Linear Regression from Uncertain Data and its Applications to Housing Price Prediction

Shijian Gao
Department of computer, Kunming University of Science and Technology Oxbridge College, Kunming, China
2138gao@163.com

Abstract: Most of the figures we get in our daily life are imprecise, a person's height is an approximation. Uncertain data is usually represented by many values around it in a limited region with probability density function. This paper first defines the uncertain data, and then proposes a linear regression algorithm based on the uncertain data. In order to verify the rationality, correctness and effectiveness of the algorithm, many experiments were carried out.

1. Introduction
There are a great number of uncertain data in science and engineer, even in our day life, such as the size of house, no matter how accurate the figure is, it’s an approximate figure, rather than an accurate one. Therefore, it is very important to study the nature of uncertain data and its impact on algorithms, and to learn knowledge from data.

Linear regression plays an important role in the field of machine learning, especially in regression analysis, which is sensitive to data; it will gain the imprecise result, even wrong result if training from the dirty or missing data. Therefore, linear regression algorithm is picked to find the effect of uncertain data.

2. Related works
Uncertain data caught a flood of scholars’ attention in recent years, many of them dedicated to explore the essence of uncertain information. For instance, summary the management of uncertain data in data modeling, query processing, and mining in the forms of XML, graph and stream, in addition including frequent patterns mining, classification and clustering[1-2]. However, most of research focused on clustering, representing and managing of uncertain data[3-4], some people interested in detecting outlier from uncertain data based on density[5]. Methods of machine learning from uncertain data have been caught little attraction, using Bayesian method to study on linear regression is beneficial attempt[6].

3. Definitions
In this section, some related definition will be listed as following.

Definition 1. (Uncertain Data). Uncertain data divided into two classes of value of uncertainty and existential uncertainty. We focus on former, which refers to a tuple whose value existence is certain, but whose value is uncertain, usually described by probability density function (PDF) in a limited region. A simple case of uncertain data showed in figure 1. Two values presented in this region, each one with 30 points in [2,3]×[2,3] and [0.5,1.5]×[0.5,1.5] respectively, and each points with a probability, sum of these probability equals to 1.
Definition 2. (Linear Regression Model). In regression analysis, if one variable and another variable have correlation, it means that one of the values depends on the other. For example, the size of house bigger, it costs more money in generally. Linear regression usually describes a formulation as

\[ y = w^T x + b \]  

(1)

Each variable considered as a vector.

Cost function is used to measured how good a linear regression mode is, that denotes the sum of error between prediction value and real value, so its value computing like that

\[
J(w) = \frac{1}{2m} \sum_{i=1}^{m} (y^{(i)} - h_w(x^{(i)}))^2
\]

(2)

Where \( y^{(i)} \) is predicted value and \( h_w(x^{(i)}) \) represents the real value. The goal of this model tries to minimize the cost function, the smaller this number, the more excellent this model.

Definition 3. (Gradient Descent). In order to minimize the cost function, an intuition idea is to setup an initialize \( w \), and declines the value of \( J(w) \) toward min direction. Similar operation as parameter \( b \), two parameters updated simultaneously. The algebra form of this idea as following

\[
w = w - \eta \frac{\partial J(w)}{\partial w}, b = b - \eta \frac{\partial J(w)}{\partial b}
\]

(3)

Where \( \eta \) is learning rate, represents the speed of declining, a very important super parameter in regression, if it is too big, cost function will miss min value, by contrary, min value of cost function will convergence too slow if too small.

4. ULR algorithm

Uncertain data cannot input into linear regression algorithm directly, centroid computed before running the traditional algorithm. Computing centroid is the pre-processing of this algorithm, achieved by each possible value multiply its corresponding probability respectively, and then put the sum of product to the first step of algorithm. The pseudo code of the algorithm shows as following, that fitting uncertain data using linear regression, call ULR algorithm for short.

Algorithm: ULR algorithm

Input: \( m \): Number of features, \( w_0 \): initial weight, \( b_0 \): initial bias, \( \eta \): learning rate, \( n \): iterations, \( x \): features, \( y \): real value set, \( p \): probability set.

Output: \( w \): final weight, \( b \): final bias.

1. \( s = \text{centroid}(x, p, m) \).
2. \( w = w_0, b = b_0 \).
3. for \( i=1 \) to \( n \) do:
   3.1 \( y_{\text{hat}}, \text{loss}, w, b = \text{U_linear_regression}(s, y, w, b) \)
   3.2 \( dw = \frac{1}{m} \sum_{i=1}^{m} \sum p_i (y_{\text{hat}}(i) - y(i))^T, \quad db = \frac{1}{m} \sum_{i=1}^{m} (y_{\text{hat}}(i) - y(i)) \)
   3.3 \( w = w - \eta * dw, b = b - \eta * db \)
4. return \( w, b \)
In this algorithm, first computing the centroid of each feature through sum of each possible value multiply by its probability, and then execute n iterations to minimize the loss function, $y_{\text{hat}}$ is the predicted value, $dw$ and $db$ denotes partial derivative for parameter $w$ and $b$ respectively.

5. Experiments
Experiments executed in Windows 7 with 4GB memory, and performed in Python. Boston Housing Price data set are used to verify the reasonable of present algorithm, the 6th feature RM, denotes the average number of rooms, is selected as feature $x$, the target of data set is real housing price considered as $y$, each feature generated 30 possible value around it in a region of $[x-0.5,x+0.5]$, because there is no feature to measure the of size of house, so it is possible that house with same rooms may be of different size, Therefore, the number of rooms is typical uncertain data.

Table 1 shows the initial value of parameters for algorithm.

| Parameters      | Initial value |
|-----------------|---------------|
| Training data set | 406           |
| Test data set   | 102           |
| Learning rate   | 0.0001        |
| Number of iterations | 3000         |

We know that the learning is the vital element of this algorithm, so the experiment explore the effect of different value of learning rate to the model, in the circumstance of Table 1, we choose three values to find that time of curve of loss function tends to be horizontal when learning rate is 0.0001, is longer than it equals to 0.0005. Furthermore, we come to the conclusions that the model convergence more faster with the learning rate increase, like the traditional algorithm.

![Figure 2. The effect of learning rate](image)

Experiment results variety with different possible value and its probability. In every single set of possible value, ten times perform to get an average number of loss functions. And the effect of number of possible value to the algorithm show as figure 3.
According to above figure, the number of possible value has a little effect on loss, the reason is that each single set of possible value has different probability, which leads to the almost same result.

In the initial experiment, we set the proportion of training set at 80%, now adjusting this parameter to find what happen, for next attempt, training set selected from 50% to 90%, according to the figure 4, we found that loss reaches to the min value at 0.5 and 0.7, and then loss climb up quickly as training set up to 80% and 90%, which imply that too much or too little training set will not obtain good result, because the algorithm have to guarantee that there are enough test set to update the parameters.

6. Conclusions
Uncertainty is the natural quality of figure that obtained from variety of measure means. In this paper, an improved algorithm is proposed, that is, linear regression is used to fit uncertain data, and the new model is applied to Boston housing data set, experiment shows that the uncertainty information indeed affect the result of algorithm, in some cases, it will get the better result because the uncertain information is closer to the reality.

Acknowledgements
The research was supported in part by Key Laboratory of Intelligent Manufacturing in Universities of Yunnan Province, Digital learning Engineering Research Center in University of Yunnan Province, Scientific Research Fund of Yunnan Education Department(#2019J0364), and Research Fund of Education and Teaching Reform of Oxbridge College, KMUST. Finally thank you for the comments of reviewer, especially the advice for grammar revise.

References
[1] Li, L., Wang, H., Li, J. et al. A survey of uncertain data management. Frontiers of Computer Science. 2020(14),162–190.
[2] D. Karne, P. Patil. A survey of Uncertain Data Handling. ICSIT 2014, Vol. 5(3), 2014, 4405-4410.
[3] E. Schubert, A. Koos, T. Emrich et al. A framework for clustering uncertain data. VLDB 2015, 1976–1979.
[4] Wan J. et al. Uncertain Data Clustering Algorithm Based on Voronoi Diagram in Obstacle Space. Journal of Computer Research and Development. 2019, 56(5):977-991.
[5] Park, Y.W., Klabjan, D. Subset selection for multiple linear regression via optimization. Journal of Global Optimization. 2020(77), 543–574.
[6] Wu D, Cheng L.B. Research on Bayesian Method based on Linear Regression Model. Journal of Changchun university of Science and Technology, 2020, 43(1): 128-132.