УДК 004.9:791.43

ДОСЛІДЖЕННЯ МЕТОДІВ ПРОГНОЗУВАННЯ ПРИБУТКОВОСТІ ФІЛЬМІВ ТА СЕРІАЛІВ

Волоховський В. Є.
Науковий керівник – доц. Назаров О. С.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
тел.: +38 (068) 438-87-08, e-mail: vitalii.volokhovskyi@nure.ua

The purpose of the research is to determine the methods of forecasting the profitability of films and series in the modern market conditions of the film industry. As a result of the research, we chose a feed-forward neural network and a polynomial regression model as methods for solving the forecasting problem, we implemented these models using the python language, and compared their effectiveness. It was determined that the neural network works faster and more accurately than the polynomial regression model and can model more complex relationships in the data.

Можливість спрогнозувати потенційний прибуток від фільму або серіалу може допомогти кінокомпанії заробити або заощадити багато грошей. Проте більшість з них не використовують сучасні методи аналізу великих даних. Оскільки кількість знятих фільмів та серіалів велика, а на прибуток впливають різні фактори, прогнозування є складною задачею. Тому для вирішення цієї проблеми є доречним використання методів машинного навчання.

Для вирішення поставленої задачі було обрано нейрону мережу прямого зв'язку та поліноміальну регресійну модель, оскільки об’єкт дослідження має багато характеристик різних типів, а характер залежності між вхідними та вихідними значеннями нелінійний.

Розглянемо математичне представлення нейронної мережі прямого зв’язку. Нехай задано набір вхідних даних у вигляді пар вхідного вектору-харатеристик $\overrightarrow{x_i}$ та вихідного значення y_i:

$$X = \{(\overrightarrow{x_1}, y_1), \dots, (\overrightarrow{x_i}, y_i)\}, i = \underline{1, m} \quad (1)$$

де $\vec{x} = (x_1, \dots, x_j), j = \underline{1, n}$ – вектор характеристик об’єкту дослідження,

n – кількість характеристик, *m* – кількість елементів тренувальної вибірки.

Задачею навчання є визначення такої функції *h*, щоб *h(x)* давала значення максимально наближене до вихідного значення *y*.

Функція *h(x)* задається наступним чином:

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x \quad (2)$$

де x_i – вектор характеристик *i*-го елементу, θ_i – вектор вагів (weights).

Для визначення різниці значення функції *h(x)* від очікуваного значення *y*, використовують функцію втрат:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} (h(x_i) - y_i)^2 \quad (3)$$

Для мінімізації функції втрат використовується метод градієнтного спуску [2]:

$$\theta_{j+1} = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (4)$$

де α – параметр швидкості навчання (learning rate).

У якості функції активації використаємо функцію ReLU.

Розглянемо математичне представлення поліноміальної регресійної моделі. Використаємо початкові умови з виразу (1). Тоді для кожної пари вектору характеристик $\overrightarrow{x_i}$ та вихідного значення y_i [3]:

$$h(x) = \beta_0 + \sum_{l_1=1}^{n} \beta_{l_1} x_{l_1} + \sum_{l_1=1}^{n} \sum_{l_2=1}^{n} \beta_{l_1 l_2} x_{l_1} x_{l_2} + \cdots + \sum_{l_1=1}^{n} \sum_{l_2=1}^{n} \ldots \sum_{l_p=1}^{n} \beta_{l_1 l_2 \ldots l_p} x_{l_1} x_{l_2} \ldots x_{l_p} \quad (5)$$

де p – порядок поліному, β_i – коефіцієнти поліному.

Для визначення різниці між розрахованим значенням функції та очікуваним, використовується метод найменших квадратів:

$$LSE(x, y) = \sum_{i=1}^{m} (y_i - h(x_i))^2 \quad (6)$$

Аналогічним чином, за допомогою методу градієнтного спуску, відбувається налаштування коефіцієнтів моделі.

Дані для аналізу було взято з бази даних Full MovieLens Dataset [4].

В результаті дослідження було визначено, що:

– нейронна мережа дозволяє моделювати більш складні нелінійні залежності у даних та прогнозує прибуток фільмів краще (коефіцієнт детермінації $R^2 = 0.75$), ніж поліноміальна регресія ($R^2 = 0.69$);

– поліноміальна модель не може працювати з великою кількістю характеристик, оскільки кількість вихідних параметрів моделі збільшується поліноміально відносно кількості вхідних параметрів та експоненціально відносно порядку полінома;

– час навчання нейронної мережі (11 сек) менший порівняно з регресійною моделлю (69 сек) для тієї самої кількості вхідних параметрів;

– нейронна мережа менш схильна до перенавчання через меншу кількість вхідних параметрів.

Список використаних джерел:

1. The Numbers. (2022, 12 грудня). Domestic Movie Theatrical Market Summary 1995 to 2022. https://www.the-numbers.com/market.

2. A. Ng. (2022, 12 грудня). CS229 Lecture notes. Supervised learning. https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pd.