

Линейная регрессия

Ельцов Данил, Михаил Михайлов

29 ноября 2020 г.

Содержание

1	Постановка задачи	2
2	Используемые данные	2
3	Описание решения	2
3.1	Идея решения	2
3.2	Подготовка данных	2
3.3	Вычисление коэффициентов	2
3.4	Построение регрессионной модели	3
4	Результаты	3

Резюме

По датасетам с Kaggle и официальной статистики по COVID-19 в мире была построена регрессионная модель, способная по демографическим данным страны предсказать кривую развития пандемии в отдельно взятой стране.

На вход модель принимает числовые характеристики конкретной страны и результатом работы программы являются коэффициенты логистической прямой

1 Постановка задачи

Предсказать динамику роста новой коронавирусной инфекции в конкретной стране, основываясь на ее географических и демографических особенностях.

2 Используемые данные

Был взят [датасет](#), содержащий числовые характеристики самых больших стран с Kaggle. Представляет из себя несколько таблиц одинакового формата: **id**, **country**, **country_code**, **feature**

Также был взят [датасет](#), содержащий официальную статистику по развитию коронавируса в разных странах.

3 Описание решения

3.1 Идея решения

В результате анализа темпов развития COVID-19 в различных странах было сделано предположение, что развитие коронавируса в целом происходит согласно логистической кривой. В следствие этого было решено построить регрессионную модель для предсказания её параметров.

3.2 Подготовка данных

Для начала необходимо было объединить все характеристики стран в один CSV-файл, что легко было сделано с помощью их трех-буквенного кода. Затем мы к каждой стране из этой таблицы сопоставили посчитанные для неё коэффициенты логистической регрессии.

3.3 Вычисление коэффициентов

Дифференциальное уравнение процесса выглядит следующим образом

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$$

где P - количество зараженных

r - коэффициент роста

K - поддерживающая емкость среды

Поддерживающую емкость среды, как критическое значение заболевших было решено взять значение, после которого рост устремится к нулю, у нас оно предполагается равным $0.75 * P_{max}$, где P_{max} - число жителей в данной стране.

После применения некоторых алгебраических операций и усреднений мы выводим формулу для расчёта коэффициента роста r

$$r = 2 * \frac{1}{n} \sum_{i=1}^n \frac{dP_i}{P_i}$$

3.4 Построение регрессионной модели

После предсчета параметров регрессии для каждой страны мы запустили обучение модели [LinearRegression](#) из популярной библиотеки для машинного обучения - **sklearn**, которая подобрала коэффициенты регрессионной кривой, минимизирующий средний квадрат ошибки. Мы сохранили данную модель для осуществления дальнейших предсказаний с ее помощью, используя библиотеку [joblib](#).

4 Результаты

- Подборка датасета - Данил Ельцов, Михаил Михайлов
- Анализ данных - Данил Ельцов, Михаил Михайлов
- Построение модели - Михаил Михайлов
- Отчёт - Данил Ельцов
- [Code](#)