Mobile Phone User Behavior Prediction Base on Multivariable Linear Regression Model

Qing YANG, Pei-ling YUAN*, Qing ZHANG and Shi-jue ZHENG

Department of Computer, Central China Normal University, Wuhan, China

*Corresponding author

Keywords: Multiple regression, Gradient descent, Application recommendation.

Abstract. With the development of the Internet industry and the smart phone industry, the amount of mobile application software is increasing and the variety is becoming more and more abundant. The ensuing question is how mobile users can choose the applications they are interested in among many mobile applications. Therefore, mobile application recommendation service for mobile phone users came into place. This paper takes mobile phone users as the starting point, uses multiple linear regression models, uses the commonly used least mean square to evaluate the error, and uses the batch gradient descent method to optimize the training analysis error. Through experiments on the usage data of a large number of mobile phone users, the relationship between gender, age, mobile phone model, mobile phone usage location and app category of the mobile phone user is analyzed, and the mobile phone user app category is predicted to personalized recommend apps for the mobile phone user.

Introduction

In recent years, with the rapid development of Internet and smart phones, mobile application market has been booming. According to App Annie, in the second quarter of 2018, global app downloads and user spending reached new highs. The total number of apps downloaded by iOS and Google Play worldwide exceeded 28.4 billion, which is a 15% year-on-year increase. Surprisingly, this number only includes the first downloads and not include reinstallation or app updates. Facing a large number of a wide variety of apps, how mobile phone users choose their favorite applications in many mobile applications has become a problem, which worth a further discussion[1]. In response to the user's habits, daily preferences, actual needs, etc., the app is recommended for users in a variety of apps that are safe and reliable and in line with the user's interests. This is an important area of the current App recommendation system[2].

In 1870, British biologist statistician F. Galton introduced the concept of regression line and correlation coefficient when he studied the inheritance of human body length, and initiated regression analysis[3]. Wang Mingchen, Liu Hongming and others applied multiple regression to the analysis of the influencing factors of radiation pneumonitis around 2002 to guide the development of lung cancer treatment plans[4]. Olweny and Shiphod used multiple linear regression methods to analyze listed banks in order to analyze the business performance of commercial banks around 2011[5].

The multiple regression analysis prediction method is to establish a prediction model by correlation analysis of multiple independent variables and one dependent variable. When there is a linear relationship between the independent variable and the dependent variable, it is called multiple linear regression analysis [6]. Multiple regression analysis is widely used in many fields. In 2005, Flor A. Castillo and Carlos M studied the potential of symbolic regression generated by GP. In the case of industrial environment, a method for mitigating multiple collinearity in multiple regression was proposed[7]. VillaIgnacio Arnaldo, Krzysztof Krawiec, Una-May O'Reilly applied multiple regression to gene programming around 2014 and proposed a new method for performing genetic programs[8].

Among the model parameters for solving machine learning algorithms, Gradient Descent is a commonly used method. In this paper, we use the multiple regression analysis and gradient descent...
method to optimize the training that analyze the gender, age, mobile phone location, mobile phone model and app relationship to recommend the app for users.

Algorithm Design for Predicting Mobile Phone User Behavior

Establishment of Multivariable Linear Regression Model

There are many influencing factors in predicting mobile phones user behavior. The choice of independent variables is based on the principle of emphasis, comprehensiveness, quantifiability and controllability. It contains effective variables and does not contain invalid variables. In the mobile phone user behavior prediction, the user's gender, age, geographic location, mobile phone model and duration of use can be used as input to the regression analysis model.

This paper proposes the multivariate linear regression algorithm model to predict app category of mobile phone users according to main influencing factors. A linear model is fitted to the given training dataset, and then predict for the new users’ behavior.

Suppose the various input factors (gender, age and other factors affecting the behavior of mobile phone users) in the equation group is represented by X, output factors (mobile application software category) with Y, investigate the relationship between the 4 variable $x_1, x_2, x_3, x_4$ and variable y and choose a total of N-sample test, each test data are $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i}), i=1,2,..., N$.

If y has the following linear relationship with $x_1, x_2, x_3, x_4$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$  \hspace{1cm} (1)

In the above formula, y is the dependent variable, $x_i (i=1,2,3,4)$ is the independent variable and $x_1, x_2, x_3, x_4$ represent gender, age, mobile phone model, mobile phone usage location.

Batch Gradient Descent

In this paper, batch gradient descent is used to solve the regression coefficients of multiple linear regression equations. Batch gradient descent is the most commonly used form of gradient descent method, which is used to update all input samples when updating input parameters.

Algorithm-related parameter initialization: Initialize $\beta_i (i=1,2,3,4)$, algorithm termination distance $\epsilon$. In the absence of any prior knowledge, $\beta_i (i=1,2,3,4)$ is initialized to 0, and the learning rate $\alpha$ is 1, which is optimized when it is tuned later. The main process of the algorithm:

1) Determine the gradient of the loss function of the current position. As for $\beta_i$, the gradient expression is as follows:

$$\frac{\partial}{\partial \beta_i} J(\beta_0, \beta_1, ..., \beta_n)$$  \hspace{1cm} (2)

2) Whether all the distance of gradient descent of $\beta_i (i=1,2,3,4)$ are less than $\epsilon$. If they are less than $\epsilon$, then the algorithm terminates and all current $\beta_i$ is the final result, otherwise go to step 3.

3) Update $\beta_i (i=0,1,2,4)$, its update expression is as follows. Continue to step 1 after the update completed.

$$\beta_i = \beta_i - \alpha \frac{\partial}{\partial \beta_i} J(\beta_0, \beta_1, ..., \beta_n)$$  \hspace{1cm} (3)

Model Test

In practical problems, it is generally impossible to determine in advance whether there is a linear relationship between the dependent variable $y$ and the independent variables $x_1, x_2, ..., x_m$. Before establishing a multiple linear regression equation based on actual observations of dependent variables and multiple independent variables, the linear relationship between dependent variables and multiple independent variables is only a hypothesis, although this assumption is often not unfounded. After the establishment of the multiple linear regression equation, the hypothesis of the
linear relationship between the dependent variable and the multiple independent variables must also be tested for significance. This paper uses the F test method.

Test whether there is a significant linear regression relationship between \( y \) and \( x_1, x_2, ..., x_{m-1} \), that is, test the hypothesis:

\[
H_0: \beta_1 = \beta_2 = ... = \beta_{m-1} = 0 \\
H_1: \text{There's at least one } \beta_i \neq 0 \ (1 \leq i \leq m-1)
\]

If \( H_0 \) is true, then \( y = \beta_0 + \varepsilon \), there is not linear regression relationship between \( y \) and \( x_1, x_2, ..., x_{m-1} \), construct the following test statistic based on analysis of variance:

\[
F = \frac{V_R}{V_e}
\]

In the above formula, \( V_R = \frac{S_R}{m-1}, V_e = \frac{S_e}{n-m}, S_R = \beta^TX^TY - \frac{1}{n}Y^TY, S_e = Y^TY - \beta^TX^TY \).

\( J \) represents an n-order matrix with an element of one. \( S_R \) is the squares sum of regression, reflecting the median deviation of the linear fit values from their mean values, that is, the fluctuation of \( y_i \) (i=1,2,...,n) caused by the change of the variables \( x_1, x_2, ..., x_{m-1} \). The larger the \( S_R \), the greater the proportion of the volatility of \( y_i \) described by the linear regression relationship, that is, the more linear the relationship between \( y \) and \( x_1, x_2, ..., x_{m-1} \). The degree of freedom of \( S_R \) is m-1.

\( S_e \) is the square sum of residuals, reflecting the data fluctuations caused by factors other than the linear relationship between \( y \) and \( x_1, x_2, ..., x_{m-1} \). The larger the \( S_e \), the larger the deviation between the observed value and the linear fitting value. The degree of freedom of \( S_e \) is n-m.

When \( H_0 \) is true, \( F \sim F(m-1, n-m) \) can be proved. When \( H_0 \) is false, the value of \( F \) has a tendency to be large. Therefore, give the significant level \( \alpha \), check the F distribution to obtain critical value \( F_\alpha(m-1, n-m) \), calculate the observation value \( F_0 \) of \( F \). If \( F_0 < F_\alpha(m-1, n-m) \), accepting \( H_0 \) assumption, that is, under the significance level \( \alpha \), the linear relationship is considered not significant; if \( F_0 \geq F_\alpha(m-1, n-m) \), rejecting \( H_0 \), that is, there is a significant linear relationship between \( y \) and \( x_1, x_2, ..., x_{m-1} \).

**Experimental Design**

**Data Introduction**

The experimental data comes from the TalkingData SDK integrated in the mobile app, which serves the service period between TalkingData and mobile app developers. Full recognition and consent from individual user of those apps have been obtained, and appropriate anonymization have been performed to protect privacy. The user attribute data includes event records of the user using the mobile phone, such as download application, location information, application type, application event, etc., and the user's own attributes include gender and age. The data set records the usage time, usage location, app category, etc. of each application (with app_id as a unique identifier) for each user (with device_id as a unique identifier). The specific description is shown in Table 1.

**Multiple Linear Regression Prediction for Mobile User Applications**

The algorithm in this paper is programmed in python, and all experiments are run on the spyder3.0.2 (anaconda3) software under win7 system.

The data set has been standardized. In the experiment, 90% of the data samples of each category are used as the training data set, and 10% of the data samples are used as the test data set. Set \( y: \text{app_id}; x_1: \text{gender}; x_2: \text{age}; x_3: \text{device_id}; x_4: \text{timestamp} \). Input the data samples into the multiple linear regression model to derive the regression equation:

\[
y = -0.447 + 0.603x_1 + 0.762x_2 + 0.059x_3 + 0.210x_4
\]

Logically, the category of mobile applications is primarily determined by the user's own attributes and external attributes. It can be reflected from the above regression equation that the
independent variable $x_1$ and the independent variable $x_2$ have a greater effect on the dependent variable $y$, and the independent variable $x_2$ has the greatest influence on the predicted value $y$, and the independent variable $x_4$ has the least influence on $y$. That is, the user's age and gender are the key factors that determine the user's mobile app usage category.

Table 1. Description of test data of mobile phone users.

<table>
<thead>
<tr>
<th>variable number</th>
<th>variable name</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>device_id</td>
<td>User's id (unique) of the mobile device</td>
</tr>
<tr>
<td>2</td>
<td>device_model</td>
<td>Mobile phone specific model</td>
</tr>
<tr>
<td>3</td>
<td>app_id</td>
<td>The ID of the app</td>
</tr>
<tr>
<td>4</td>
<td>category</td>
<td>Types of apps: online shopping, games, etc.</td>
</tr>
<tr>
<td>5</td>
<td>gender</td>
<td>User’s gender</td>
</tr>
<tr>
<td>6</td>
<td>age</td>
<td>User’s age</td>
</tr>
<tr>
<td>7</td>
<td>longitude</td>
<td>The longitude of the location where the user uses the phone</td>
</tr>
<tr>
<td>8</td>
<td>latitude</td>
<td>The latitude of the location where the user uses the phone</td>
</tr>
<tr>
<td>9</td>
<td>timestamp</td>
<td>Duration of time</td>
</tr>
</tbody>
</table>

After the regression equation is fitted, the following problems need to be solved: (1) the metric of total effect of the equation given; (2) when the effect of the regression equation is significant, the significant regression test is performed on each regression coefficient, and the independent variables whose effects are not obvious are deleted to optimize the model; (3) the equation significance test is given to judge whether the regression equation is valid or not.

The regression square sum and $S_R$ and residual square sun and $S_e$ can be obtained by formula, and the test statistic $F$ can be obtained by formula, in which $F_\alpha$ is obtained by look-up table, and the results are shown in table 2.

As can be seen from Table 2, since $F = 37.4456 > F_{0.05} = 2.322$, the linear regression is significant.

Table 2. Results of regression variance analysis.

<table>
<thead>
<tr>
<th>source of variance</th>
<th>Deviation squared sum</th>
<th>Degree of freedom</th>
<th>variance</th>
<th>$F$</th>
<th>$F_{0.05}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression</td>
<td>0.014104</td>
<td>4</td>
<td>3.3487e-4</td>
<td>37.4456</td>
<td>2.322</td>
</tr>
<tr>
<td>Remaining</td>
<td>0.014274</td>
<td>148</td>
<td>9.4651e-5</td>
<td>----</td>
<td>3.5627</td>
</tr>
</tbody>
</table>

Summary

In this paper, the multivariate linear regression model is used to predict the mobile app category. Because the multivariate linear test $F=37.4456 \geq 2.322$, it is significant to use multivariate linear regression to analyze the prediction problem, which shows that it is effective and feasible to use multivariate regression designed in this paper to predict and analyze.

In actual situations, there are many factors that affect which type of mobile app users use. In this paper, the users’ attribute characteristics age, gender, usage time, and phone model are selected as dependent variables to predict application category. From the experimental calculation results of this paper, the age and gender of the user's own attributes have a relatively large impact on the use of mobile app. In this paper, the batch gradient descent method is used to solve the equation coefficients. When the sample are large, the training speed will be slower. In the future work, we plan to design more sophisticated method to solve the regression coefficients for multivariable Linear Regression Model.

Acknowledgement

This research was financially supported by National Science and Technology Support Program, under NO.2015BAK33B0303.
References


