知能情報実験 III（データマイニング班）国内における新型コロナウイルスの感染者予測

グループメンバー
185719D 上田瑞宜, 185725J 識名俊希, 185757G 鈴木海土, 185768B 佐藤優

提出日：2020 年 8 月 13 日

目次

概要

概要本文書は知能情報実験 III（データマイニング班）におけるレポートのテンプレートとして用意したものである。一般的な実験レポートに関する補足と共に、データマイニング班における実験レポートに求められる内容を確認するために用意した。ここに書いてある事柄は全てを必須とするわけではなく、適宜取捨選択や追加編集してもらって構わないが、実験報告書としての位置づけを忘れずに利用すること。

0.1　実験の目的と達成目標

知能情報実験 III は、情報工学分野のより専門的な知識を理解・習得することを目的として、半年間でシステムの開発やデータ解析等に取り組む実施される。その中の一つデータマイニング班においては機械学習外観ならびにその応用を通し、対象問題への理解、特徴量抽出等の前処理、バージョン管理やデバッグ・テスト等を含む仕様が定まっていない状況下における開発方法、コード解説や実験再現のためのドキュメント作成等の習得を目指す。

0.2　テーマ

本グループでは、新規コロナウィルス感染者を予測することを対象問題として設定した。これは、新型コロナウィルスの 1 週間や 1 ヶ月先の近未来の新規感染者数を予測することを目的として設定された。
このテーマを設定した理由は、近未来における新規感染者を予測することでコロナウィルスの傾向を掴むのと同時に第二波や第三波を予期することで、被害を少なくしたいと考えた。

1　実験方法

1.1　実験目的

新規コロナウィルス感染者数を正確に予測できるか検証する。

1.2　データセット構築

日本国内におけるコロナ感染者数、死者数、入院者数などが含まれているデータを探す。
実際に個人でコロナ感染者予測を行っているサイトで使用しているデータや国などが出しているデータを使用。
1 日ごとの新規感染者数や総感染者数、入院数、死者数など含まれているデータによって使い分けて使用。 SARIMA モデルの場合は新規感染者数データを使用した。
(人口密度や人の移動率を NTTdocomo が提供しているデータを使用)。

新規感染者数：`https://github.com/kaz-ogiwara/covid19`
都市間による人口密度：`https://www.nttdocomo.co.jp/utility/demographic_analytics/20200509.html` 予測を行っているサイト：https://signate.jp/competitions/272

1.3　モデル選定

1. SARIMA モデル
2. SEIR モデル
3. 仮設モデル

SARIMA モデル
SARIMA モデルとは ARIMA モデルに季節的な周期パターンを加えたモデルであり、非時系列データに対して適用できる。

SEIR モデル
SEIR モデルとは感染症流行の数流モデルである。

1. S:感染しうるもの、免疫を持たず感染可能
2. E:感染症が潜伏期間中の者
3. I:感染者が発症しているもの、接触した感染しうるもの
4. R:感染後死亡者、もしくは免疫を獲得したもの
5. N: 全人口 S+E+I+R
6. β: 感染率
7. ϵ: 暴露後に感染書を得る率
8. γ: 除去率

から構成される。

仮設モデル

SARIMA モデルに外部的要因を考慮した結果を求めるために独自で作った仮設モデルと SARIMA モデルで得られたデータを合成して外部的要因を考慮するために用意したモデルである。

人口密度を使った検証モデル (仮設モデル)
今回立てた仮説モデルは予測する 5 日前の新宿駅の人口密度の増減から線形回帰を用いて新規感染者数を予測するものである。
線形回帰を求めるのに用いた説明変数は、5 日前の新規感染者数の移動平均、 5 日前の感染拡大以前との比較した新宿駅の人口密度のデータ、 5 日前の前日との比較した人口密度のデータを説明変数として線形回帰を行なった。

1.4　パラメータ調整

パラメーターの内の order(p,d,q) と seasonal_order(sp,sd,sq) の値の組み合わせを
総当たりで求めその中で最も AIC 値の低かったものを採用した。

2　実験結果

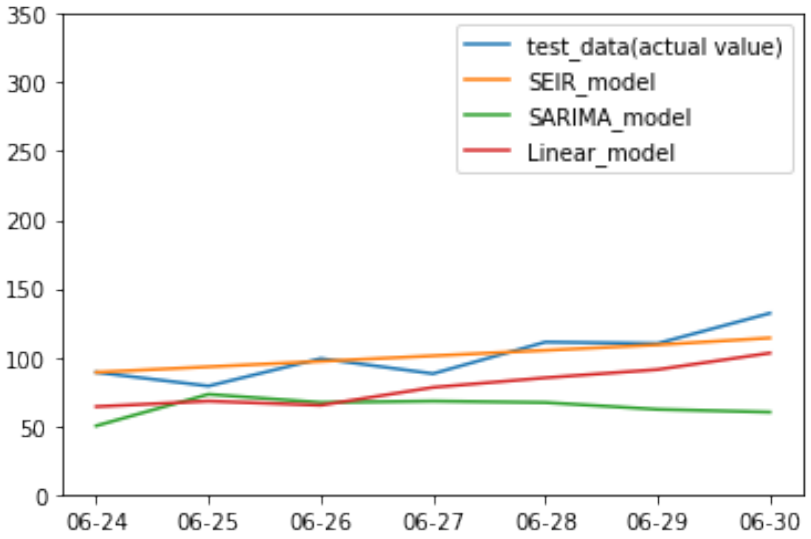

図 1　6 月 24 日 6 月 30 日までの予測結果

図 1 は 6 月 24 日から 6 月 30 日までの各モデルによる予測結果である。
縦に新規感染者数、横に日付をとっている。
図 1 から見てわかるように SEIR モデルの予測結果が最も精度が高く予測できている。

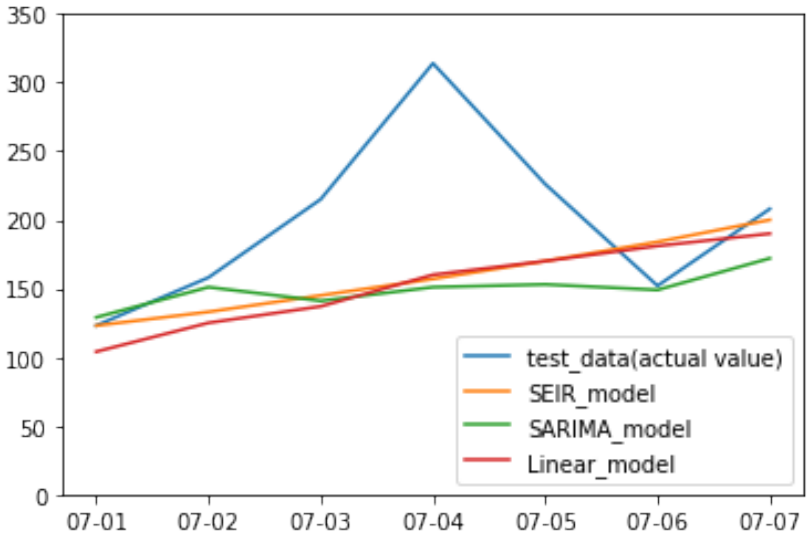

図 2　7 月 1 日 7 月 7 日までの予測結果

図 2 は 7 月 1 日から 7 月 7 日までの各モデルによる予測結果である。
縦に新規感染者数、横に日付をとっている
図 2 のようにどの予測結果も爆発的に新規感染者数が増えたときは予測できていない。

SARIMA モデルと SEIR モデルで予測した値を決定係数で評価した結果
6 月 24 日 6 月 30 日の区間：SARIMA モデル→-5.384 SEIR モデル→ 0.629
7 月 1 日 7 月 7 日の区間：SARIMA モデル→-0.618 SEIR モデル→-0.44
どちらの区間も SEIR モデルの方が 1 に近い値となり、SEIR モデルの方が新規感染者数を予測するには良い結果となった。

* 決定係数とは、標本知から求めたモデルの当てはまり度の良さを表す。(一般的に 1 に近い方が良いと言われている)

3　考察

今回は 2 つのモデルを用意しどちらのモデルが良いかを判断したが、新規感染者数を予測する上で推定期間によって決定係数の値は異なった。パラメータの調整によってより精度をあげられることもあり、どちらのモデルの方が良いのかは断言することはまだできない。これらの最適化された推定期間やパラメータを用意し検討することも重要である。第二波や PCR 検査数の違いによって値が変わるため継続的に予測することが必要である。爆発的な感染者数の増減はどちらのモデルでも予測できなかったため、爆発的な増減を予測することも考えなければならない。

4　意図していた実験計画との違い

実験計画は、最初に新規感染者数を予測する上でどのようなモデルがあるのかを調べた。そのモデルが複数個見つかり、2 つのモデルが良さそうだと判断したため、どちらのモデルがより正確に予測できているかを調べた。また、緊急事態宣言による外部的要因なども考慮して精度改善を試みた。予測する区間は 1 週間とし、実際に得られた予測データから評価モデルにより、どちらのモデルが良いか判断した。

5　まとめ

オンライン上でグループの作業を共有することは難しかった。2 つのモデルがあったため、2 つのグループで分かれて作業を進め終了時に各モデルの進捗具合を確かめ合った。実行結果より、推定の期間やパラメータの値を変えると決定係数も大きく異なるためどの機関のデータを使って学習するのがよいのか、最適解を調べる必要があると考えている。また、いずれのモデルでも爆発的に増加した場合に対応することができなかったため、より詳細なパラメータを含めたり、人の行動を予測した要因をデータに組み込んで爆発的な増加にも対応できるようにしたいなと考えた。

参考文献

[1] 新型コロナウイルス　国内感染の状況　https://github.com/kaz-ogiwara/covid19,2020/08/13

[2] 都道府県別の詳細　https://www.nttdocomo.co.jp/utility/demographic_analytics/20200509.html,2020/08/13

[3] COVID-19 Challenge https://signate.jp/competitions/272,2020/08/13

[4] ももものきとデータ解析をはじめよう　python で時系列分析の練習 https://momonoki2017.blogspot.com/2018/02/python1.html

[5] SIGNATECOVID-19 最適化モデル https://github.com/ShuichiOhtsuka/SIGNATE_COVID_19/blob/master/SIGNATE_COVID-19_最適化モデル.ipynb,2020/08/13

[6] 感染病の数学予測モデル　(SIR モデル):事例紹介 https://qiita.com/kotai2003/items/d74583b588841e6427a2#3-パラメータ推定,2020/08/13

[7] SEIR モデル　https://ja.wikipedia.org/wiki/SEIR モデル, 2020/08/13

[8] SEIR モデルで新型コロナウイルスの挙動を予測して見た。 https://qiita.com/kotai2003/items/ed28fb723a335a873061,2020/08/13

[9] minimize 入門　モデルでデータ解析する https://qiita.com/MuAuan/items/88d5d0c416abb9e9915e