

Линейная регрессия

Ельцов Данил, Михаил Михайлов

8 января 2021 г.

Содержание

1	Постановка задачи	3
2	Используемые данные	3
3	Описание решения	3
3.1	Идея решения	3
3.2	Подготовка данных	3
3.3	Вычисление коэффициентов	3
3.3.1	Теоретический расчет параметров	3
3.3.2	Предсказание целевой переменной	4
3.4	Построение регрессионной модели	4
4	Результаты	4

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, такие как средний возраст, индекс урбанизации и проч. и результатом работы программы являются коэффициенты логистической прямой

Постановка задачи

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

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.

Описание решения

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 **параметров логистической кривой**.

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. В результате получился файл **clear_data.csv**, имеющий следующую структуру:

(ISO – code|rfactor|Median – age|Sex – ratio|Urbanization – rate)

Вычисление коэффициентов

Дифференциальное уравнение задающее логистическую кривую выглядит следующим образом

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

где параметры кривой это:

- P - количество зараженных
- r - коэффициент роста
- K - поддерживающая емкость среды

Поддерживающую емкость среды в различных моделях оценивают как число из промежутка $[0.75P_{max}; P_{max}]$ где P_{max} — число жителей в данной стране, поэтому значение K нас не интересует. Таким образом интересует только коэффициент роста r .

Теоретический расчет параметров

Поймем как можно вычислять параметр r . Проведя преобразования над уравнением кривой получаем

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

Проведем усреднение, *предположив*, что выражение в скобках будет равно примерно $\frac{1}{2}$:

$$\left\langle \frac{dP}{P} \right\rangle = \frac{r dt}{2}$$

Теперь положим для удобства приращение $dt = 1$ (так как мы сами выбираем единицы относительно которых считаем), получим:

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

Где P_i, dP_i — общее число больных и число заболевших на i -ый день.

Предсказание целевой переменной

Для предсказания коэффициента r логистической кривой конкретной страны будут использованы её следующие демографические признаки:

- sex_ratio
- median_age
- urbanization_rate

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 методом наименьших квадратов.

Результаты

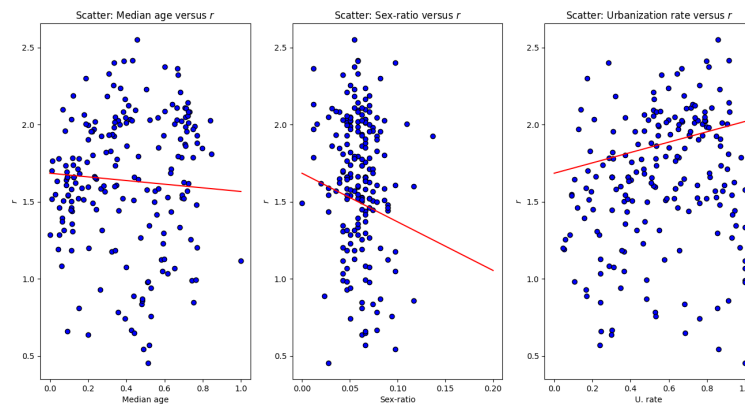


Рис. 1: Результаты работы регрессии

На рисунке 1 представлены нормализованные данные (значение r помножено на 100, рассматриваемые параметры — минмакс нормализация) и прямая регрессии. Видно, что вообще говоря коэффициент r не имеет линейной зависимости от исследуемых параметров. Возможно, это связано с методом подсчета значения r — хотя получились вполне реалистичные числа, такой метод подсчета не имеет качественного обоснования. Возможно, это связано с самой природой числа r и индекс урбанизации, распределение по полу, средний возраст

не влияют на значение параметра r или влияют незначительно. Для проверки роли вклада рассматриваемых параметров в значение r можно рассмотреть другие параметры

На текущий момент вывод таков — **зависимость r от рассматриваемых параметров не является линейной.**

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