش‌پردازش داده، متن شماره یک به متنی مانند شماره دو تبدیل شود.

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

---

۱ Twitter

۲ Learning rate

۳ Lemmatization

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

۲-۲. ساخت ماتریس جاسازی[۱]

(۱۰ نمره)

در این بخش همان‌طور که در مقاله ذکر شده لازم است از مدل از پیش‌آموزش دیده شده word2vec ماتریس جاسازی را بسازید. دلیل استفاده و ویژگی‌های این ماتریس را در گزارش به صورت مختصر توضیح دهید.

۲-۳. آموزش مدل‌های یادگیری عمیق

(۵۰ 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.

(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.

[۱] Embedding

۲-۴. مقایسه نتایج

(۱۰ نمره)

در این بخش نتایج سه مدل آموزش داده شده را مقایسه کنید و استدلال خود را جهت توجیه نتایج در گزارش بنویسید.