پروژه‌ی درس علوم شبکه پیشرفته

پیش‌بینی شیوع کووید-۱۹ با استفاده از داده‌های تردد افراد

مهدی صادقی

استاد درس: دکتر حشمتی

۱. خلاصه:

بیماری کووید-۱۹ از زمان شیوع اولیه در ووهان چین تا به حال به بیش از ۱۰۰ کشور شیوع پیدا کرده است. این سرعت شیوع و تعداد بالای افراد فوت شده، باعث شده استراتژی‌های مختلفی که مهمترین آن فاصله‌گزاری اجتماعی بوده برای کاهش نرخ شیوع استفاده شود اما به دلایل اقتصادی امکان تعطیلی و محدودیت گسترده برای طولانی مدت وجود ندارد، به همین دلیل سعی در آن است که اجتماع را تنها در خطرناک‌ترین مکان‌ها را کاهش دهیم. در این تحقیق سعی کردیم با استفاده از داده‌های مسیریابی افراد به مکان‌های مختلف و ترکیب آن با روش SIR، میزان شیوع را پیش‌بینی کنیم تا با استفاده از آن میزان تاثیر هر کدام از این مکان‌ها در شیوع را به شکل مستقل به دست آوریم.

۲. پیشینه‌ی تحقیق:

در طول تاریخ پاندمی‌های گوناگون جهان را تحت تاثیر قرار داده است. کرونا ویروس ۲۰۱۹ (کووید ۱۹) در ابتدا در شهر ووهان چین مشاهده شد اما در مدتی کوتاه و به سرعت در کل چین و جهان منتشر شد به طوری که در ۱۱ مارچ سازمان جهانی بهداشت شرایط پاندمی جهانی اعلام کرد و کرونا به تهدید شماره یک در بیش از ۱۵۰ کشور تبدیل شده است. این ویروس نه تنها بر سلامت جهان تاثیر گذاشته بلکه مشکلات زیادی هم در بخش‌های اقتصادی ایجاد کرده است.

از آنجایی که انتقال بیماری از طریق ارتباط بیمار با فرد سالم صورت می‌گیرد، کشورهای مختلف سعی کردند با تعطیلی‌های گسترده و یا محدودیت‌های دیگر میزان شیوع را کاهش دهند اما به دلایل عمدتا اقتصادی اعمال این محدودیت‌ها برای طولانی مدت و در همه جا امکان‌پذیر نبوده، و کشوری مانند ایران که در ابتدا با تعطیلی‌های سراسری در کاهش بیماری موفق بوده، پس از بازگشایی‌ها درگیر موج دیگری از بیماری شده است. به همین دلیل مقامات کشورها سعی دارند با درک رفتار ویروس و شناسایی مکان‌های خطرناک‌تر (نسبت به سایر مکان‌ها) با تعطیلی‌های موردی نرخ شیوع را کاهش دهند. برای شناسایی و پیش‌بینی شیوع بیماری مدل‌های مختلفی پیشنهاد شده است که می‌توان آن‌ها را در یکی از ۴ دسته‌ی زیر قرارداد (Mahalle et al., n.d.):

- مدل‌های مبتنی بر داده‌های بزرگ
- مدل‌های مبتنی بر داده‌های بدست آمده از شبکه‌های اجتماعی
- مدل‌های آماری/ریاضیاتی
- مدل‌های مبتنی بر یادگیری ماشین

البته ممکن است یک مدل در چند دسته‌ی مختلف قرار گیرد. در اینجا ما سعی می‌کنیم پیشینه‌ای از دو دسته از این مدل‌ها را ارائه دهیم. مدل‌هایی که از یکی از مشتقات مدل SIR استفاده می‌کنند و مدل‌هایی که با استفاده از داده‌های رفت و آمد مردم به پیش‌بینی شیوع می‌پردازند.

۲.۱. مدل SIR و مشتقات آن:

مدل SIR و مشتقات آن زیر مجموعه‌ای از مدل‌های اپیدمیک هستند که تعداد افراد بیمار و فوت شده توسط یک بیماری واگیردار را در طول زمان و با استفاده از روابط ریاضی نشان می‌دهند. در این مدل‌ها کل جمعیت مورد بررسی به چند بخش جدا تقسیم می‌شوند و با استفاده از روابط ریاضی انتقال افراد میان این بخش‌ها مدل می‌شود. برای مثال در مدل SIR استاندارد، کل جمعیت یک شهر به سه بخش مستعد، بیمار و بهبود یافته تقسیم می‌شود. در ابتدا اکثر مردم در بخش مستعد قرار دارند و تعداد محدودی هم در بخش بیمار. به مرور زمان افرادی که در بخش بیمار قرار دارند، بیماری را به افراد مستعد انتقال می‌دهند. همچنین افراد بیمار پس از مدتی یا فوت میکنند و یا بهبود میابند که در هر دو حالت به بخش بهبودیافته منتقل می‌شوند. برای مدل‌سازی ریاضی SIR در طول زمان، برای انتقال از هر بخش به بخش دیگر یک نرخ انتقال در نظر گرفته می‌شود که نشان می‌دهد در هر واحد زمانی چه تعداد از افراد از بخش اول به بخش دوم انتقال میابند.

بر پایه‌ی همین مدل SIR مدل‌های دیگری که تعداد و نوع بخش‌های مختلفی دارند معرفی شده‌اند. برای مثال در بیماری‌هایی که فرد آلوده شده است اما نمی‌تواند بیماری را به فرد دیگری انتقال دهد، از مدل SEIR استفاده می‌کنند که یک مرحله‌ی Exposed بین مستعد و بیمار اضافه می‌کند.

از SIR و مشتقاتش برای مدل کردن کرونا هم استفاده شده است. برای مثال در (Wangping et al. 2020) با استفاده از مدل SIR و استفاده از نرخ انتقال متغیر بجای نرخ انتقال ثابت شیوع کووید ۱۹ را در ایتالیا پیش‌بینی کرده است. همچنین در (Dye 2003) برای بیماری سارس که همخوانواده‌ی کرونا است یک مدل بر پایه‌ی SEIR پیشنهاد شده است.

۲.۲. مدل بر پایه‌ی رفت و آمدهای مردم

برخی از مدل‌های پیشنهاد داده شده برای پیش‌بینی کرونا از داده‌های رفت و آمدهای مردم برای پیش‌بینی استفاده می‌کنند. برای مثال در (Dye 2003; Siwiak, Szczesny, and Siwiak, n.d.) ابتدا یک مدل اپیدمی ۷ مرحله‌ای مشتق شده از SIR طراحی شده است و آن را به یک نرم‌افزار به نام GLEAMviz داده‌اند. این نرم‌افزار با استفاده از داده‌های رفت و آمد افراد بین کشورها (داده‌های خطوط هوایی) و همچنین رفت و آمد افراد درون یک شهر در هر لحظه پارامترهای مدل اپیدمی ورودی خود را تعیین می‌کند (Balcan et al. 2010). همچنین در (Balcan et al. 2010; Hossain et al., n.d.) سعی شده با استفاده از داده‌های مربوط به رفت و آمد هواپیماها و ترکیب آن با یک مدل SIR ساده، تاثیر کاهش رفت و آمد بر روی شیوع بیماری را مدل کند.

۳. روش کار:

در این تحقیق سعی داریم با استفاده از داده‌هایی که توسط گوگل در مورد تردد مردم در مکان‌های عمومی منتشر شده است نرخ انتشار را در یک مدل SIR ساده تخمین بزنیم و بررسی کنیم که آیا می‌توان پیش‌بینی بهتر از مدل SIR با نرخ انتشار ساده ارائه داد.

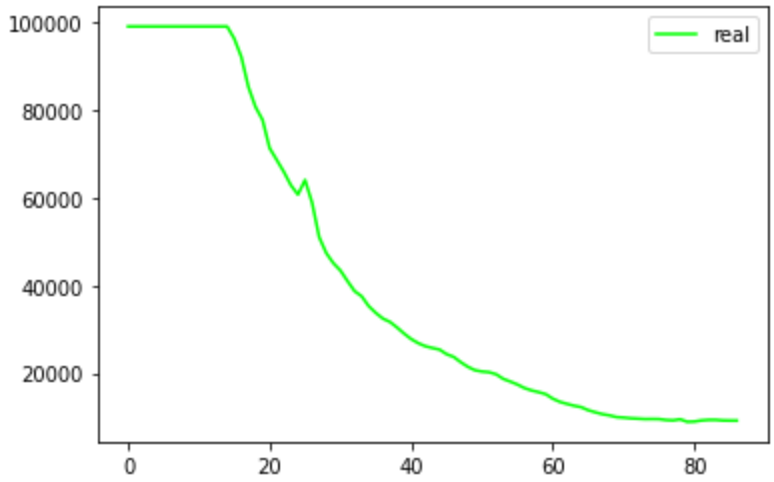
۳.۱. آماده‌سازی داده‌های تردد

برای بررسی میزان تردد افراد در مکان‌های مختلف از داده‌های تردد انجمنی منتشر شده توسط گوگل استفاده شده است. این داده‌ها با استفاده از مسیریابی‌های افراد مختلف در برنامه‌ی نقشه‌ی گوگل بدست آمده. البته داده‌ها به صورت میزان درصد افزایش یا کاهش نسبت به یک عدد پایه است که بسته به شهر متفاوت است. این قضیه باعث عدم هماهنگی بین داده‌های شهرها شده است و نمی‌توان از داده‌های شهرهای مختلف همزمان استفاده کرد.
باید توجه داشت که منطقا تنها تردد امروز مستقیما بر تعداد افراد مبتلا شده‌ی امروز تاثیر نمی‌گذارد و ما باید تردد چند روز را با هم بررسی کرد. به همین منظور داده‌های تردد در یک پنجره‌ی ۷ روزه میانگین گرفته شده‌اند. این کار باعث می‌شود افزایش تردد در یک مکان خاص (مثلا مراکز فروشگاهی) و داده‌ی پرت باعث خراب شدن الگوریتم نشوند.

```
def get_averaged_mobility_data(state_data, average_period=7):
    m = []
    for i, row in state_data.iterrows():
        end = i
        start = max(end - 7, 0)
        window = state_data.loc[start:end]
        m.append({
            'date': row['date'],
            'retail_and_recreation_percent_change_from_baseline':
window['retail_and_recreation_percent_change_from_baseline'].mean(),
            'grocery_and_pharmacy_percent_change_from_baseline':
window['grocery_and_pharmacy_percent_change_from_baseline'].mean(),
            'parks_percent_change_from_baseline':
window['parks_percent_change_from_baseline'].mean(),
            'transit_stations_percent_change_from_baseline':
window['transit_stations_percent_change_from_baseline'].mean(),
            'workplaces_percent_change_from_baseline':
window['workplaces_percent_change_from_baseline'].mean(),
            'residential_percent_change_from_baseline':
window['residential_percent_change_from_baseline'].mean()
        })

    return pd.DataFrame(m)
```

۳.۲. آمادهسازی دادههای کرونا

در میان کشورهای درگیر، بیشترین آمار مبتلایان و فوت شدگان، مربوط به کشور آمریکا بوده است. سیستم درمانی این کشور تا به حال بیش از ۵ میلیون مبتلا را تا به حال شناسایی کرده است که ۱۶۰ هزار نفر آنها فوت شدهاند که این آمار به تنهایی بیش از ۲۰ درصد کل موارد گزارش شده در جهان است. البته تعداد تستهای انجام شده در این کشور هم بسیار بیشتر از سایر کشورهاست که این خود در نتایج بدست آمده بسیار تاثیرگذار است. این تعداد بالای تست و کیفیت بالای دادههای تردد به تفکیک استان باعث شد که تصمیم بگیریم بر روی دادههای آمریکا و استانهایش کار کنیم. به همین منظور از دادههای دانشگاه جان هاپکینز استفاده کردیم. نکتهی مهم این است که در مدل SIR در بخش بیماران منظور تنها بیماران جدید نیست و بیمارانی که روزهای قبل تشخیص داده شدهاند و هنوز درمان نشدهاند هم شامل میشود. به همین منظور با فرض این که افرادی که کرونا میگیرند به صورت میانگین پس از ۱۴ روز یا میمیرند یا درمان میشوند، به همین خاطر باید دادهها به صورتی پردازش شوند که در هر لحظه تخمینی از تعداد بیماران فعال جامعه داشته باشیم.



پس از آن نرخ انتقال ایدهآل β برای گذار از مستعد به بیمار در هر لحظه با استفاده از فرمول زیر بدست میآوریم:

$$\beta = \frac{population * newcases}{S * activecases}$$

که در آن S تعداد افراد مستعد است.

```
def get_state_covid_data(state_name, population):
    raw_data = pd.read_csv('../data/covid19/us.csv')
    state_data = raw_data[raw_data['state']==state_name].copy()
    state_data['new_cases'] = state_data['total_confirmed'].diff()
    state_data['active_cases'] = state_data['total_confirmed'].diff(14)
    state_data['infection_rate'] = (population * state_data['new_cases']) /
state_data['active_cases']
    state_data = state_data.fillna(method='bfill')
```

```
    sus = population - raw_data['total_confirmed'][0]

    m = []
    for i, row in state_data.iterrows():
        sus = sus - row['new_cases']
        m.append({
            'date': row['date'],
            'active_cases': row['active_cases'],
            'new_cases': row['new_cases'],
            'infection_rate': row['infection_rate'] / sus,
        })

    return pd.DataFrame(m)
```

۳.۳. الگوریتم

پس از اینکه داده‌ها پاکسازی شد، با استفاده از الگوریتم‌های یادگیری ماشین مختلف سعی کردیم یک مدل بین داده‌های ترددی میانگین گرفته شده و نرخ انتقال‌های بدست آمده در مرحله‌ی قبل بدست آوریم. در بین الگوریتم‌های یادگیری ماشین تست شده بهترین نتیجه مربوط به SVM با کرنل ROF بود.

```
def prediction_svm(x, y):
    model = SVR(kernel='rbf')
    model.fit(x, y)
    return model.predict(x)
```

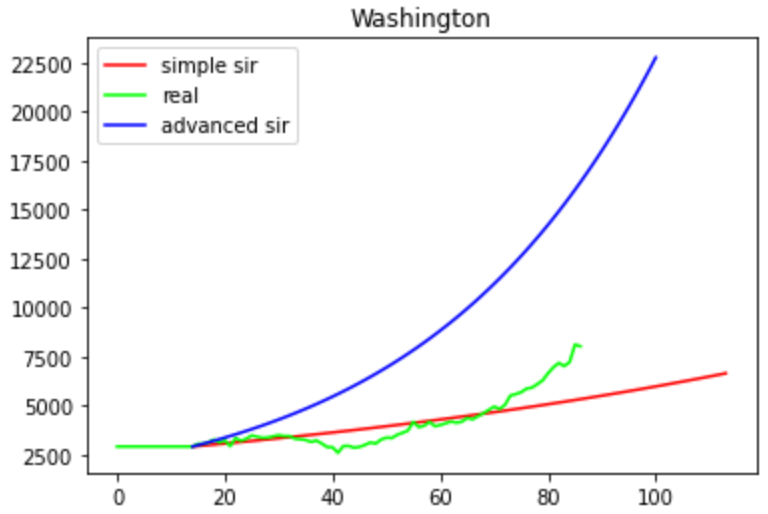
پس از آن این مدل را در یک مدل SIR استفاده می‌کنیم به این صورت که همان SIR ساده را می‌نویسیم اما بجای قرار دادن نرخ انتقال به صورت ثابت از محتوای تردد را به این مدل پاس داده و نرخ انتقال خروجی را استفاده می‌کند.

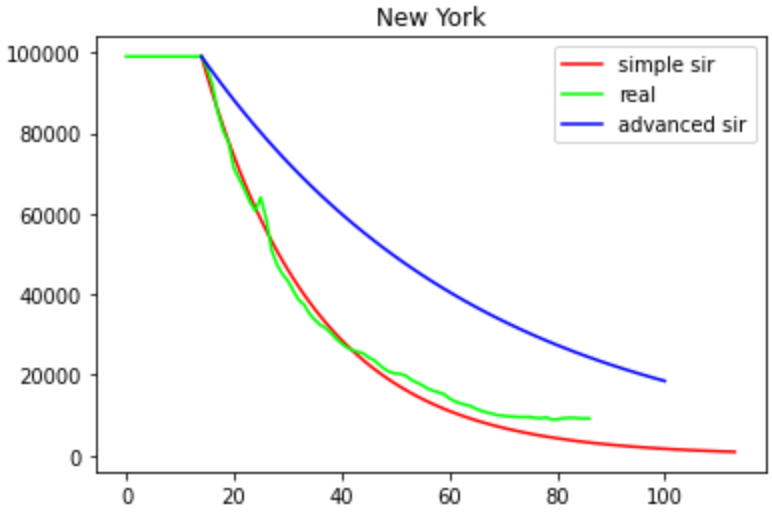
```
def adv_sir(population, initial_infected, initial_dead, adv_beta, gamma,
rounds=100):
    index = [0]
    susceptiple = [population - initial_infected]
    infected = [initial_infected]
    dead = [initial_dead]

    for t in range(1, rounds):
        index.append(t)
        new_infected = adv_beta[t-1] * (susceptiple[t-1]/population) *
infected[t-1]
        new_dead = gamma * infected[t-1]
```

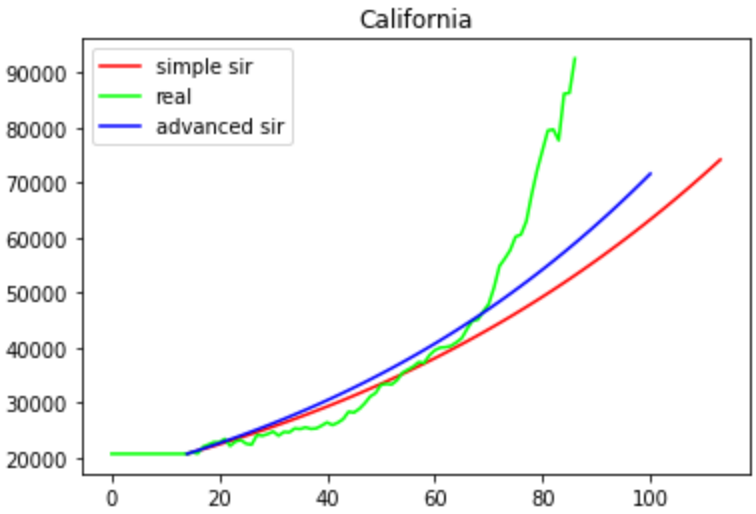
```
        susceptiple.append(susceptiple[t-1] - new_infected)
        infected.append(infected[t-1] + new_infected - new_dead)
        dead.append(dead[t-1]+new_dead)

    return pd.DataFrame({'date_number': index, 'susceptiple': susceptiple,
'infected': infected, 'removed': dead}, index=index)
```

۴. نتایج:

برای استان نیویورک، کالیفرنیا و واشنگتن فرایند بالا انجام شد. همچنین یک مدل SIR ساده با استفاده از نرخ ثابت (برای هر شهر متفاوت است) هم پیادهسازی و اجرا شده است. در زیر نمودارهای این اجراها آورده شده است.







همانطور که از نمودارها پیداست، در تمام موارد بجز کالیفرنیا SIR ساده نتایج بهتری نسبت به مدل SIR با کمک داده‌های تردد داده است. این مورد می‌تواند به دلیل کم بودن میزان داده‌ها باشد. تعداد داده‌های ترددی که در اختیار ما است حدود ۱۴۰ روز است که برای آموزش مدل کافی نیست. همچنین کیفیت پایین داده‌های ترددی هم می‌تواند دلیل دیگر این نتایج باشد.

۵. منابع:

Balcan, Duygu, Bruno Gonçalves, Hao Hu, José J. Ramasco, Vittoria Colizza, and Alessandro Vespignani. 2010. "Modeling the Spatial Spread of Infectious Diseases: The GLobal Epidemic and Mobility Computational Model." *Journal of Computational Science* 1 (3): 132–45.

Dye, C. 2003. "EPIDEMIOLOGY: Modeling the SARS Epidemic." *Science*. https://doi.org/10.1126/science.1086925.

Hossain, M. Pear, M. Pear Hossain, Alvin Junus, Xiaolin Zhu, Pengfei Jia, Tzai-Hung Wen, Dirk Pfeiffer, and Hsiang-Yu Yuan. n.d. "The Effects of Border Control and Quarantine Measures on Global Spread of COVID-19." https://doi.org/10.1101/2020.03.13.20035261.

Mahalle, Parikshit, Asmita B. Kalamkar, Nilanjan Dey, Jyotismita Chaki, Aboul Ella Hassanien, and Gitanjali R. Shinde. n.d. "Forecasting Models for Coronavirus (COVID-19): A Survey of the State-of-the-Art." https://doi.org/10.36227/techrxiv.12101547.v1.

Siwiak, Marlena M., Pawel Szczesny, and Marian P. Siwiak. n.d. "From a Single Host to Global Spread. The Global Mobility Based Modelling of the COVID-19 Pandemic Implies Higher Infection and Lower Detection Rates than Current Estimates." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3562477.

Wangping, Jia, Han Ke, Song Yang, Cao Wenzhe, Wang Shengshu, Yang Shanshan, Wang Jianwei, et al. 2020. "Extended SIR Prediction of the Epidemics Trend of COVID-19 in Italy and Compared With Hunan, China." *Frontiers of Medicine* 7 (May): 169.