Київський національний університет
імені Тараса Шевченка
Факультет комп'ютерних наук та кібернетики
Кафедра обчислювальної математики

Кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 113 Прикладна математика
на тему:

Дослідження поведінки епідемій за допомогою машинного навчання

Виконала студентка 4-го курсу
Кодирова Заріна Іномівна _____

Науковий керівник:
асистент
Денисов Сергій Вікторович _____

Засвідчую, що в цій роботі немає запозичень з праць інших авторів без відповідних посилань.

Студентка _____

Роботу розглянуто й допущено до захисту на засіданні кафедри обчислювальної математики
«___» _____ 202_ р.,
протокол № ___
Завідувач кафедри
С. І. Ляшко _____

Київ — 2021

РЕФЕРАТ

Обсяг роботи 39 сторінок, 11 ілюстрацій, 12 джерел посилань, 1 додаток.

МАШИННЕ НАВЧАННЯ,ЕПІДЕМІЯ, COVID-19, ПРОГНОЗУВАННЯ, SIR-МОДЕЛЬ, МЕТОД ГРАДІЄНТНОГО БУСТІНГУ, НАЇВНИЙ БАЙЄСОВСЬКИЙ КЛАСИФІКАТОР, ЛОГІСТИЧНА РЕГРЕСІЯ, МЕТОД ОПОРНИХ ВЕКТОРІВ

Об’єктом роботи є процес порівняння методів машинного начання та SIR-моделі для прогнозування поведінки поточної пандемії Covid-19.

Метою роботи є визначення найкращого способу прогнозування кількості інфікованих на прикладі даних поточної пандемії.

Методи розроблення: комп’ютерне моделювання, алгоритми машинного навчання.

Інструменти розроблення: Jupyter Notebook, Google Colaboratory, мова програмування Python, середовище програмування PyCharm.

Результати роботи: ця робота представляє собою огляд розроблених підходів до прогнозування інфекційно захворюваності і розвитку епідемічного процесу на прикладі Covid-19. Зроблено висновки про недоліки та переваги існуючих методів прогнозування епідемій. Дане дослідження може бути використано у подальших епідеміологічних дослідженнях пов’язаних з Covid-19, пов’язаних з вакцинацією, зі створенням ліків та введенням карантинних обмежень.

ЗМІСТ

1 Вступ

Епідемії здавна погрожували людству, і тільки в XX столітті були розроблені ефективні засоби боротьби з інфекціями. До числа цих засобів належать і системи диференціальних рівнянь та сучасні методи машинного навчання. Математика допомагає моделювати поширення епідемій і допомагає зрозуміти, як з ними боротися.

У XXI столітті світ вже встиг зіткнутися з епідемією пташиного грипу в Південно-Східній Азії (в 2013 році) і спалахом захворювань лихоманкою Ебола в Африці (2015). Але в історії людства бували і більш масштабні епідемії [11].

У 551-580 роках нашої ери в Східної Римської імперії вибухнула перша задокументована пандемія чуми, що отримала назву Юстиніанової, в результаті якої загинуло близько 100 мільйонів чоловік. Ще через 800 років в Євразію і Північну Африку прийшла Чорна смерть - пандемія чуми, вбила від третини до половини тодішнього населення цих регіонів.

В результаті Першої світової війни, що викликала переміщення великої кількості людей, в 1918 році поширився іспанський грип, що охопив понад 500 мільйонів чоловік і вбив кожного десятого хворого. Ця пандемія стала наймасштабнішою за всю історію людської цивілізації, торкнувшись до 30% населення Землі.

Першу спробу використовувати математичний апарат для дослідження механізмів поширення захворювань зробив Данило Бернуллі, раніше відкрив перші закони гідродинаміки. Наступний крок зробив Вільям Фарр, що застосував в 1840 році нормальний розподіл до аналізу смертності від віспи.

У 2020 році тема епідемій знову знайшла популярність, оскільки світ охопила нова пандемія.

Актуальність роботи полягає у дослідженні переваг та недоліків сучасних методів машинного навчання при прогнозуванні епідемій.

Метою роботи є дослідження поведінки епідемій та визначення найкращого способу прогнозування поточної пандемії.

Об’єктом дослідження є різні алгоритми методів машинного навчання, модель-SIR та її варіації, а також поточна пандемія Covid-19 та дані, які можуть

бути пов’язані з нею.

Методи та засобами розробки: Jupyter Notebook, Google Colaboratory, средовище програмування PyCharm, мова програмування Python, бібілотеки numpy, sklearn, matplotlib та pandas для роботи з даними.

Можливі сфери застосування: дослідження механізмів розвитку і поширення епідемій є важливим способом боротьби із захворюваннями поряд з пошуком нових ліків, вакцинацією і профілактичними заходами, тому даже дослідження може бути використано у сферах медицини.

2 Дослідження предметної області

2.1 Епідемії

Визначення. Епідемією називається спалах хвороби, що виникає на широкій географічній території і вражає винятково високу частку населення [12].

Зазвичай універсальним епідеміологічним порогом вважається захворювання 5% жителів території. Однак багато медичних відомств розраховують власні епідемічні пороги для звичайних захворювань, виходячи з середньостатистичного рівня цього захворювання протягом багатьох років.

Визначення. Пандемія - епідемія, що виникає у всьому світі або на дуже широкій території, перетинаючи міжнародні кордони і зазвичай вражаючи велику кількість людей. Відповідно до критеріїв Всесвітньої організації охорони здоров'я, пандемія - поширення нового захворювання в світових масштабах. Наприклад, пандемія грипу відбувається, коли з'являється новий вірус грипу і поширюється по всьому світу і більшість людей не мають імунітет. Відомі пандемії та епідемії:

- чорна віспа
- чума
- холера
- тиф
- грип
- туберкульоз
- малярія
- проказа
- ВІЛ інфекція
- Коронавірусна інфекція COVID-19.

2.2 Загальна інформація про Covid-19

Визначення. Пандемія COVID-19 - поточна пандемія коронавірусної інфекції, викликана коронавірусами SARS-CoV-2. Спалах вперше було зафіксована в грудні 2019 року в Китаї, в місті Ухань.

Визначення. COVID-19 (абревіатура від англ. COronaVIrus Disease 2019 - коронавірусна інфекція 2019 року), - потенційно важка гостра респіраторна інфекція, що викликається коронавірусів SARS-CoV-2.

Являє собою небезпечне захворювання, яке може протікати як у формі гострої респіраторної вірусної інфекції легкого перебігу, так і у важкій формі. Вірус здатний вражати різні органи через пряме інфікування або за допомогою імунної відповіді організму. Найбільш частим ускладненням захворювання є вірусна пневмонія, здатна призводити до гострої дихальної недостатності, при яких найчастіше необхідні киснева терапія і респіраторна підтримка.

До найбільш поширених симптомів захворювання відносяться підвищена температура тіла, стомлюваність і сухий кашель. У рідкісних випадках ураження вірусом дітей і підлітків, ймовірно, може призводити до розвитку запального синдрому.

Поширюється вірус повітряно-крапельним шляхом через вдихання розпорошених в повітрі при кашлі, чханні, а також через потрапляння вірусу на поверхні з подальшим занесенням в очі, ніс або рот.

30 січня 2020 року Всесвітня організація охорони здоров’я оголосила цей спалах надзвичайною ситуацією у сфері охорони здоров’я, що має міжнародне значення, а 11 березня - пандемією. Станом на 6 травня 2021 року зареєстровано понад 155 млн випадків захворювання по всьому світу; більше 3,2 млн чоловік померло і більше 133 млн видужало.

Багато перших хворих мали відношення до ринку Ухань, на якому продаються морепродукти, а також птиця, змії, кажани і сільськогосподарські тварини.

Для правильної оцінки ризиків потрібні додаткові дослідження для виявлення поширеності вірусу серед населення в цілому, в тому числі серед людей без симптомів захворювання. Реальна поширеність COVID-19, спектр його поширення і реальний рівень смертності залишаються невідомими. Зареєстровані

показники смертності в різних країнах дуже неоднорідні: наприклад, у Німеччині повідомляється про дуже невелику кількість смертей у порівнянні з іншими європейськими країнами з аналогічним населенням і системами охорони здоров'я, що свідчить про відсутність єдиних критеріїв оцінки смертних випадків.

2.3 Дані про Covid-19

Дані було взято з вільного сховища даних для візуальної інформаційної панелі Novel Coronavirus 2019, що керується Центром системної науки та техніки університету Джонса Хопкінса (JHU CSSE).

Ресурсний центр коронавірусу Джонса Хопкінса (CRC) - це постійно оновлюване джерело даних COVID-19 та рекомендації експертів. Вони збирають та аналізують найкращі наявні дані про випадки, смерті, тести, госпіталізації та вакцини, щоб допомогти населенню, політикам та медичним працівникам у всьому світі реагувати на пандемію. TIME визнав CRC ”джерелом даних”для COVID-19 і назвав його 100 найкращих винаходів 2020 року. У 2021 році Research! America назвав CRC лауреатом премії ”Зустріч моменту для громадського здоров’я”.

Інформаційна панель вперше була опублікована 22 січня. Вона ілюструє місце та кількість підтверджених випадків COVID-19, смертей та відновлення для всіх постраждалих країн. Вона була розроблена з метою забезпечення дослідників, органів охорони здоров’я та широку громадськість зручним інструментом для відстеження спалаху в процесі його розвитку. Всі зібрані та відображені дані є у вільному доступі через сховище GitHub.

Також, було використано дані про населення кожної країни з платформи Kaggle.

3 Методи прогнозування епідемій

3.1 Класична SIR модель

Ця абревіатура походить від англійських слів Susceptible - Infected - Recovered, що означають «сприйнятливі - інфіковані - одужавші».

Модель SIR, розроблена Рональдом Россом, Вільямом Хамером та іншими на початку XX століття [1], складається із системи трьох звичайних диференціальних рівнянь.

Теоретичні роботи Кермака і Маккендрінка про моделі моделей інфекційних захворювань між 1927 і 1933 роками мали великий вплив на розвиток математичних моделей епідеміології [8]. Математичні моделі часто використовуються для з'ясування передачі захворювань. Ці моделі, як правило, засновані на компартментних моделях, можуть бути досить простими, але їх вивчення має вирішальне значення для отримання важливих знань про основні аспекти розповсюдження інфекційних хвороб.

Математичні моделі - це спрощене уявлення про те, як з часом інфекція поширюється серед населення. Більшість моделей епідемій засновані на розподілі населення в невелику кількість відділень. Кожен з них містить особи, однакові з точки зору свого статусу щодо хвороба, про яку йдеться. У моделі SIR виділяють три типи осіб:

- Сприйнятливі (S): це клас осіб, які є сприйнятливий до зараження; сюди може входити і пасивно імунні, як тільки вони втрачають імунітет, або, будь-яке новонароджене немовля, чия мати ніколи не була інфікована і, отже, не передала жодного імунітету;
- Інфіковані (I): у цьому класі рівень паразитів досить великий у хазяїна і існує потенціал для передачі інфекції іншим людям;
- Одужавші (R): включає всіх осіб, які були заражені та одужали.

Ця епідеміологічна модель фіксує динаміку гострого захворювання інфекції, які надають довічний імунітет після відновлення. Захворювання, при яких особи набувають постійний імунітет, а також для до якої може застосовуватись ця модель, включають кір, віспу, вітрянка, свинка, черевний тиф та дифтерія.

Як правило, загальний розмір населення вважається постійним, тобто $N =$

$S + I + R$. Надалі буде розглядатися модель, на яку впливає демографічний фактор.

Найпростіший і найпоширеніший спосіб введення демографії в модель SIR полягає в припущенні, що існує довжина життя - $\frac{1}{\mu}$ років. Потім, швидкість, з якою особини в будь-якому епідеміологічному відділі страждають від природної смертності визначається параметром μ. Важливо підкреслити, що цей фактор не залежить від захворювання і не призначений відображати патогенність збудника інфекції. Історично вважалося, що μ також представляє грубу народжуваність населення, забезпечуючи тим самим, що загальна чисельність населення не змінюється з часом, або іншими словами, $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$. Отже, модель SIR, яка включає демографічні показники, такі як народження та смерть, можна визначити як:

$$\begin{aligned} \frac{dS}{dt} &= \mu - \beta SI - \mu S \\ \frac{dI}{dt} &= \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} &= \gamma I - \mu R \end{aligned} \tag{3.1}$$

з початковими умовами $S(0) > 0$, $I(0) \geq 0$ та $R(0) \geq 0$.

Параметр β представляє швидкість передачі на інфекційний та негативний члени рівняння говорять нам про це кожна людина витрачає в середньому одиниці часу $\frac{1}{\gamma+\mu}$ у цьому класі.

Є три найпоширеніші порогові значення в епідеміології: $\mathscr{R}_0$, σ і R. Найпоширеніший і, мабуть,найважливішим є базове число відтворення [5] [4]. Основне число розмноження, що позначається $\mathscr{R}_0$, визначається як середня кількість вторинних інфекцій, що виникає, коли один інфекційний інфекційний організм потрапляє в повністю сприйнятливе населення.

Цей показник, $\mathscr{R}_0$, є відомим результатом завдяки Кермаку та Маккендрік-ку і називається ”пороговим явищем що дає межу між стікими до хвороби або смертю від хвороби. $\mathscr{R}_0$ ще називають базовим коефіцієнтом відтворення.

Номер контакту, σ - це середня кількість адекватних контактів типового інфекціоніста протягом інфекційного періоду. Адекватний контакт - це той, який є достатнім для передачі, якщо особа, з якою контактує сприйнятливий, є інфекційним. Неявно припускається, що заражений аутсайдер знаходиться в

популяції господаря протягом усього інфекційного періоду і змішується з популяцією господаря точно так само, як і місцеве населення.

Число R, є середньою кількістю вторинних інфекцій, спричинених типовим інфекціоністом протягом усього періоду зараження. Ці три величини $\mathscr{R}_0$, σ і R однакові на початку розповсюдження інфекційного захворювання, коли сприйнятлива вся популяція (крім загарбника). $\mathscr{R}_0$ визначається лише під час зараження, тоді як σ та R визначаються постійно. Номер заміщення R - це фактична кількість вторинних випадків від типового заражника, так що після того, як інфекція вторглася в популяцію, і всі вже не сприйнятливі, R завжди менше базового числа розмноження $\mathscr{R}_0$.

Отже, якщо припустити, що сприйнятлива до вірусу вся популяція, то середня кількість нових інфекційних інфекцій на одну інфекційну особину визначається як

$$\mathscr{R}_0 = \frac{\beta}{\gamma + \mu}.$$

Включення демографічної динаміки може дозволити хворобі вимерти або зберегтись серед населення в довгостроковій перспективі. З цієї причини важливо дослідити, що відбувається, коли працює система знаходиться в рівновазі.

Модель, визначена SIR, має точку рівноваги, якщо $E^* = (S^*, I^*, R^*)$ задовольняє систему:

$$\begin{cases} \dfrac{dS}{dt} = 0 \\ \dfrac{dI}{dt} = 0 \\ \dfrac{dR}{dt} = 0 \end{cases}$$

Якщо в точці рівноваги інфекційний компонент дорівнює нулю $(I^* = 0)$, це означає, що збудник зазнав вимирання, а E^* називається рівновагою, вільною від хвороб (DFE). Якщо $I^* > 0$, хвороба зберігається в популяції і E^* називається ендемічною рівновагою (EE).

За допомогою деяких обчислень та алгебраїчних маніпуляцій можна отримати дві рівноваги для системи (3.1):

DFE: $E_1^* = (1, 0, 0)$

EE: $E_2^* = (\frac{1}{\mathscr{R}_0}, \frac{\mu}{\beta}(\mathscr{R}_0 - 1), 1 - \frac{1}{\mathscr{R}_0} - (\frac{\mu}{\beta}(\mathscr{R}_0 - 1))$

Коли $\mathscr{R}_0 < 1$, кожна заражена особина виробляє в середньому менше однієї нової зараженої особини, а отже, передбачувано, що зараження буде видалено від популяції. Якщо $\mathscr{R}_0 > 1$, збудник може вторгнутися в сприйнятливу популяцію [5]. Можна довести, що для того, щоб ендемічна рівновага була стабільною, $\mathscr{R}_0$ має бути більше одиниці, інакше рівновага, вільна від хвороб, є стабільною. Ця порогова поведінка є дуже корисною, коли ми зможемо визначити, які заходи контролю та з якою величиною будуть найбільш ефективними у зменшенні $\mathscr{R}_0$ нижче одного, забезпечуючи важливі вказівки для ініціатив у галузі охорони здоров'я.

SIR-модель перестає працювати в разі необхідності враховувати неоднорідність популяції (наприклад, різну щільність населення в різних районах), різні шляхи передачі інфекції та фактори випадковості, значимі в малих популяціях і на початковій фазі поширення захворювання.

Також існують наступні варіації SIR-моделі:

- SIRS — «Сприйнятливі - інфіковані - видужали - сприйнятливі»: модель опису динаміки захворювань с тимчасовим імунітетом (люди, які перехворіли з часом знову можуть заразитися);
- SEIR — «Сприйнятливі - контактні (Exposed) - інфіковані - видужали»: модель для опису поширення захворювань з інкубаційним періодом;
- SIS - «сприйнятливі - інфіковані - сприйнятливі»: модель для поширення захворювання, до якого не виробляється імунітет;
- MSEIR - «наділені імунітетом від народження (Maternally derived immunity) - сприйнятливі - контактні - інфіковані - видужали»: модель, що враховує імунітет дітей, набутий внутрішньоутробно.

3.2 Машинне навчання

Визначення. Машинне навчання вважається гілкою штучного інтелекту, основна ідея якого полягає в тому, щоб комп'ютер не просто використовував заздалегідь написаний алгоритм, а сам навчився розв'язувати поставлену задачу [7].

У 1959 році Артур Самуель, дослідник штучного інтелекту, ввів термін «машинне навчання». Він винайшов першу комп'ютерну програму з гри в шашки, яка навчалась сама. Самуель визначив машинне навчання як процес, в результаті якого комп'ютери здатні показати таку поведінку, яу в них не було запрограмовано спочатку [10]. Джозеф Вейцбаум разом з Дональдом Кнутом прцював над проектом TeX, результатом якого стала система комп'ютерної верстки, ось уже майже 40 років не має собі рівних для підготовки математичних текстів.

Всі моделі машинного навчання поділяються на навчання з учителем (supervised) і без вчителя (unsupervised).

- Навчання з вчителем - пошук залежності між кінцевим результатом і початковим описом завдання;
- Навчання без вчителя - в цьому випадку кінцевий результат не відомий заздалегідь і потрібно шукати залежності між об'єктами, тобто стоїть завдання організувати інформацію або описати їх структуру.

Класичні задачі машинного навчання:

- Класифікація, як правило, виконується за допомогою навчання з учителем на етапі власне навчання.
- Кластеризація, як правило, виконується за допомогою навчання без учителя
- Регресія, як правило, виконується за допомогою навчання з учителем на етапі тестування, є окремим випадком задач прогнозування.
- Зниження розмірності даних і їх візуалізація виконується за допомогою навчання без учителя
- Побудова рангових залежностей виявлення аномалій

Далі розглянемо алгоритми, які будуть використовуватись при прогнозуванні епідемії. Всі алгоритми, які будуть розглядатись відносяться до навчання з учителем.

Метод градієнтного бустінгу

Визначення. Метод градієнтного бустінгу - це техніка машинного навчання для задач регресії та класифікації, яка створює модель прогнозування у вигляді ансамблю слабких моделей прогнозування, як правило, дерев рішень.

Ідея градієнтного бустінгу виникла в спостереженні Лео Бреймана, що бустінг можна інтерпретувати як алгоритм оптимізації відповідної до функції витрат. Явні алгоритми посилення градієнта регресії згодом були розроблені Джеромом Х. Фрідманом [3] одночасно з більш загальною перспективою посилення градієнта Леу Мейсона, Джонатана Бакстера, Пітера Бартлетта та Маркуса Фріна.

У багатьох навчальних задачах з учителем є вихідна змінна y та вектор вхідних змінних x, пов'язаних між собою з деяким імовірнісним розподілом. Мета - знайти якусь функцію $\hat{F}(x)$, яка найкраще апроксимує вихідну змінну зі значень вхідних змінних. Це формалізується введенням деякої функції втрат $L(y, F(x))$ та мінімізацією її:

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))]. \tag{3.2}$$

Метод грдієнтного бустінгу приймає дійсне значення y і шукає наближення $\hat{F}(x)$ у вигляді зваженої суми функцій $h_i(x)$ з якогось класу $\mathcal{H}$, що називається базовими (або слабкими) учнями:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const}. \tag{3.3}$$

Зазвичай дано навчальний набір $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ відомих зразкових значень x та відповідних значень y . Відповідно до емпіричного принципу мінімізації ризику, метод намагається знайти апроксимацію $\hat{F}(x)$, яка мінімізує середнє значення функції втрат на навчальному наборі, тобто мінімізує емпіричний ризик. Це робиться, починаючи з моделі, що складається з постійної функції $F_0(x)$, і поступово розширює її:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma), \tag{3.4}$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h_m \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right], \tag{3.5}$$

де $h_m \in \mathcal{H}$ є базовою навчальною функцією.

На жаль, вибір найкращої функції h на кожному кроці для довільної функції втрат L є загалом обчислювально нездійсненною задачею оптимізації. Тому ми обмежуємо наш підхід спрощеною версією проблеми.

Ідея полягає в тому, щоб застосувати найкрутіший крок спуску до цієї проблеми мінімізації (функціональний градієнтний спуск). Якби ми розглянули неперервний випадок, тобто де $\mathcal{H}$ - це набір довільних диференційованих функцій на $\mathbb{R}$, ми б оновили модель відповідно до наступних рівнянь

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i)), \tag{3.6}$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))\right), \tag{3.7}$$

де похідні беруться функцій F_i для $i \in \{1, .., m\}$, а γ_m - довжина кроку. Однак у дискретному випадку, тобто коли множина $\mathcal{H}$ скінченна, ми вибираємо функцію-кандидат h, найближчу до градієнта L, для якої тоді коефіцієнт γ можна обчислити за допомогою пошуку рядків на вищезазначеному рівняння. Зауважимо, що такий підхід є евристичним , отже, не дає точного розв'язання даної задачі, а швидше наближення.

Метод градієнтного бустінгу над дерева рішень

Визначення. Дерево рішень - це інструмент підтримки прийняття рішень, який використовує деревоподібну модель рішень та їх можливі наслідки, включаючи результати випадкових подій, витрати на ресурси та корисність. Це один із способів відобразити алгоритм, який містить лише умовні оператори управління.

Визначення. Градієнтний бустінг зазвичай використовується з деревами рішень фіксованого розміру в якості базових учнів. Для цього особливого випадку Фрідман пропонує модифікацію методу, яка покращує якість підгонки кожного базового учня.

Загальне посилення градієнта на m-му кроці відповідало б дереву рішень $h_m(x)$ до псевдо-залишків. Нехай J_m - кількість його листків. Дерево розділяє вхідний простір на J_m непересічні області $R_{1m}, \ldots, R_{J_m m}$ і передбачає постійне

значення в кожному регіоні. Використовуючи позначення індикатора, результат $h_m(x)$ для вводу x можна записати як суму:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} \mathbf{1}_{R_{jm}}(x) \tag{3.8}$$

де b_{jm} - значення, передбачене в регіоні R_{jm}.[10]

Потім коефіцієнти b_{jm} множаться на деяке значення γ_m, вибране за допомогою пошуку рядків, щоб мінімізувати функцію втрат, і модель оновлюється наступним чином:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad \gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{3.9}$$

Фрідман пропонує змінити цей алгоритм таким чином, щоб він вибирав окреме оптимальне значення γ_{jm} для кожного з регіонів дерева, замість одного γ_m для всього дерева. Він називає модифікований алгоритм "TreeBoost". Потім коефіцієнти b_{jm} з процедури підгонки дерева можна просто відкинути, і правило оновлення моделі стає:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x), \quad \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma). \tag{3.10}$$

Розмір дерев

J - кількість кінцевих вузлів у деревах, є параметром методу, який можна налаштувати для відомого набору даних. Він контролює максимально допустимий рівень взаємодії між змінними в моделі. При $J = 2$ не допускається взаємодія між змінними. При $J = 3$ модель може включати ефекти взаємодії до двох змінних тощо.

Хасті та ін. [2] говорив про те, що зазвичай $4 \leq J \leq 8$ добре працюють для підвищення, а результати досить нечутливі до вибору J у цьому діапазоні, $J = 2$ недостатньо для багатьох додатків, а $J > 10$ навряд чи є можливим варіантом.

Наївний байєсовський класифікатор

Визначення. Наївний байєсовський класифікатор - простий імовірнісний класифікатор, заснований на застосуванні теореми Байэса зі строгими (наївними) припущеннями про незалежність [9].

Незважаючи на дуже спрощені умови, наївні байєсовські класифікатори іноді можуть бути більш вдалим вибором, ніж нейронні мережі. Перевагою наївного байєсівського класифікатора є мала кількість даних, необхідних для навчання, оцінки параметрів та класифікації.

Модель наївного байєсівського класифікатора

Імовірнісна модель для класифікатора - це умовна модель $p(C_1, \ldots, F_n)$ над залежною змінною класу C з невеликою кількістю результатів або класів, залежна від кількох змінних $F_1, \ldots, F_n$. Проблема полягає в тому, що коли кількість властивостей n дуже велике або коли властивість може приймати велику кількість значень, тоді будувати таку модель на імовірнісних таблицях стає неможливо. Тому використовуючи теорему Байеса, запишемо модель наступним чином:

$$p(C \mid F_1, \ldots, F_n) = \frac{p(C)\, p(F_1, \ldots, F_n \mid C)}{p(F_1, \ldots, F_n)}. \tag{3.11}$$

Нас цікавить лише чисельник цього дробу, так як знаменник не залежить від C і значення властивостей F_i дані, так що знаменник - константа.

Чисельник еквівалентний спільній ймовірності моделі $p(C, F_1, \ldots, F_n)$ яка може бути переписана наступним чином:

$$\begin{aligned} p(C, F_1, \ldots, F_n) &= p(C)\, p(F_1, \ldots, F_n \mid C) = \\ &= p(C)\, p(F_1 \mid C)\, p(F_2, \ldots, F_n \mid C, F_1) = \\ &= p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1)\, p(F_3, \ldots, F_n \mid C, F_1, F_2) = \\ &= p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1) \cdot \ldots \cdot p(F_n \mid C, F_1, F_2, F_3, \ldots, F_{n-1}) \end{aligned} \tag{3.12}$$

і т. д. Тепер можна використовувати «наївні» припущення умовної незалежності: припустимо, що кожна властивість F_i умовно незалежно від будь-якого іншого властивості F_j при $j \neq i$. Це означає:

$$p(F_i \mid C, F_j) = p(F_i \mid C) \tag{3.13}$$

таким чином, спільна модель може бути записана як:

$$\begin{aligned} p(C, F_1, \ldots, F_n) &= p(C)\, p(F_1 \mid C)\, p(F_2 \mid C)\, p(F_3 \mid C) \cdot \ldots \cdot p(F_n \mid C) = \\ &= p(C) \prod_{i=1}^{n} p(F_i \mid C). \end{aligned} \tag{3.14}$$

Це означає, що з припущення про незалежність, умовний розподіл по класовій змінній C може бути виражено так:

$$p(C \mid F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i \mid C) \tag{3.15}$$

де $Z = p(F_1, \ldots, F_n)$ - це масштабний множник, що залежить тільки від $F_1, \ldots, F_n$, тобто константа, якщо значення змінних відомі.

Побудова класифікатора по ймовірнісної моделі

Наївний байесовский класифікатор об'єднує модель з правилом рішення. Одне загальне правило має вибрати найбільш ймовірну гіпотезу; воно відоме як апостеріорне правило прийняття рішення (MAP). Відповідний класифікатор - це функція classify, визначена наступним способом:

$$\text{classify}(f_1, \ldots, f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i \mid C = c) \tag{3.16}$$

Логістична регресія

Визначення. Логістична регресія (Logistic regression) - метод побудови лінійного класифікатора, що дозволяє оцінювати апостеріорні ймовірності приналежності об'єктів класів [6].

Логістична регресія застосовується для прогнозування ймовірності виникнення деякої події за значеннями безлічі ознак. Для цього вводиться залежна змінна y, яка приймає лише одне з двох значень - як правило, це числа 0 (подія не відбулося) і 1 (подія відбулася), множина незалежних змінних (також називають ознаками, предикторами або регресорами) - дійсних $x_1, x_2, ..., x_n$, на основі значень яких потрібно обчислити ймовірність прийняття того чи іншого значення залежної змінної. Для простоти запису вводиться фіктивна ознака $x_0 = 1$.

Робиться припущення про те, що ймовірність настання події $y = 1$ дорівнює:

$$\mathbb{P}\{y = 1 \mid x\} = f(z), \tag{3.17}$$

де $z = \theta^T x = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$, x та θ — вектори-стовпці значень незалежних змінних $1, x_1, \ldots, x_n$ і параметрів (коефіцієнтів регресії) - дійсних чисел $\theta_0, ..., \theta_n$, відповідно, а $f(z)$ — логістична функція:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{3.18}$$

Так як y приймає лише значення 0 і 1, то ймовірність прийняти значення 0 дорівнює:

$$\mathbb{P}\{y = 0 \mid x\} = 1 - f(z) = 1 - f(\theta^T x). \tag{3.19}$$

Функцію розподілу y при заданому x можна записати в такому вигляді:

$$\mathbb{P}\{y \mid x\} = f(\theta^T x)^y (1 - f(\theta^T x))^{1-y}, \quad y \in \{0, 1\}. \tag{3.20}$$

Фактично, це є розподіл Бернуллі з параметром, рівним $f(\theta^T x)$.

Підбір параметрів

Для підбору параметрів $\theta_0, ..., \theta_n$ необхідно скласти навчальну вибірку, що складається з наборів значень незалежних змінних і відповідних їм значень залежної змінної y. Формально, це множина пар $(x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})$, де $x^{(i)} \in \mathbb{R}^n$ — вектор значень незалежних змінних, а $y^{(i)} \in \{0, 1\}$ — відповідне їм значення y. Кожна така пара називається навчальним прикладом.

Зазвичай використовується метод максимальної правдоподібності, згідно з яким вибираються параметри θ, максимізує значення функції правдоподібності на навчальній вибірці:

$$\hat{\theta} = \text{argmax}_\theta \, L(\theta) = \text{argmax}_\theta \prod_{i=1}^{m} \mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\}. \tag{3.21}$$

Максимізація функції правдоподібності еквівалентна максимізації її логарифма:

$$\begin{aligned} \ln L(\theta) &= \sum_{i=1}^{m} \log \mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\} = \\ &= \sum_{i=1}^{m} y^{(i)} \ln f(\theta^T x^{(i)}) + (1 - y^{(i)}) \ln(1 - f(\theta^T x^{(i)})), \end{aligned} \tag{3.22}$$

де

$$\theta^T x^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)}. \tag{3.23}$$

Для максимізації цієї функції може бути застосований, наприклад, метод градієнтного спуску. Він полягає у виконанні наступних ітерацій, починаючи з деякого початкового значення параметрів θ:

$$\theta := \theta + \alpha \nabla \ln L(\theta) = \theta + \alpha \sum_{i=1}^{m} (y^{(i)} - f(\theta^T x^{(i)})) x^{(i)}, \alpha > 0. \tag{3.24}$$

На практиці також застосовують стохастичний градієнтний спуск або його варіанти

Регуляризація

Для зменшення ефекту перенавчання, на практиці часто розглядається логістична регресія з регуляризацією.

Регуляризація полягає в тому, що вектор параметрів θ розглядається як випадковий вектор з деякою заданою апріорною щільністю розподілу $p(\theta)$. Для навчання моделі замість методу найбільшої правдоподібності при цьому використовується метод максимізації апостеріорної оцінки, тобто шукаються параметри θ, що максимізує величину:

$$\prod_{i=1}^{m} \mathbb{P}\{y^{(i)} \mid x^{(i)}, \theta\} \cdot p(\theta). \tag{3.25}$$

В якості апріорного розподілу часто виступає багатовимірний нормальний розподіл $\mathcal{N}(0, \sigma^2 I)$ з нульовим середнім і матрицею коваріації $\sigma^2 I$, відповідне апріорному переконанню про те , що всі коефіцієнти регресії повинні бути невеликими числами, ідеально - більшість маловагомих коефіцієнтів повинніа бути нулями. Підставивши щільність цього апріорного розподілу в формулу вище, і прологаривмувавши, отримаємо наступну оптимізаційну задачу:

$$\sum_{i=1}^{m} \log \mathbb{P}\{y^{(i)} \mid x^{(i)}, \theta\} - \lambda \|\theta\|^2 \to \max, \tag{3.26}$$

де $\lambda = \text{const}/\sigma^2$ - параметр регуляризації. Цей метод відомий як $L2$-регуляризована логістична регресія, так як в цільову функцію входить $L2$-норма вектора параметрів для регуляризації.

Якщо замість $L2$-норми використовувати $L1$ норма, що еквівалентно використанню розподілу Лапласа, як апріорного, замість нормального, то вийде інший поширений варіант методу - $L1$ -регуляризована логістична регресія:

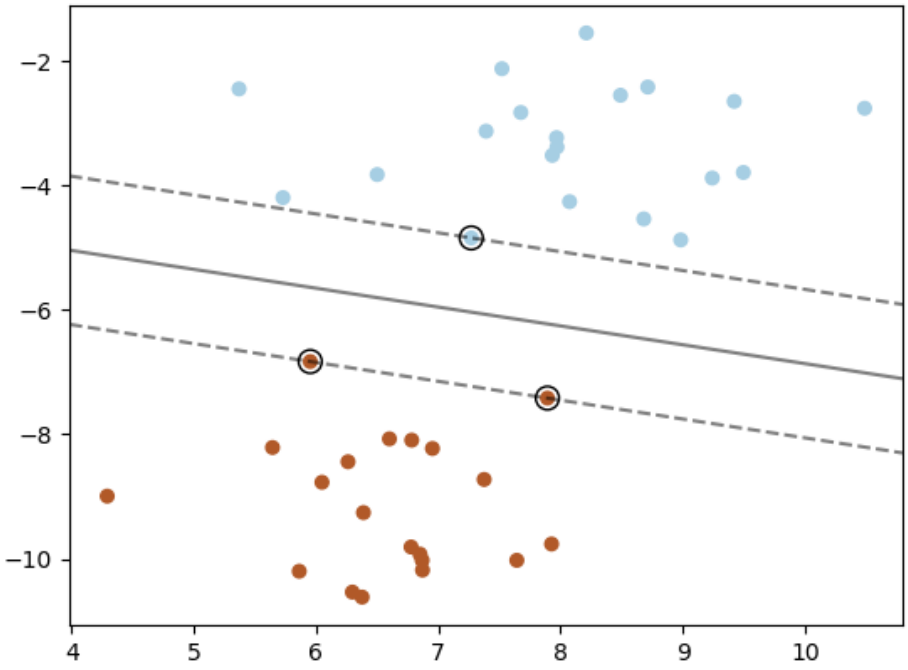
$$\sum_{i=1}^{m} \log \mathbb{P}\{y^{(i)} \mid x^{(i)}, \theta\} - \lambda \|\theta\|_1 \to \max. \tag{3.27}$$

Метод опорних векторів

Визначення. Метод опорних векторів (SVM) - це набір методів навчання з учителем, що використовуються для задач класифікації та регресії [2]

Визначення. Гіперплощиною називається підпростір з розмірністю, на одиницю меншою, ніж осяжний простір.

Метод опорних векторів конструює гіперплощину або набір гіперплощин у високому або нескінченному розмірному просторі. Інтуїтивно, хорошого розділення досягає гіперплощина, яка має найбільшу відстань до найближчих точок будь-якого класу (так званий функціональний запас), оскільки загалом, чим більший запас, тим нижча похибка узагальнення класифікатора. На малюнку нижче показано функцію прийняття рішення для лінійно відокремлюваної задачі з трьома зразками на границях полів, які називаються “опорними векторами”:



Перші ідеї методу були запропоновані ще в 1950-ті роки. Метод був створений на основі статистичної теорії навчання. Метод став відомий і популярний після статті Володимира Вапника в 1992 році. В даний час метод успішно використовується у багатьох областях. Метод також був модифікований для задач регресії.

Опис алгоритму

Нехай нам дано точки, що мають наступний вигляд:

$$\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}, \tag{3.28}$$

де c_i мають значення 1 або -1, в залежності від класу точки $\mathbf{x}_i$. $\mathbf{x}_i$ — це p-мірний вектор, нормалізований значеннями $[0, 1]$ або $[-1, 1]$. Якщо точки не будуть нормалізовані, то точка з великими відхиленнями від середніх значень координат точок занадто сильно вплине на класифікатор. Розглянемо це як навчальну вибірку, в якій для кожного елемента вже задано клас, до якого він належить.Алгоритм методу опорних векторів повинен класифікувати їх таким же чином. Для цього ми будуємо розділяє гіперплощину, яка має вигляд:

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \tag{3.29}$$

Вектор **w** - перпендикуляр до розділяючої гіперплощини. Параметр $\frac{b}{\|\mathbf{w}\|}$ дорівнює по модулю відстані від гіперплощини до початку координат. Якщо параметр b дорівнює нулю, гіперплощина проходить через початок координат.

Оскільки нас цікавить оптимальний розподіл, то нас цікавлять опорні вектори і гіперплощини, паралельні оптимальній і найближчі до опорних векторах двох класів. Ці паралельні гіперплощини можуть бути описані наступними рівняннями.

$$\mathbf{w} \cdot \mathbf{x} - b = 1, \tag{3.30}$$

$$\mathbf{w} \cdot \mathbf{x} - b = -1 \tag{3.31}$$

Якщо навчальна вибірка лінійно розподільна, то гіперплощини потрібно обрати таким чином, щоб між ними не лежала жодна точка навчальної вибірки і потім максимізувати відстань між гіперплощинами. Ширина смуги дорівнює

$\frac{2}{\|\mathbf{w}\|}$, таким чином наше завдання мінімізувати $\|\mathbf{w}\|$. Щоб виключити всі точки зі смуги, ми повинні переконатися, що для всіх i виконується наступне:

$$\mathbf{w} \cdot \mathbf{x_i} - b \geq 1, \; c_i = 1 \tag{3.32}$$

$$\mathbf{w} \cdot \mathbf{x_i} - b \leq -1, \; c_i = -1 \tag{3.33}$$

Або можна записати:

$$c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n. \tag{3.34}$$

Побудова оптимальної розділяючої гіперплощини зводиться до мінімізаці **w**, за умови 3.34. Задача має наступний вигляд:

$$\begin{cases} \|\mathbf{w}\|^2 \to \min \\ c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n. \end{cases} \tag{3.35}$$

За теоремою Куна - Таккера (ссылка на теорему) ця задача еквівалентна двоїстій задачі пошуку сідлової точки функції Лагранжа

$$\begin{cases} \mathbf{L}(\mathbf{w}, \mathbf{b}; \lambda) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \lambda_\mathbf{i}(c_i((\mathbf{w} \cdot \mathbf{x_i}) - b) - 1) \to \min_{w,b} \max_{\lambda} \\ \lambda_\mathbf{i} \geq 0, \quad 1 \leq i \leq n \end{cases} \tag{3.36}$$

де $\lambda = (\lambda_\mathbf{1}, \ldots, \lambda_\mathbf{n})$ — вектор двоїстих змінних.

Зведемо цю задачу до еквівалентної задачі квадратичного програмування, що містить тільки двоїсті змінні:

$$\begin{cases} -\mathbf{L}(\lambda) = -\sum_{i=1}^{n} \lambda_\mathbf{i} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_\mathbf{i}\lambda_\mathbf{j}c_ic_j(\mathbf{x_i} \cdot \mathbf{x_j}) \to \min_{\lambda} \\ \lambda_\mathbf{i} \geq 0, \quad 1 \leq i \leq n \\ \sum_{i=1}^{n} \lambda_\mathbf{i}c_i = 0 \end{cases} \tag{3.37}$$

Нехай задача розв'язана, тоді **w** та **b** шукаємо наступним чином:

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_\mathbf{i}c_i\mathbf{x_i} \tag{3.38}$$

$$\mathbf{b} = \mathbf{w} \cdot \mathbf{x_i} - c_i, \quad \lambda_i > 0 \tag{3.39}$$

Алгоритм класифікації має наступний вигляд:

$$a(x) = sign\left(\sum_{i=1}^{n} \lambda_\mathbf{i}c_i\mathbf{x_i} \cdot \mathbf{x} - b\right) \tag{3.40}$$

Сумування здійснюється по опорним векторам, для яких $\lambda_{\mathbf{i}} \neq 0$ Для того, щоб алгоритм міг працювати в разі, якщо класи лінійно нерозподільні, дозволимо йому допускати помилки на навчальній вибірці. Введемо набір додаткових змінних $\xi_i \geq 0$, що характеризують величину похибки $\mathbf{x}_i, \quad 1 \leq i \leq n$. Візьмемо за відправну точку 3.36, пом'якшимо обмеження нерівності, так само введемо в функціонал, що мінімізується штраф за сумарну помилку:

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{w,b,\xi_i} \\ c_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \\ \xi_i \geq 0, \quad 1 \leq i \leq n \end{cases} \tag{3.41}$$

Коефіцієнт C — параметр методу, який дозволяє регулювати відношення між максимізацією ширини розділяючої смуги і мінімізацією сумарної похибки.

Аналогічно, за з теоремою Куна-Таккера зводимо задачу до пошуку сідлової точки функції Лагранжа:

$$\begin{cases} \mathbf{L}(\mathbf{w},\mathbf{b},\xi;\lambda,\eta) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\lambda_{\mathbf{i}}(c_i((\mathbf{w}\cdot\mathbf{x_i}) - b) - 1) - \\ -\sum_{i=1}^{n}\xi_{\mathbf{i}}(\lambda_{\mathbf{i}} + \eta_{\mathbf{i}} - C) \to \min_{w,b,\xi}\max_{\lambda,\eta} \\ \xi_{\mathbf{i}} \geq 0, \lambda_{\mathbf{i}} \geq 0, \eta_{\mathbf{i}} \geq 0, \quad 1 \leq i \leq n \\ \left[\begin{array}{l} \lambda_{\mathbf{i}} = 0 \\ c_i(\mathbf{w}\cdot\mathbf{x_i} - b) = 1 - \xi_i, \end{array}\right. \quad 1 \leq i \leq n \\ \left[\begin{array}{l} \eta_{\mathbf{i}} = 0 \\ \xi_{\mathbf{i}} = 0, \end{array}\right. \quad 1 \leq i \leq n \end{cases} \tag{3.42}$$

По аналогії зведемо задачу до наступної:

$$\begin{cases} -\mathbf{L}(\lambda) = -\sum_{i=1}^{n}\lambda_{\mathbf{i}} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_{\mathbf{i}}\lambda_{\mathbf{j}}c_ic_j(\mathbf{x_i}\cdot\mathbf{x_j}) \to \min_{\lambda} \\ 0 \leq \lambda_{\mathbf{i}} \leq \mathbf{C}, \quad 1 \leq i \leq n \\ \sum_{i=1}^{n}\lambda_{\mathbf{i}}c_i = 0 \end{cases} \tag{3.43}$$

На практиці для побудови методу опорних векторів вирішують саме завдання, де класи лінійно нерозподільні, а не 3.37, так як гарантувати лінійну роздільність точок на два класи в загальному випадку не представляється можливим. Цей варіант алгоритму називають алгоритмом з м'яким зазором (soft-margin SVM), тоді як в лінійно розподільному випадку говорять про жорсткий зазор (hard-margin SVM).

Для алгоритму класифікації зберігається формула 3.40, з тією лише різницею, що тепер ненульовими λ_i володіють не тільки опорні об'єкти, але і об'єкти-порушники. У певному сенсі це недолік, оскільки порушниками часто виявляються шумові викиди, і побудоване на них вирішальне правило, по суті справи, спирається на шум.

Константу C зазвичай вибирають за критерієм змінного контролю (крос валідація). Це трудомісткий спосіб, так як завдання доводиться вирішувати спочатку при кожному значенні C.

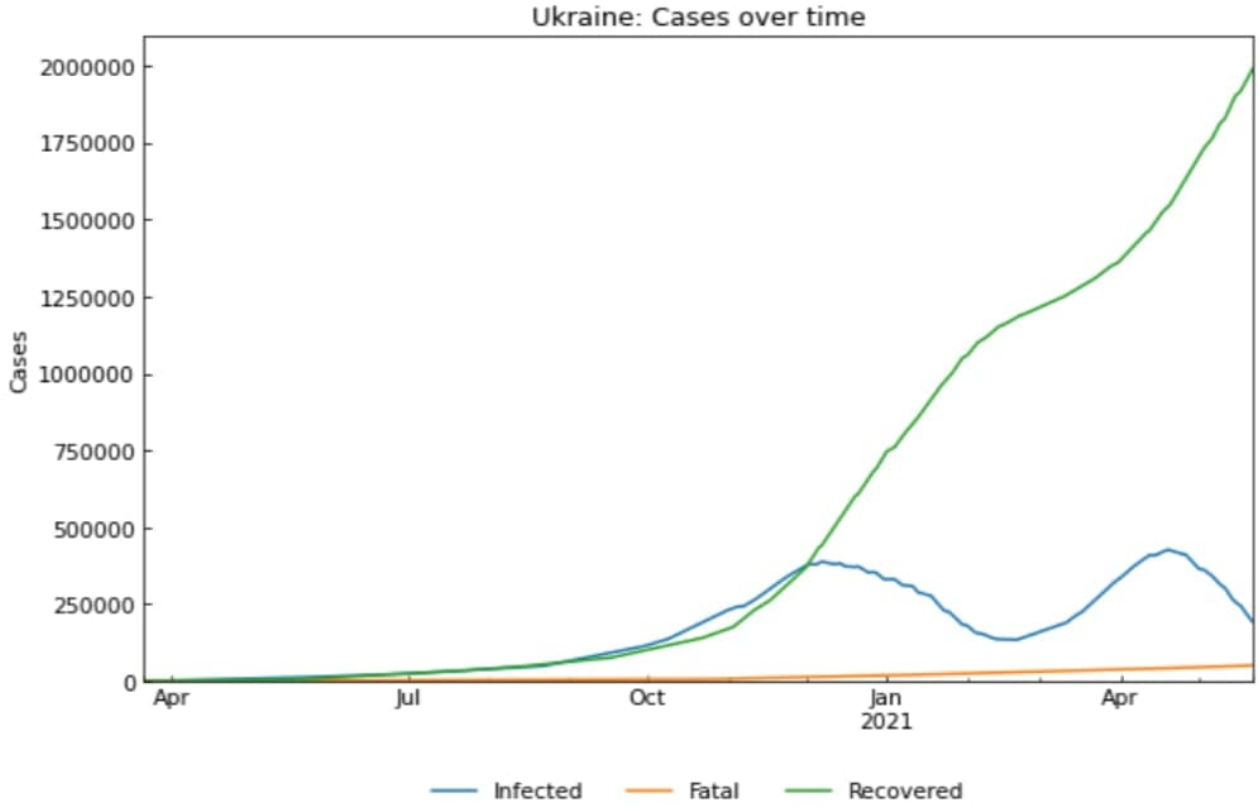
Якщо є підстави вважати, що вибірка майже лінійно роздільна, і лише об'єкти-порушники класифікуються невірно, то можна застосувати фільтрацію викидів. Спочатку завдання вирішується при деякому C, і з вибірки видаляється невелика частка об'єктів, що мають найбільшу величину помилки ξ_i. Після цього завдання вирішується заново при зменшеній вибірці. Можливо, доведеться виконати кілька таких ітерацій, поки що залишилися об'єкти не виявляться лінійно нерозподільні.

4 Результати

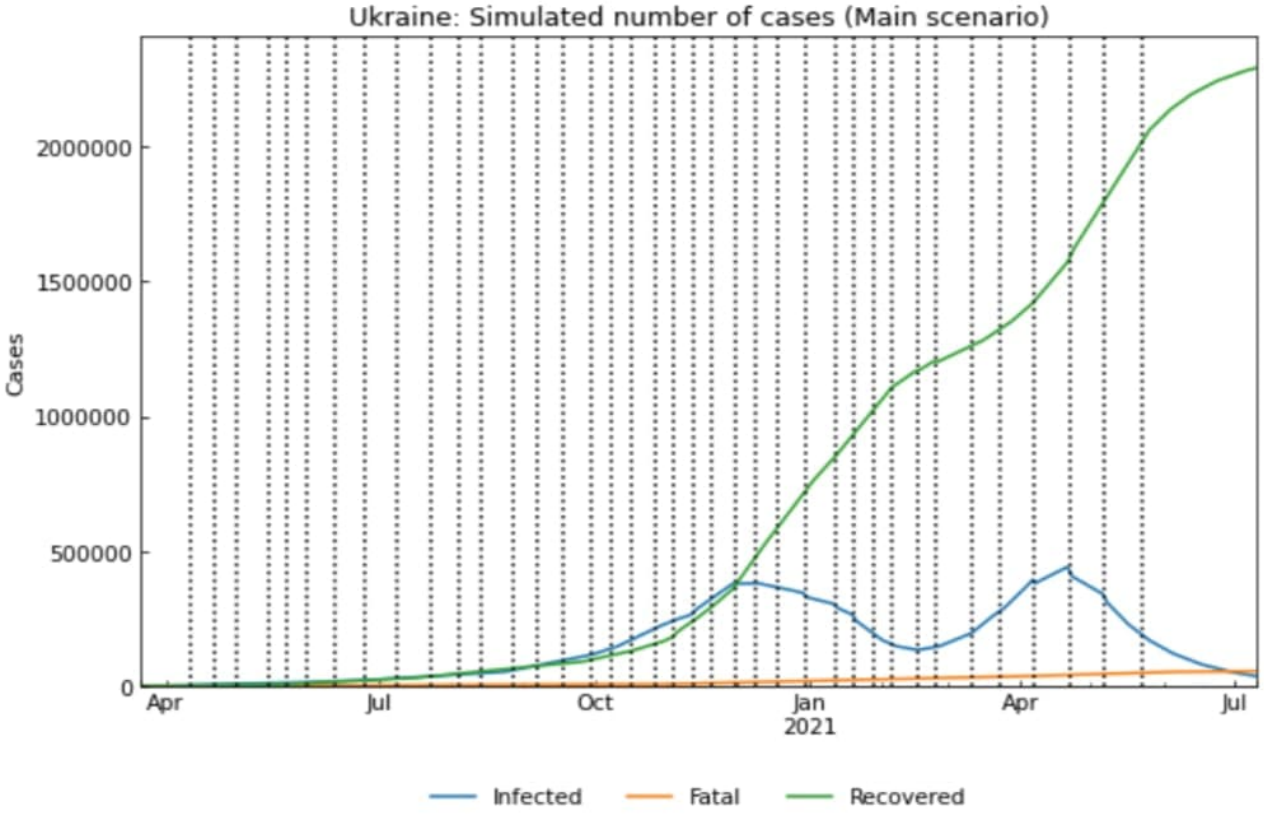
Для всіх методів прогнозування епідеміїї Covid-19 використовувались однакові дані, приклад яких можна знайти у Додатку. На результатах зображено кількість інфікованиї, яка вже є відомою та прогнозовані результати.

4.1 SIR Model

На даному графіку зображено кількість інфікованих, одужавших та летальні випадки, які вже є відомими. Даний графік охоплює територію України.

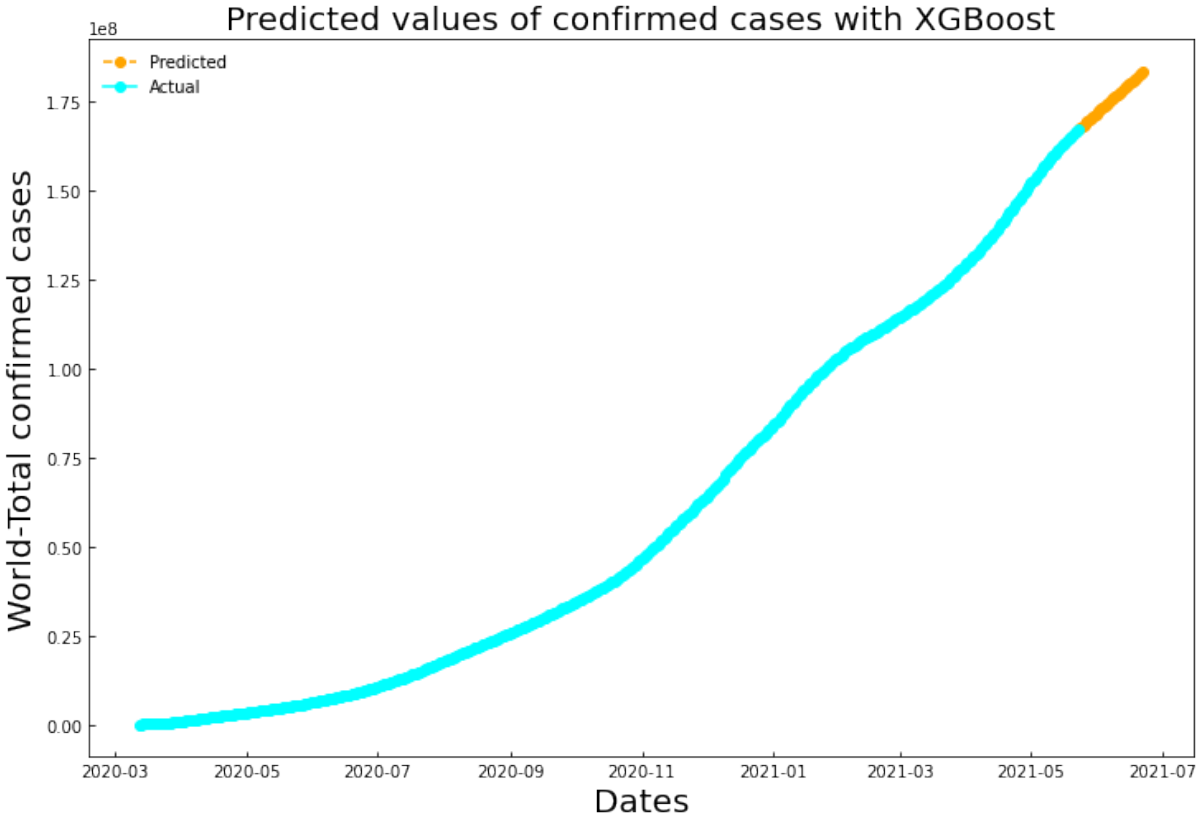


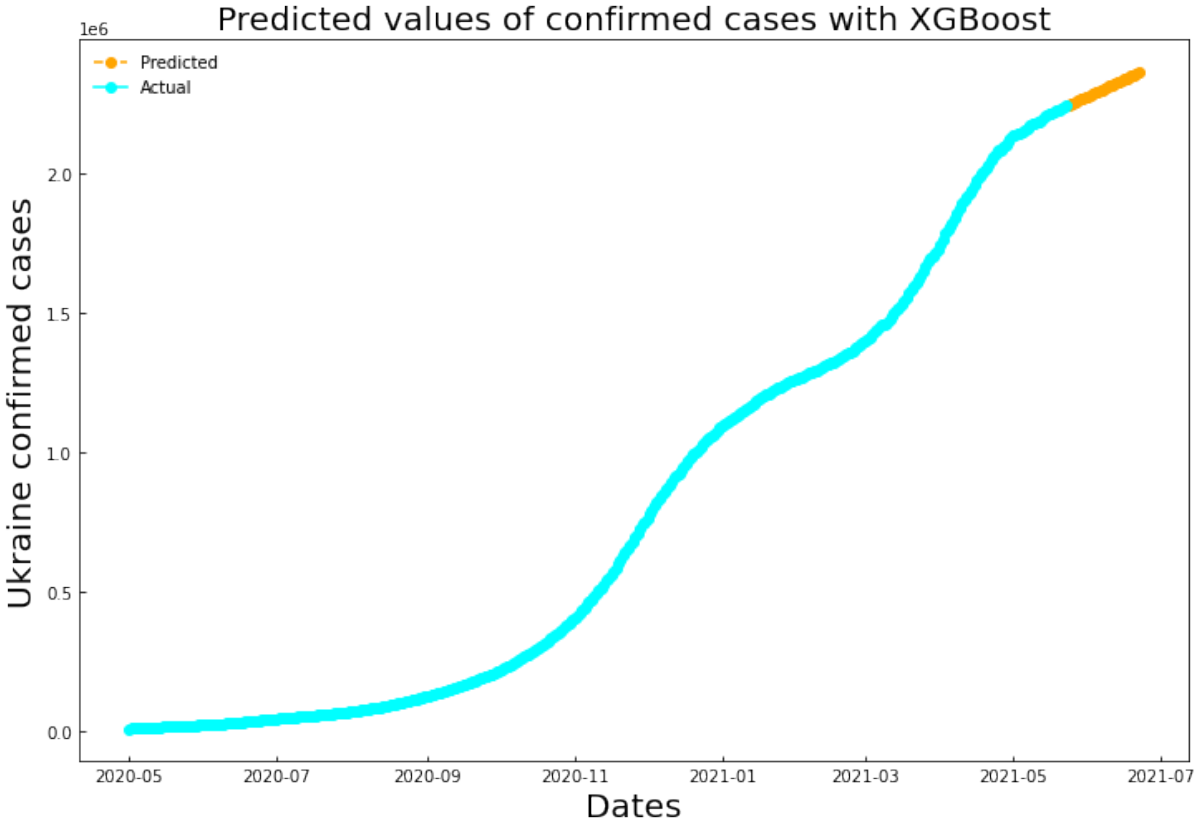
На даному графіку зображено прогнозовані результати на території України.



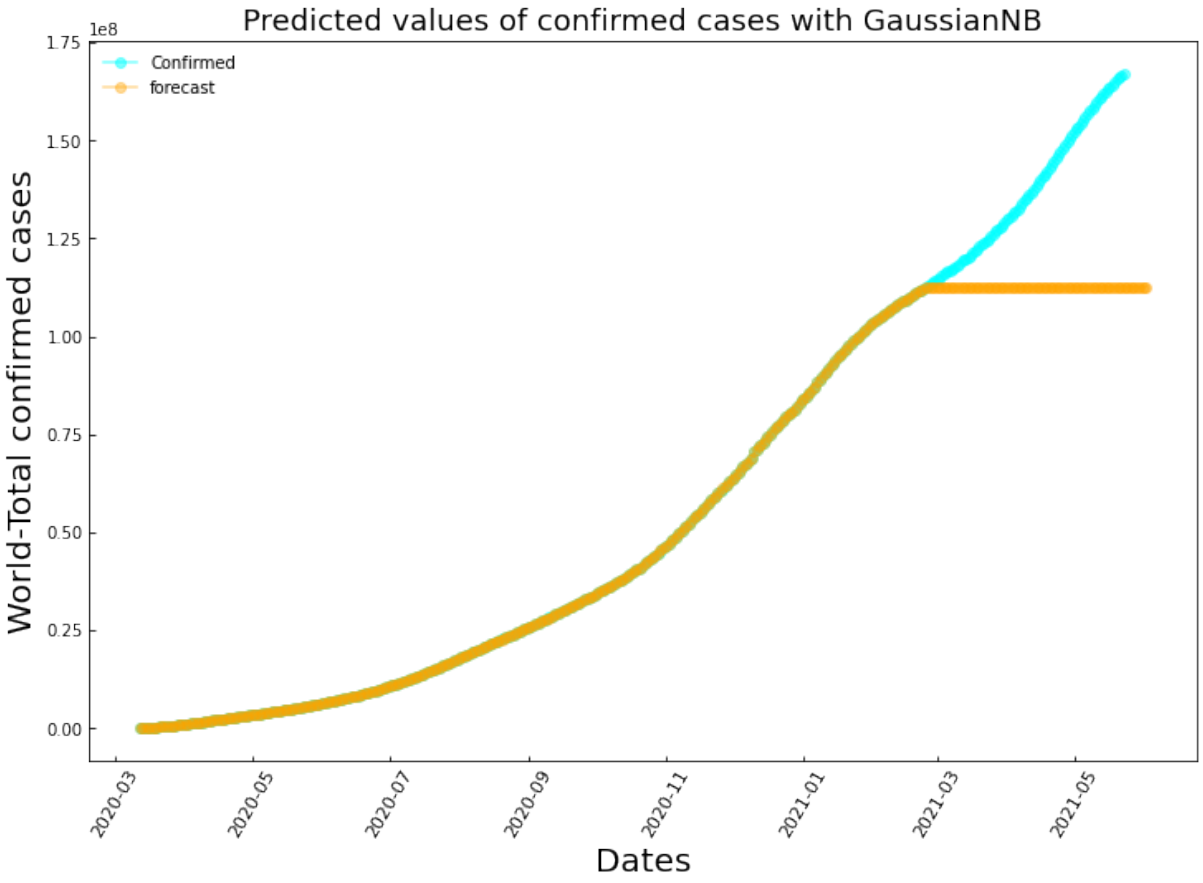
4.2 Метод градієнтного бустінгу над деревами рішень

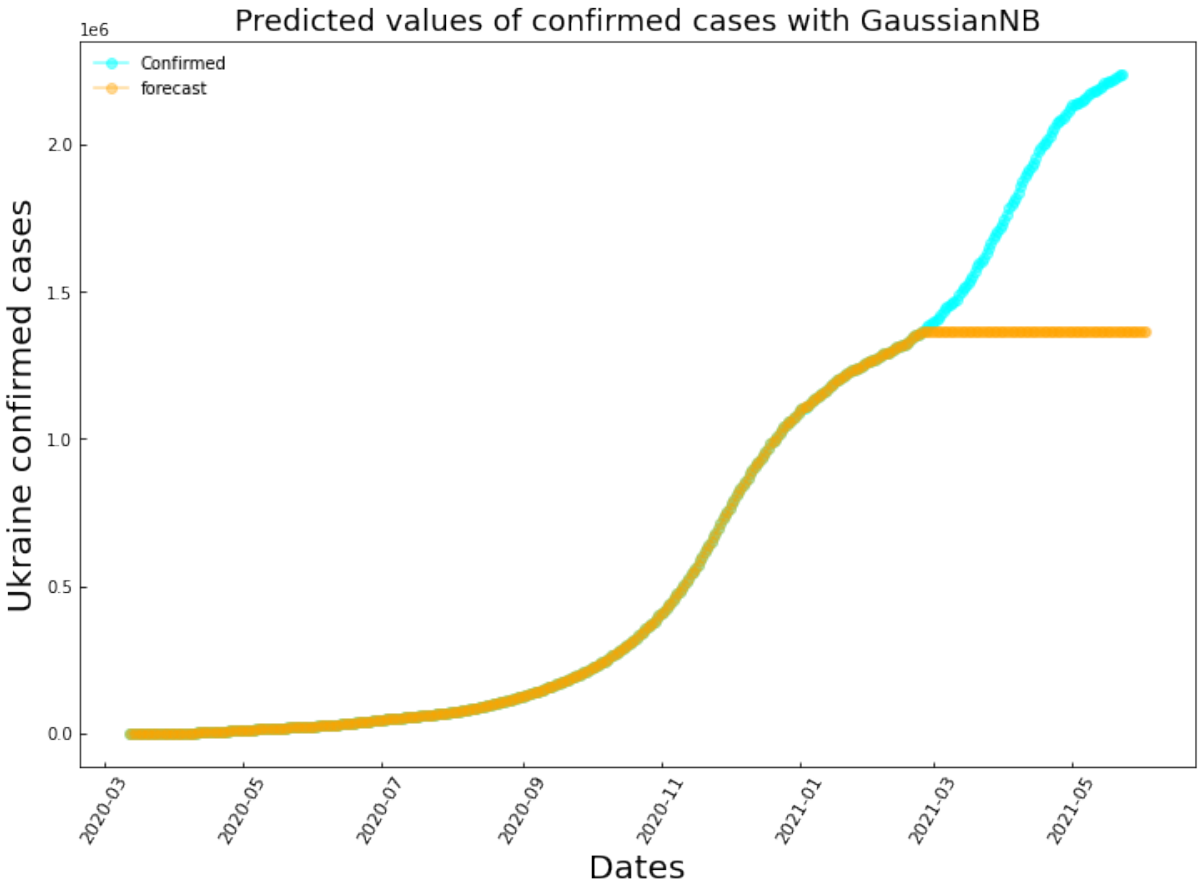
Всі наступні результати будут включати прогнозування підтверджених випадків захворювання для України та для всього світу.



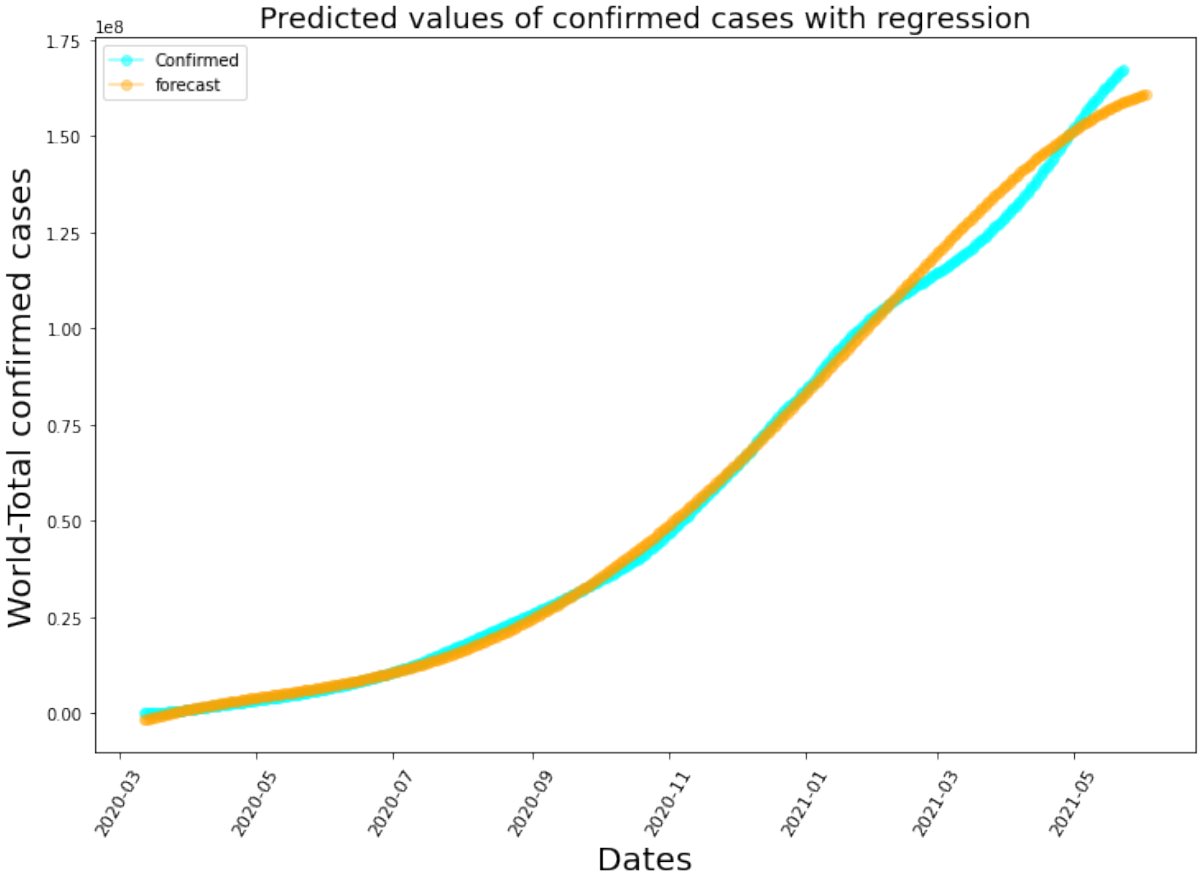


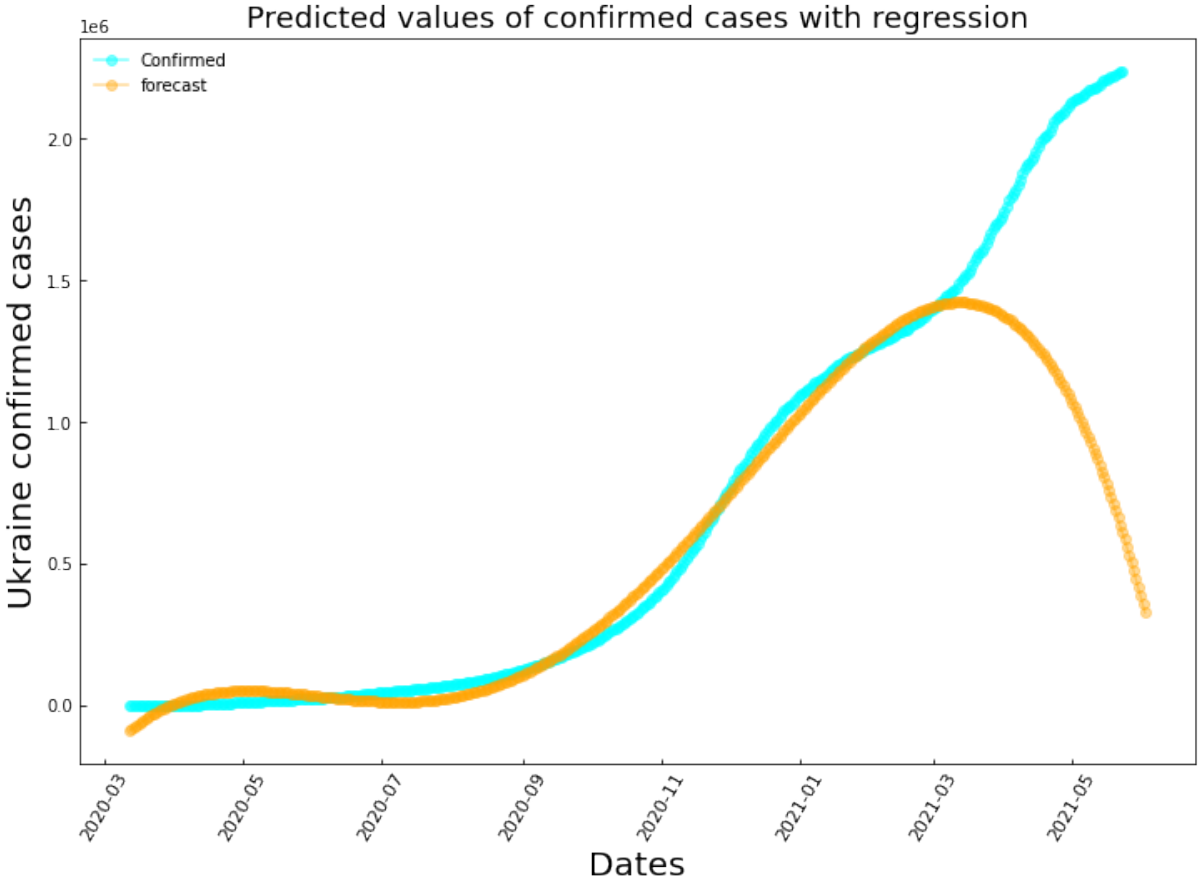
4.3 Наївний байєсовський класифікатор



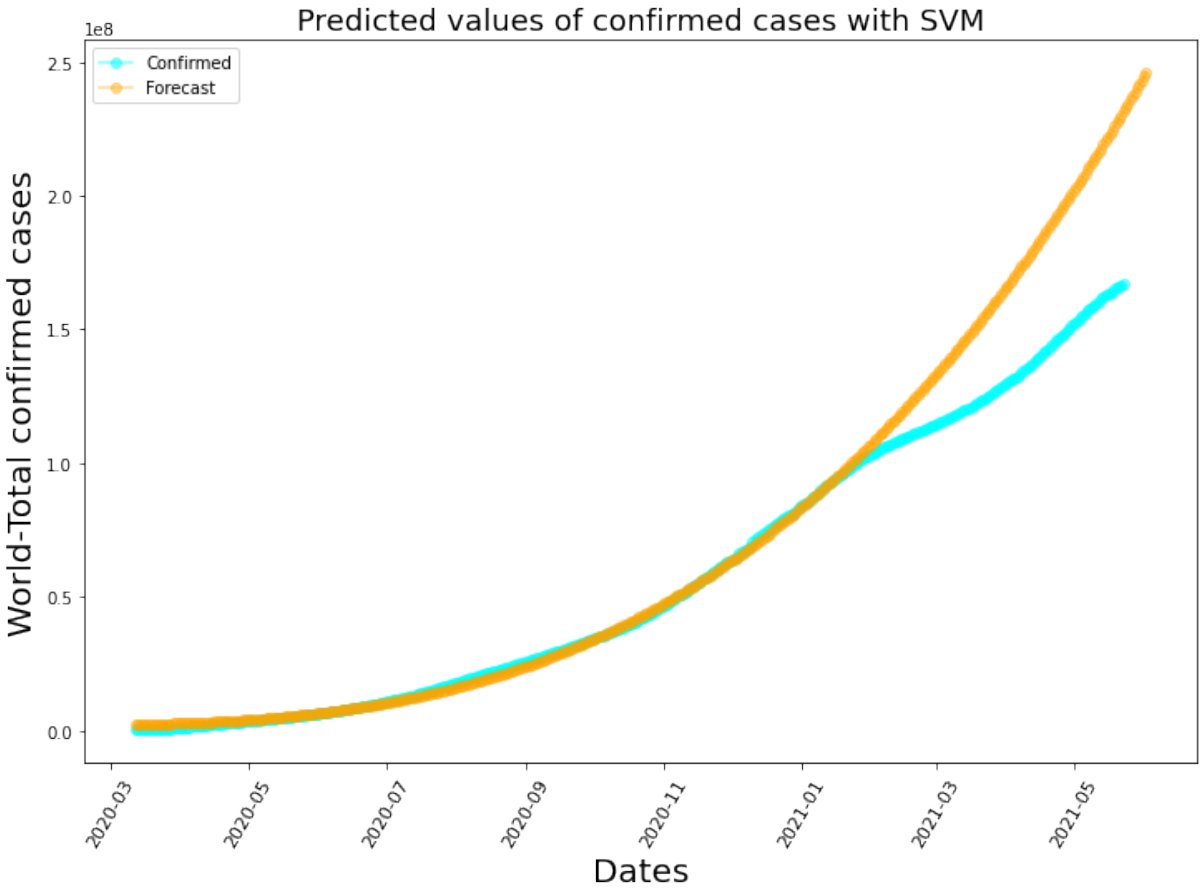


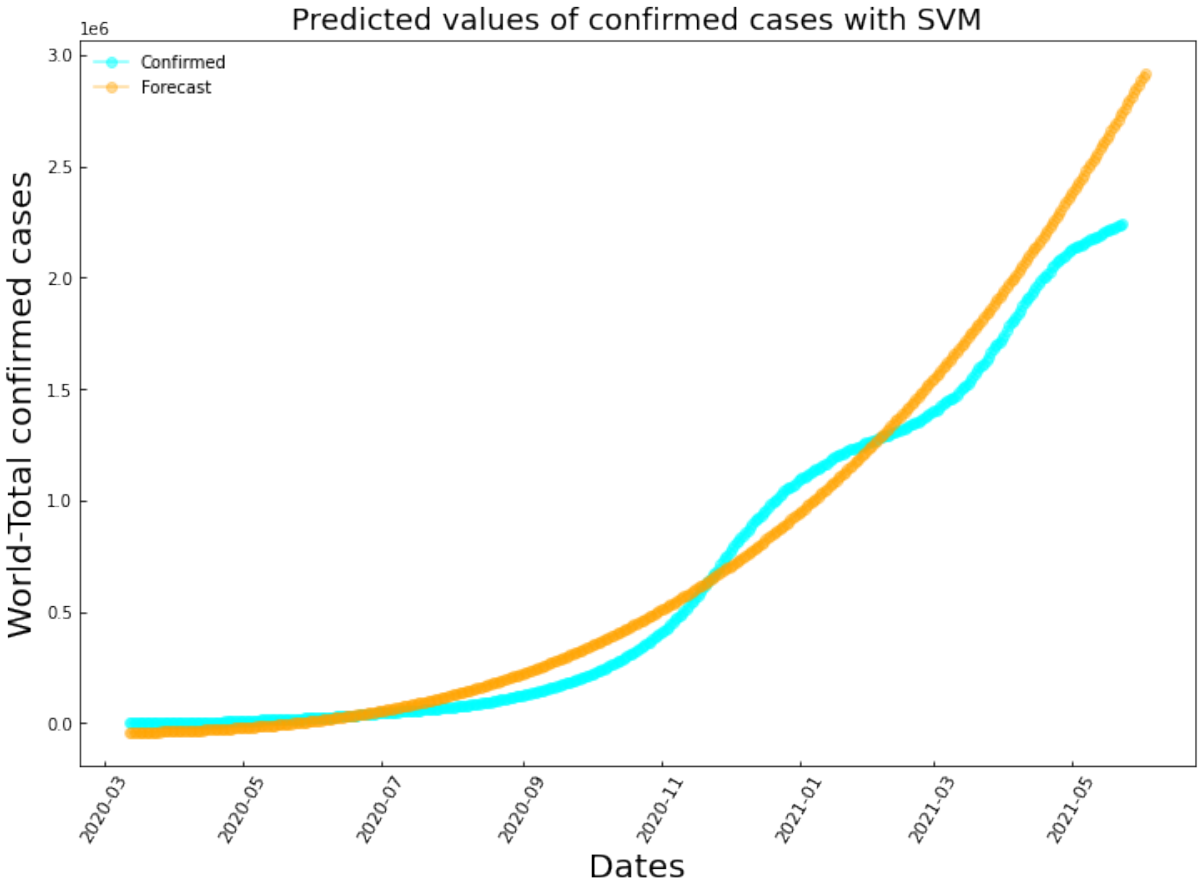
4.4 Логістична регресія





4.5 Метод опорних векторів





5 ВИСНОВКИ

Отже, епідеміологія є важливим питанням для сучасності суспільство. Для математика епідеміологія є джерелом багатьох цікавих проблем та моделей, тоді як для епідеміолога математичне моделювання є важливим інструментом дослідження при вивченні хвороб.

Інфекційні хвороби є основною причиною смерті у всьому світі, і в минулому вбили набагато більше людей, ніж усі війни, наприклад, іспанський грип). Дослідження механізмів розвитку і поширення епідемій є важливим способом боротьби із захворюваннями поряд з пошуком нових ліків, вакцинацією і профілактичними заходами.

Машинне навчання - перспективна та потенційно потужна техніка виявлення та прогнозування захворювань. Методи машинного навчання, в тому числі, де візуалізація та інші потоки даних поєднуються з великими електронними базами даних про здоров’я, можуть забезпечити персоналізований підхід до медицини завдяки вдосконаленій діагностиці та прогнозуванню індивідуальних реакцій на терапію.

Інфекційні хвороби є основною причиною смерті у всьому світі, і в минулому вбили набагато більше людей, ніж усі війни, наприклад, іспанський грип). Дослідження механізмів розвитку і поширення епідемій є важливим способом боротьби із захворюваннями поряд з пошуком нових ліків, вакцинацією і профілактичними заходами.

В даній роботі було опрацьовано дані з 23.01.2020 року до 24.05.2021 року.

Порівнюючи надані методи машинного навчання найкраще спрацював Метод грідєнтного бустінгу над деревами рішень. Модель на кожному кроці оцінює дані за вибраний день в конкретному регіоні - населення, площа, урбанізацію та зміну числа заражених за минулий день - по якомусь параметру, щоб в кінці прийти до найкращих даних. Наприклад, спочатку модель може подивитися, чи був введений карантин у країні. Якщо так, модель йде по першій гілці, якщо немає, то по другій. Далі запитає, чи пройшло більше 30 днів після першого зараженого, і знову відбувається розгалуження, і так далі. Поря-

док цих питань вона визначає сама, намагаючись на кожному кроці зменшити ентропію.

SIR-модель виявилась другою за успішністю. Методи машинного навчання використовують відкриті дані, і в цьому є суттєвий недолік методів. Оскільки немає впевненості, що дані є правдивими, тому відповідно немає впевненості в правильності прогнозованих результатів. Перевага SIR-моделі в тому, що для неї не потрібно використовувати дані, але коефіцієнти моделі також залишаються невідомими, оскільки Covid-19 ще не є цілком дослідженим, тому один з варіантів - займатися підбором параметрів, і в цьому можуть допомогти методи машинного навчання.

Бібліографія

[1] Anderson, R. M. discussion: the kermack-mckendrick epidemic threshold theorem / R. M. Anderson // Bulletin of mathematical biology. — 1991. — Vol. 53, no. 1. — P. 1 – 32.

[2] C. Cortes and V. Vapnik. support-vector networks / C. Cortes and V. Vapnik // Machine Learning. — 1995. — Vol. 20. — P. 273 — 29.

[3] Friedman, J. H. Greedy function approximation: A gradient boosting machine / J. H. Friedman. — 1999. — https://statweb.stanford.edu/~jhf/ftp/trebst.pdf.

[4] Hethcote, H. W. the basic epidemiology models: models, expressions for r0, parameter estimation, and applications / H. W. Hethcote // Mathematical Understanding of Infectous Disease Dynamics. — 2008. — P. 1 – 61.

[5] J. R. Heffernan, R. J. Smith and L. M. Wahl. Perspective on the basic reproductive ratio / J. R. Heffernan, R. J. Smith and L. M. Wahl // J. R. Soc. Interface. — 2005. — Vol. 2. — P. 281 – 293.

[6] J. Tolles and W. J. Meurer. Logistic regression relating patient characteristics to outcomes / J. Tolles and W. J. Meurer. — 2016. — https://jamanetwork.com/journals/jama/article-abstract/2540383.

[7] Mitchell, T. M. Machine Learning / T. M. Mitchell. — McGraw-Hill, 1997.

[8] Murray, J. D. Mathematical Biology / J. D. Murray. — Springer-Verlag, 2002.

[9] P. Domingos and M. Pazzani. on the optimality of the simple bayesian classifier under zero-one loss / P. Domingos and M. Pazzani // Machine Learning. — 1997. — Vol. 29. — P. 103 – 137.

[10] Samuel, A. some studies in machine learning using the game of checkers / A. Samuel // IBM Journal of Research and Development. — 1959. — Vol. 3. — P. 210 – 229.

[11] А. Н. Устинов. К истории эпидемий древнего мира / А. Н. Устинов. — Товарищество типографии А. И. Мамонтова, 1894.

[12] П.Н. Бургасов, А.А. Сумароков. Эпидемия / П.Н. Бургасов, А.А. Сумароков. — 1986. — Vol. 30, no. 3. — P. 544.

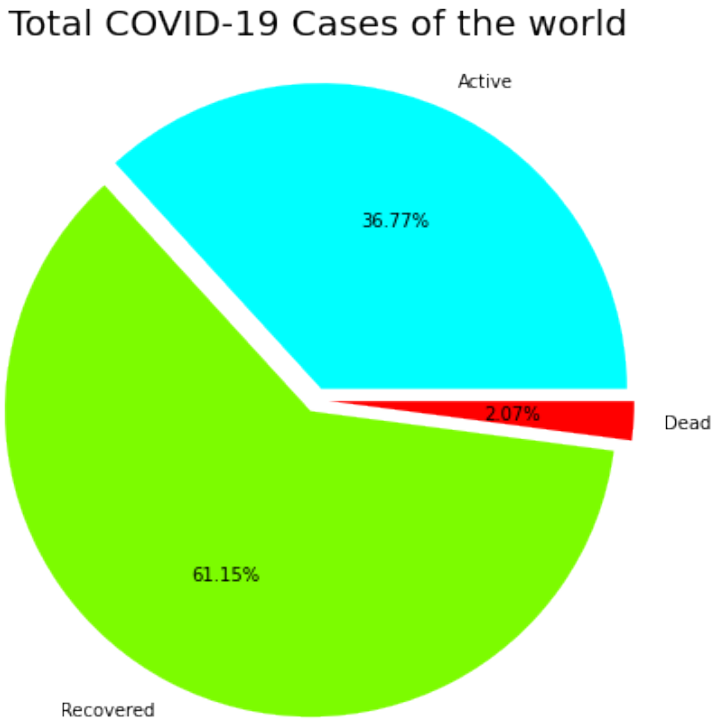
Додаток А Таблиці

Далі наводимо приклад даних станом на 24.05.2021, які використовувались при дослідженнях:

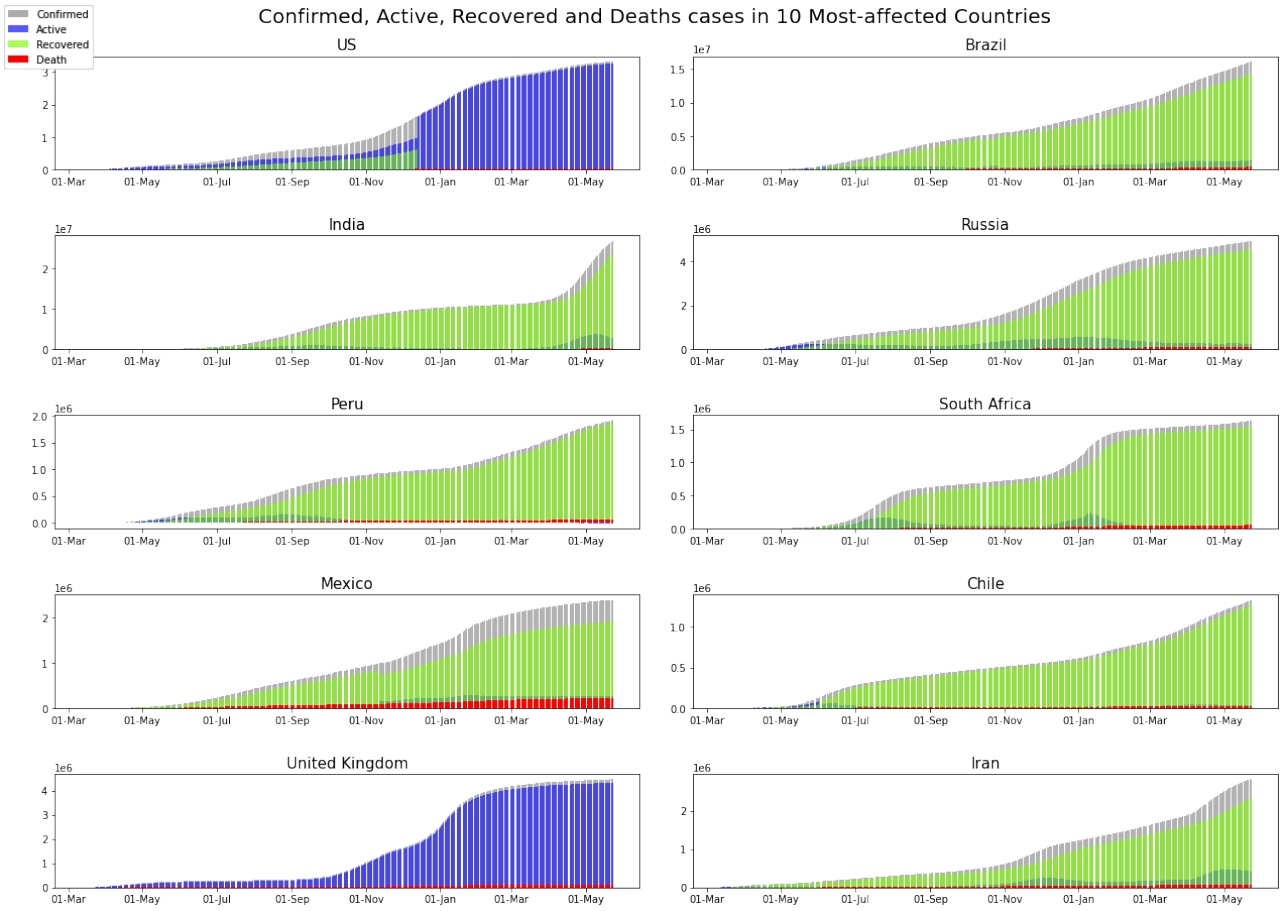
	Date	Country/Region	Confirmed	Deaths	Recovered	Active	New confirmed	New deaths	New recovered	WHO region
93203	2021-05-23	Vietnam	5275	43	2721	2511	156	1	0	WPRO
93204	2021-05-23	West Bank and Gaza	305201	3459	297201	4541	0	0	0	EMRO
93205	2021-05-23	Yemen	6658	1307	3245	2106	9	3	44	EMRO
93206	2021-05-23	Zambia	93201	1268	91156	777	95	1	54	AFRO
93207	2021-05-23	Zimbabwe	38682	1586	36453	643	3	0	8	AFRO

Country/Region	Confirmed	Deaths	Recovered	Active	New confirmed	New deaths	New recovered	Recovery rate(per 100)	Mortality rate(per 100)	WHO region
Afghanistan	65728	2802	56035	6891	242	10	146	85.25	4.26	EMRO
Albania	132209	2444	128732	1033	33	2	131	97.37	1.85	EURO
Algeria	126860	3418	88346	35096	209	7	138	69.64	2.69	AFRO
Andorra	13569	127	13234	208	0	0	0	97.53	0.94	EURO
Angola	32441	725	26778	4938	292	10	3	82.54	2.23	AFRO

Відсоткове значення одужавших, хворіючих та летальних випадків по всьому світу:



Статистика випадків Covid-19 в найбільш постраждалих країнах:



Статистика випадків Covid-19 в Україні: