Шарапов Р.В.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: sharapov76@gmail.com

Прогнозирование погоды с помощью линейной регрессии

Линейная регрессия – регрессионная модель зависимости одной (зависимой) переменной от другой или нескольких других переменных (независимых переменных) с линейной функцией зависимости.

Модель линейной регрессии основана на выражении:

$$\hat{y} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_{p-n} x_{p-n} + \epsilon$$

где $\hat{y}$ – прогнозируемая результирующая переменная (зависимая переменная);

x_i – переменные-предикторы (независимые переменные) для параметров i=1,2,…,p-n;

β_0 точка пересечения или значение $\hat{y}$, когда все $x_i = 0$;

β_i – изменение в $\hat{y}$, основанное на изменении на одну единицу одного из соответствующих x_i;

ϵ – случайная погрешность, связанная с разницей между прогнозируемым значением $\hat{y}_i$ и фактическим значением y_i.

В приведенном уравнении линейной регрессии очень важен последний член ϵ. Простейшая форма построения модели линейной регрессии основывается на алгоритме обычных наименьших квадратов, находящим такую комбинацию значений β_i, которая минимизирует ϵ.

Линейная регрессия основана на предположении, что между зависимой переменной $\hat{y}$ и каждой независимой переменной x_i имеется линейная зависимость. Для того, чтобы оценить линейную зависимость между независимой переменной и зависимыми переменными используется коэффициент корреляции Пирсона (r) [1].

Коэффициент представляет собой измерение степени линейной корреляции между массивами одинаковой длины. Коэффициент корреляции Пирсона принимает значения в диапазоне от -1 до 1. При этом значения корреляции от 0 до 1 представляют собой сильную положительную корреляцию. Два ряда данных обладают положительной корреляцией если значения в одном ряду данных увеличиваются одновременно со значениями в другом ряду. Значения корреляции от 0 до -1, называются обратно (отрицательно) коррелированными. При увеличении значений одного ряда, соответствующие значения в другом ряду уменьшаются. Когда изменения в величине между рядами становятся равными (с противоположными знаками), значение корреляции приближается к -1. Значения коэффициента корреляции Пирсона, близкие к нулю (как положительные, так и отрицательные), предполагают слабую линейную зависимость, все более ослабевающую по мере приближения к нулю.

При значении коэффициента корреляции Пирсона 0,8–1,0 корреляция считается очень сильной, 0,6–0,8 сильной, 0,4–0,6 умеренной 0,2–0,4 слабой, 0,0–0,2 очень слабой [2].

Был проведен корреляционный анализ зависимости целевой переменной (температуры) от значений различных переменных-предикторов. Значения коэффициента корреляции были отсортированы от наиболее отрицательно коррелированных до наиболее положительно коррелированных.

Как можно заметить, наиболее коррелированы с целевой переменной значения максимальной, минимальной и средней температуры за прошедшие дни, а также максимальные, минимальные и средние значения точки росы. Корреляция постепенно снижается по мере отдаления от рассматриваемой даты: значения за 1, 2 и 3 дня назад более коррелированы с целевой переменной, чем значения за 4 и 5 дней назад [3].

Коэффициенты корреляции (по модулю) для разности температур, точек росы и давления, а также осадков, максимальных, минимальных и средних давлений, максимальной влажности лежат в диапазоне 0.2. Поэтому влияние этих параметров на целевую переменную минимально.

Значение коэффициента корреляции для минимальной влажности близко к -0.5, что говорит об умеренном влиянии этого параметра на целевую переменную. Однако, знак «-» указывает на то, что имеет место обратная корреляция.

В работе и использовалась реализации линейной регрессии LinearRegression из библиотеки Scikit-learn, реализованной для Python [4].

Весь набор данных был разделен на две части – тренировочную (80% всех данных) и тестовую (20%). На тренировочной части линейная регрессия обучалась, а на тестовой проводилась оценка качества предсказаний погоды.

Для оценки качества предсказаний использовались следующие метрики:

- R^2 – коэффициент детерминации:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}|a_i - y_i|^2}{\sum_{i=1}^{m}|\bar{y} - y_i|^2}$$

где $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$

- MAE (Mean Absolute Error) – средний модуль отклонения:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|a_i - y_i|$$

- MedAE (Median Absolute Error) – средняя абсолютная ошибка:

$$MedAE = median(|a_1 - y_1|, \dots, |a_m - y_m|)$$

Метрики подсчитывались средствами, предоставляемыми пакетом Scikit-learn.

Результаты точности предсказаний погоды при использовании различных свойств

Признаки	R^2	MAE	MedAE
72 признака (5 дней)	0.95	1.92	1.46
42 признака (3 дня)	0.95	1.89	1.52
27 признаков (2 дня)	0.95	1.91	1.48
12 признаков (1 день)	0.95	1.92	1.47
35 признаков (|r|=0.5, 5 дней)	0.95	1.92	1.47
21 признак (|r|=0.5, 3 дня)	0.95	1.91	1.51
30 признаков (|r|=0.6, 5 дней)	0.95	1.93	1.47
18 признаков (|r|=0.6, 3 дня)	0.95	1.9	1.49
6 признаков (|r|=0.6, 1 день)	0.94	1.93	1.44
7 признаков (|r|=0.9)	0.94	2.06	1.56
1 признак	0.93	2.18	1.76

Таким образом, использование линейной регрессии для прогнозирования погоды дает вполне приемлемые результаты. Средняя абсолютная ошибка составляет около 1.5 градусов. Конечно, точность предсказаний пока невелика. Тем не менее, регрессионные методы не требуют таких вычислительных ресурсов как методы, построенные на глобальном моделировании атмосферы (GFS, ECMWF) [5, 6]. Кроме того, методы линейной регрессии неплохо справляются со своей задачей даже при использовании небольшого числа свойств (переменных-предикторов), взятых за предшествующий прогнозу день.

Литература

1. McQuistan A. Using Machine Learning to Predict the Weather: Part 1 // https://stackabuse.com/using-machine-learning-to-predict-the-weather-part-1/

2. McQuistan A. Using Machine Learning to Predict the Weather: Part 2 // https://stackabuse.com/using-machine-learning-to-predict-the-weather-part-2/

3. Шарапов Р.В. Использование линейной регрессии для прогнозирования погоды // Машиностроение и безопасность жизнедеятельности, № 1, 2021. – С.47-55.

4. Scikit-learn. Machine Learning in Python [Электронный ресурс]. Режим доступа: https://scikit-learn.org/stable/

5. Шарапов Р.В. К вопросу автоматизации обработки данных погодных наблюдений // Машиностроение и безопасность жизнедеятельности, № 1, 2018. – С.53-56.

6. Шарапов Р.В. Математические модели для составления прогнозов погоды // Машиностроение и безопасность жизнедеятельности, № 4, 2018. – С.24-31.