Logistic Regression with Expert Intervention

Pavla Pecherková and Ivan Nagy

Abstract—**This paper deals with problem of analysis of traffic data. A traffic network has several types of roads: historical centre, peripherals, arterial roads, etc. They have specific properties. For a traffic analysis, large amounts of data are needed. Some traffic data are difficult to obtain due to their rare occurrence. Typical example is the investigation of traffic accidents. In these cases, data from other similar roads can be used. In such cases, an expert intervention added to the general analysis is very important. In this paper, the logistic regression with two types of expert intervention is briefly introduced. The performance of these methods is demonstrated on examples concerning seriousness of traffic accidents.**

Index Terms— **logistic regression, Newton method, traffic accidents**

I. Introduction

THIS paper deals with logistic regression used for analysis of traffic data. Its use in the traffic area, especially in the analysis of traffic accidents, is very broad. Here the modelled variable is discrete – mostly the seriousness of accidents (e.g. minor, serious, with injuries, with loss of lives) that depends on both discrete (visibility: daylight, gloom, dark) and continuous (speed – real value) variables.

Here we are going to tackle rather specific problem: we have a data sample and we perform logistic regression analysis. As a result we produce an output estimation based on the estimated logistic regression model and we consult our result with an expert in transportation. He checks the results – i.e. the regression vectors used and the corresponding estimates of the output and he agrees up to some items which, according to his opinion, should have some reversed form. Our goal is to suggest how to fulfil the requirements of the expert and to leave the model with minimal changes.

The task is similar to that of incorporating the expert knowledge into the model produced in a process of Bayesian estimation. The difference is that the prior knowledge in Bayesian estimation enters the estimation before the measured data, it is usually rather weak and influences the estimates of all model parameters while here the expert knowledge is used after the model estimation from measured data, its effect is demanded with 100% certainty and it concerns only some parameters.

To achieve the mentioned goal, we will demonstrate the following two methods:

1. The use of expert–based relations between regression vectors and the value of the output.
2. Direct enforcement of the bounds between regression vectors and values of the output declared by the expert.

Each method has its pros and cons that will be described in the further text.
This paper will be divided into several parts: (i) traffic problem, (ii) logistic regression with expert intervention (with the description and discussion of both methods), (iii) experiments with traffic data and (iv) conclusion.

II. Traffic Problem

The road traffic is one of the areas where the progress is very rapid. This progress concerns the improvement of the car parameters development of both active and passive elements of traffic safety and also the quality of roads.

The design of new roads or the reconstruction of the old ones is always a compromise between the demands of the drivers that want the roads as short as possible and in a quality as best as possible, and the available budget. The correctness and the effective of the design is reviewed by experts.

The traffic expert is a person who is able, on the basis of the data and his own experience, to design a correct and optimal solution to traffic problems in a specified locality. However, each expert is only a human and if the region where the problems are to be solved is too large or the situation depends on a large number of variables, the human is not able to take into account all the necessary information and can fail. In such case, it is necessary to construct some automatic solution minimizing the damages or maximizing the utility. Present–day computers are able to do such work and the present algorithms are able to construct such solutions based on the whole amount of information gained from the region.

A solution to such traffic optimization is usually based on a mathematical model of the traffic system to be optimized. The model is estimated using a data sample measured on the system and further it is used for prediction (to be able to inform the

This work was supported in by the project GAČR GA15–03564S
.

Pavla Pecherková is with Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 110 00 Prague, Czech Republic and Department of Adaptive Systems, Institute of Information Theory and Automation of the ASCR, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic (e-mail: pecherkova@fd.cvut.cz).

Ivan Nagy is with Faculty of Transportation Sciences, Czech Technical University, Na Florenci 25, 110 00 Prague, Czech Republic and Department. of Signal Processing, Institute of Information Theory and Automation of the ASCR, Pod Vodárenskou věží 4, 182 08 Prague, Czech Republic (e-mail: nagy@fd.cvut.cz).

978-1-5090-1116-2/16/$31.00 ©2016 European Union

traffic operators about the situation that will come) or for direct control (to influence the system so that its behaviour would correspond to our preferences).

An example of such situations that are to be solved is a judgement of traffic accidents with determining which variables from the neighbourhood influence these accidents and with what importance. A traffic expert can help in constructing such model even if he is not familiar with relevant mathematical algorithm or even with the basics of mathematics, at all. However, such a task in a large traffic region where we need to include relations of variables across the whole region is usually not a work only for an expert but it needs an automatic solution.

However, even a correctly constructed model (which is not always the case) can bring results that are not suitable in some aspects. Let us mention a simple example. We are to define maximum allowed speed in a specific point of communication. We are not able to design the solution based on the measured data because we want to prevent accidents, not to wait and measure them as data sample. So, instead of measuring our own data sample, we use some other data measured on a road with similar properties. Such a solution deals with a general problem which does not need to correspond to our situation in each detail. The traffic expert is to judge the submitted solution in every detail and to approve it or to have some objections.

III. Logistic regression

Logistic regression is a well–known offshoot of the ordinary regression for the case that the modelled variable is discrete. It can depend on other discrete or/and continuous variables forming so–called regression vector [2]. In case that input variables are discrete only, we obtain a completely discrete model. However, even in this case, the logistic regression is used if it is necessary to realize a radical reduction of parameters. On the other hand, if the variables in the regression vector are only continuous, the continuous model cannot be used because the output is discrete. In these cases, the solution is to use the model of logistic regression.

A. Model

The logistic model has the form

$$\text{logit}\,(p_t) = x_t b + e_t \qquad (1)$$

where

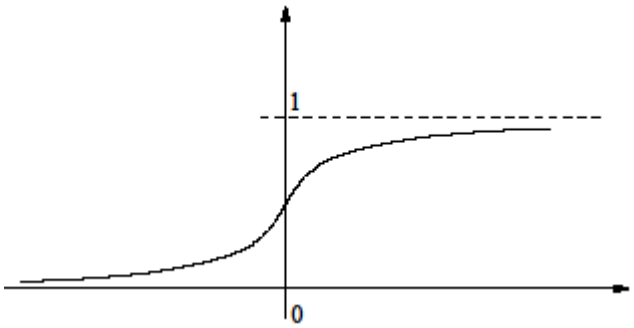
$$p_t = P(y_t = 1 | x_t, b) \qquad (2)$$

and $y_t \in \{0,1\}$ is the model output, $x_t = [1, x_1, x_2, \dots, x_n]$ is the regression vector, $b = [b_0, b_1, b_2, \dots, n_n]'$ is the vector of model parameters, $P(\cdot \mid \cdot)$is the conditional probability function (pf) and $\text{logit}\,(\cdot)$ is the logistic function defined as follows

$$\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right) \qquad (3)$$

It accepts a probabilistic argument $p \in (0,1)$ and returns the result from the whole real axis R.

The main trick hidden in this model is, that it does not model directly the output but the probability that the output is one. Thus, the right–hand side of the equation (1) is an ordinary regression computing a real value, say $z_t \in R$. This value is transformed by the inverse logistic function into the probability interval $(0,1)$. The value of this probability is interpreted according to the relation (2) as a probability that $y_t = 1$. The inverse logistic function is depicted in the following figure:



For the estimation of the model parameters b, it is necessary to construct a likelihood function and numerically search for its maximum. As the first and second derivative of the likelihood can be computed analytically, the most convenient numerical method is Newton's formula for optimization [1].

To perform the output estimation, we can proceed as described above. With the estimated model, we know $\hat{b}$ as the estimates of model parameters. For a specific regression vector x_t we are able to compute the regression

$$z_t = x_t \hat{b} \qquad (4)$$

and to transform it to the probability

$$p_t = \text{logit}^{-1}\,(z_t) \qquad (5)$$

According to its value, we have a point estimate of the output:

$$\hat{y}_t = \begin{cases} 0 & \text{if } p_t \le 0.5 \\ 1 & \text{if } p_t > 0.5 \end{cases}$$

Remark

In this paragraph, we introduced a logistic regression with binary output and we will deal only with it. There is also a possibility to use the logistic regression for the output with more than two values. However, in this case our approach

needs to be much more general. We strongly hope, that some results from this area appear soon.

B. Expert intervention into logistic regression

By the intervention of an expert into the estimated model of logistic regression we mean that the expert denotes some regression vector or several regression vectors for which she/he disagrees with the corresponding output values which were determined by the logistic model. To respect the expert demands, the model has to be changed. As the expert has no objections to other conclusions performed by the model, we want to respect the expert and in the same time to change the model as less as possible.

1) Method one – construction of expert data

This is a very simple method, however, is does not guarantee success in respecting the expert conditions. Let us demonstrate it in a very simple example.

Let us have the following data x and y for training the logistic model, together with the output predictions computed on the basis of the estimated model (Table 1(a)).

TABLE I
LOGISTIC MODEL AND EXPERT DATA

x	y	$\hat{y}$
x_1	y_1	$\hat{y}_1$
x_2	y_2	$\hat{y}_2$
…	…	…
x_i	y_i	$\hat{y}_i$
…	…	…
x_N	y_N	$\hat{y}_N$

(a) original data

x	y	$\hat{y}$
x_1	0	0
x_1	0	0
…	…	…
x_1	0	0

(b) expert data

The expert disagrees with the prediction $\hat{y}_i$ from the model. Say that the prediction is $\hat{y}_i = 0$ and the expert is convinced that to the regression vector x_i, the value of the output should be $\hat{y}_i = 1$.

Then a possibility, how to convert $\hat{y}_i$ from 0 to 1 is to add the "expert data" (Table 1(b)). The larger is the number of the expert data added, the stronger is the demand for fulfilling the expert conditions. However, the precise number of the data added to fulfil the expert demand is not known and must be determined on the basis of experimenting.

2) Method two – enforced fulfilling the expert demand

This method is much more complex. It guarantees the change demanded by the expert and moreover it changes the estimated model in a minimal way. The method will be again demonstrated in an example.

Let us have the same situation as we mentioned in the exposition of the previous method. Again, the expert wants to change $\hat{y}_i$ from 0 to 1.

It is necessary to change the model which means to change the estimates of its parameters $\hat{b}$. So we introduce new parameter estimates $\tilde{b}$ as

$$\tilde{b} = \hat{b} + \beta$$

where β represents the change. For the logistic model computing the estimates of the output (i.e. logistic regression model without noise), now holds

$$\text{logit}\,(p_t) = z_t = x_t(\hat{b} + \beta) \qquad (6)$$

where we denoted the regression part of the model according to (4) by z_t.

Now, for the regression vector x_t we should have the value of the output equal to one. It implies that the corresponding probability p_i must be greater than 0.5. Inserting into (6) we obtain the condition

$$x_t(\hat{b} + \beta) > 0 \qquad (7)$$

where β is a searched variable and the rest are known numbers.

As we want to do as small change in the model as possible, we choose the criterion

$$\beta'\beta = \sum_{j=1}^{n+1} \beta_j^2 \qquad (8)$$

which is to be minimized [1].

So, the task enforcing the expert condition with minimal model change is given as a task of minimization the criterion (8) with the restriction given by the condition (7).

Remark 1
For more values of the output estimate to change the situation is identical, only we have more restricting conditions.

Remark 2
If the demands of the expert are not consistent with the model as a whole, the optimization can lead to degradation of the model. It means that the resulting parameter $\tilde{b} = \hat{b} + \beta$ has all its items practically equal to zero. Such model carries no useful information about the system.

IV. EXPERIMENTS

In this section, the logistic regression is discussed for two types of expert intervention: (i) construction of expert data and (ii) enforced fulfilling the expert demand. At first, the logistic regression without any intervention will be made.

Let us assume that the new peripheral road has been constructed. With respect to the parameters of the road, there is

an increased risk of accidents. The aim is to propose maximum speed to minimize accidents with injury or death. An excessive decrease of the speed allowed is undesirable, because the drivers would go an alternative road, which goes through a city.

For an analysis, the data from other similar roads are used. In the table II (column 1 and 2), the parameters of accidents are shown. The input x means the speed the car drives when the accident occurred. The Output y means seriousness of the accident, where $y = 0$ means no or minor injury in the crashed car and $y = 1$ means serious or fatal injury. It is assumed, that the seriousness of the accident is dependent on speed only. This example is simplified because in the real situation, it depends on several parameters such as visibility, road surface, drivers and weather.

In this example, we have binary logistic regression. Using the data we computed the regression equation in the form

$$z_t = -13.6716 + 0.1823x_t \qquad (9)$$

where $z_t > 0$ means that output $y_t = 1$ and $z_t \leq 0$ means that $y_t = 0$.

The model and the original data (Table II), help to estimate the output (assumed seriousness of the accident). This data give us the information that "secure" speed is 70 km/h. The passengers are slightly injured or unharmed during an accident with this speed.

TABLE II
ASSUMED SERIOUSNESS OF ACCIDENT ($\hat{y}$)

x	y	z	$\hat{y}$
20	0	-10.0263	0
30	0	-8.2004	0
40	0	-6.3810	0
50	0	-4.5583	0
60	{0, 0}	-2.7324	0
70	{1, 0}	-0.9130	0
80	{0, 1}	0.9097	1
90	{1, 1}	2.7324	1
100	1	4.5551	1
120	1	8.2004	1

Expert opinion:

An expert has objections and disagrees. Fixed obstacles are near the new road and so the seriousness of the injury will be much higher. The traffic expert says that the "secure" speed is about 30% lower. The next task is to change (increase) the seriousness of the accidents from the speed 50 km/h. In this sense it is necessary to change the regression model (i.e. its coefficients).

A. Method one – construction of expert data

This method is based on the principle that the expert information is used as data. Expert data: speed 50 $km/h \rightarrow$ serious or fatal injury $\rightarrow y = 1$. This new data are added several times (Table III.). For this case, expert data had to be added at least 7 times to achieve the required result.

TABLE III
SPEED (x) AND SERIOUSNESS OF THE ACCIDENT (y) WITH EXPERT DATA

Original data				Expert data	
x	y	x	y	x	y
20	0	30	0	50	1
40	0	70	0	50	1
60	0	80	0	50	1
70	1	100	1	50	1
100	1	50	0	50	1
120	1	90	1	50	1
80	1	90	1	50	1

This method is simple but the result is not guaranteed. In case that the new data will be added (speed or other parameters such visibility, etc.), larger number of expert data must be constructed and added. The new regression is

$$z_t = -1.9572 + 0.0411x_t \qquad (10)$$

The slope parameter $b_1 = 0.0411$ and there is a risk because there is not any evidence that a change in x is associated with a change in y.

B. Method two – enforced fulfilling the expert demand

This method is based on a principle that the coefficients of the model are recomputed with respect to the equation (6). To the original regression the new coefficients are added:

$$z_t = (-13.6716 + \beta_0) + (0.1823 + \beta_1)x$$

A Newton optimization method [1] was used for finding a local minimum for parameter β. In this case, the parameters are $\beta_0 = 0.001825$ and $\beta_1 = 0.091296$. The new regression is:

$$z_t = -13.6700 + 0.2736x_t \qquad (11)$$

With respect to z, the new assumed "unsafe" speed is 50 km/h (Table IV.). It is evident that $z \approx 0$ for speed 50 km/h and b_1 is still sufficiently large. Graph of logistic regression curve showing the probability of minor or serious/fatal injury, see Figure 1. The black cross indicates original data. The blue curve is $logit$ for data without expert intervention. The red one is after expert intervention. Its

position is more to the left to meet the requirement of the expert. Also, its course is steeper.

Obviously, the expert requirements are not without limits. If e.g. her/his demand it that the serious accident occurs simultaneously at low as well as in high speeds, the enforced logic to the problem cannot be met and, again, the resulting parameters are near to zero, which means that the informational content of the model is suppressed. In such case, the solution is iterative until the inner contradiction is removed.

V. CONCLUSION

The presented paper concerns some methods of enforcing expert information into the logistic model after its estimation from the measured data. The first method recollects a possibility of using so called expertly constructed data vectors $[y_i , x_i]$ respecting the expert demands. The second method is newly introduced. It changes the logistic regression model so that it meets the expert knowledge and at the same time it differs from the estimated model as less as possible. Both the methods are demonstrated on simulated examples.

The demonstrated theory is just the first step to the overall goal – embedding the expert knowledge the multivariate logistic regression (i.e. with the output taking more than two possible values) and prior analysis of the expert knowledge excluding its inner contradictions. This work will be, hopefully, published soon.

REFERENCES

[1] P.E. Gill, W.A. Murray, and W. Murray., “Numerical methods for constrained optimization.”, Academic Press, 1974

[2] David W. Hosmer and Stanley Lemeshow. “Applied logistic regression”, John Wiley & Sons, Inc. A Wiley-Interscience Publication, New York, Chichester, Weinheim, 2000.