Let Us Use Negative Examples in Regression-Type Problems Too

Jonatan Contreras[1], Francisco Zapata[2], Olga Kosheleva[2],
Vladik Kreinovich[1], and Martine Ceberio[1]
[1]Department of Computer Science
[2]Department of Department of Industrial,
Manufacturing, and Systems Engineering
[3]Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
jmcontreras2@utep.edu, fcozpt@outlook.com, olgak@utep.edu,
vladik@utep.edu, mceberio@utep.edu

Abstract

In many practical situations, we need to reconstruct the dependence between quantities x and y based on several situations in which we know both x and y values. Such problems are known as *regression* problems. Usually, this reconstruction is based on *positive examples*, when we know y – at least, with some accuracy. However, in addition, we often also know some examples in which we have *negative* information about y – e.g., we know that y does not belong to a certain interval. In this paper, we show how such negative examples can be used to make the solution to a regression problem more accurate.

1 Formulation of the Problem

What we want: a general description. From the practical viewpoint, in a rough approximation, the main objective of science is to enable people to predict what will happen in the world, and the main objective of engineering is to find out what changes we need to make in the world to make it better. To select the appropriate changes, we need to be able to predict how each possible change will affect the world.

Thus, in both cases, we need to be able, given the initial conditions x (which include the information about the change), to predict the value of each quantity y characterizing the future state.

Often, we do not know the dependence of y on x. In some cases – e.g., in celestial mechanics – we know the equations (or even explicit formulas) that relate the available information x and the desired quantity y. In such cases, in principle, we have an algorithm for predicting y.

In some situations, this algorithm may not be practical. For example, the fastest we can reasonably reliably predict where the tornado will go in the next 15 minutes is after several hours of computations on a high-performance computer – which makes these computations useless. However, as computers get faster and faster, we will eventually be able to make the corresponding computations practical.

In many other situations, however, we do not know how y depends on x. In such situations, we need to determine this dependence based on the known examples $\left(x^{(k)}, y^{(k)}\right)$ of past situations, in which we know both x and y.

Comment. Of course, this knowledge comes from measurements, and measurements are never absolutely accurate. So, in reality, instead of knowing the exact value y, we usually know an interval containing y (see, e.g., [3, 7, 9, 14]), and sometimes a probability distribution on this interval describing the relative frequency of different measurement errors [14].

Classification vs. regression. In some cases, the desired variable y takes only finite many values – e.g., sick or healthy; poor, medium, or rich, etc. Such problems are known as *classification problems.*

In other cases, the variable y can take all possible values within a certain interval. Such problems are known as *regression problems.*

Positive and negative examples. In addition to cases when we know both x and y – which we will call *positive examples*, there are also some cases in which we know x, but we only have partial information about y – e.g., we know that y *does not belong* to a certain interval. We will call such examples *negative examples.*

In classification problems – especially in binary classification problems, when we have only two possible values y_1 and y_2 of the quantity y – negative example are ubiquitous: indeed, every positive example in which we know that $y = y_2$ can be interpreted as a negative example in which we know that y is *not* equal to y_1.

However, in regression problems, negative examples are usually not used. In principle, they provide an additional information about the dependence, so it would be beneficial to use them – however, they are not used because it is not clear how to use them.

What we do in this paper. In this paper, we show how to use negative examples, and we show cases when the use of negative examples help.

In our analysis, we will cover all three major types of uncertainty: interval, fuzzy, and probabilistic. In our analysis, we will assume, for simplicity, that the x values are known exactly (i.e., to be more precise, that the inaccuracy in x can be safely ignored), but that the values of y are known with uncertainty. In all three cases, we assume that we know the family of dependencies $y = f(x, c_1, \ldots, c_n)$

– e.g., the family of all linear functions or the family of all quadratic functions – and we want to find the values $c = (c_1, \ldots, c_n)$ of the parameters for which the corresponding dependence is the best fit with the available data.

Important comment: negative examples in education. Another application area where negative examples are useful is education. A significant part of knowledge is taught by presenting examples $\left(x^{(k)}, y^{(k)}\right)$ of a problem x and of its correct solution y. It is well know, however, that learning can be enhanced if, in addition to correct solutions, student also see example of typical mistakes – e.g., pairs $\left(x^{(k)}, y^{(k)}\right)$ in which we know that $y^{(k)}$ is *not* a correct solution.

2 Case of Interval Uncertainty

Regression under interval uncertainty: a brief reminder. Following the general simplifying assumption, let us first consider the case when the values $x^{(k)}$ are known exactly, but the values $y^{(k)}$ are known with interval uncertainty – i.e., that for each k, we know the interval $\left[\underline{y}^{(k)}, \overline{y}^{(k)}\right]$ that contains the actual (unknown) value $y^{(k)}$.

Based on these measurement results, we select the values $c = (c_1, \ldots, c_n)$ for which the following condition is satisfied for all k:

$$\underline{y}^{(k)} \le f\left(x^{(k)}, c_1, \ldots, c_n\right) \le \overline{y}^{(k)}, 1 \le k \le K. \tag{1}$$

Regression under interval uncertainty: algorithms. For each i, we want to find the range $[\underline{c}_i, \overline{c}_i]$ of possible values of c_i. This range can be obtained by solving the following two constraint optimization problems:

- to find $\underline{c}_i$, we minimize c_i under the linear constraints (1); and
- to find $\overline{c}_i$, we maximize c_i under the linear constraints (1).

In the general non-linear case, this problem is NP-hard (even finding one single combination c that satisfies all the constraints (1) is, in general, NP-hard); see, e.g., [5]. In such cases, constraint solving algorithms (see, e.g., [3]) can lead to an approximate ranges: e.g., to enclosures $[\underline{c}_i', \overline{c}_i'] \supseteq [\underline{c}_i, \overline{c}_i]$ for the actual range.

The problem of computing the ranges $[\underline{c}_i, \overline{c}_i]$ becomes feasible if we consider families that linearly depend on the parameters c_i, i.e., families of the type

$$f(x, c_1, \ldots, c_n) = f_0(x) + c_1 \cdot f_1(x) + \ldots + c_n \cdot f_n(x). \tag{2}$$

In this case, inequalities (1) become linear inequalities in terms of the unknowns c_i:

$$\underline{y}^{(k)} \le f_0(x) + c_1 \cdot f_1\left(x^{(k)}\right) + \ldots + c_n \cdot f_n\left(x^{(k)}\right) \le \overline{y}^{(k)}, 1 \le k \le K \tag{3}$$

In this case, e.g., the range $[\underline{c}_i, \overline{c}_i]$ of possible values of c_i can be obtained by solving the following two linear programming problems – i.e., problems of optimizing a linear function under linear constraints:

- to find $\underline{c}_i$, we minimize c_i under the linear constraints (3); and
- to find $\overline{c}_i$, we maximize c_i under the linear constraints (3).

There exist feasible algorithms for solving linear programming problems; see, e.g., [2, 6]. Thus, the corresponding regression problem can indeed be feasibly solved.

What if we have "negative" intervals? What if, in addition to "positive" intervals – i.e., intervals that contain the y-values $y^{(k)}$, $k = 1, \ldots, K$ – we also have "negative" intervals $\left(\underline{y}^{(k)}, \overline{y}^{(k)}\right)$, $k = K+1, \ldots, L$ – i.e., intervals that are known *not* to contain the corresponding values $y^{(k)}$. In this case, in addition to the condition (1) satisfied for all k from 1 to K, we also have an additional condition that must be satisfied for each ℓ from $K+1$ to L:

$$f\left(x^{(\ell)}, c_1, \ldots, c_n\right) \le \underline{y}^{(\ell)} \text{ or } \overline{y}^{(\ell)} \le f\left(x^{(\ell)}, c_1, \ldots, c_n\right). \tag{4}$$

In this case, the question is to find the values $c = (c_1, \ldots, c_n)$ that satisfy all the constraints (1) and (4).

Negative intervals can help. Suppose that for a linear model $y = c_1 \cdot x$, we have two observations: for $x = -1$ and for $x = 1$, we have $y \in [-1, 1]$. One can easily see that in this case, the set of possible values of c_1 is the interval $[-1, 1]$.

In particular, for $x = 2$, the only information that we can extract from this data is that $y \in [-2, 2]$.

Now, if we know that for $x = 2$, the value y cannot be in the interval $(-3, 2)$, then the set of possible values of y narrow down to a single value $y = 2$, and the set $[-1, 1]$ of possible values of c_1 narrows down to a single value $c_1 = 1$.

With negative intervals, the problem becomes NP-hard already in the linear case. Indeed, it is known that the following problem is NP-hard (see, e.g., [5, 13]): given natural numbers $s_1, \ldots, s_n$ and s, find a subset of the values s_i that adds up to s. In other words, we need to find the values $c_i \in \{0, 1\}$ (describing whether to take the i-th value s_i or not) for which $\sum\limits_{i=1}^{n} c_i \cdot s_i = s$.

This problem can be easily reformulated as an interval problem with positive and negative examples. For this purpose, we take a linear model

$$y = c_1 \cdot x_1 + \ldots + c_n \cdot x_n$$

and the following examples:

- a positive example in which $x_i = s_i$ for all i and $y \in [s, s]$; consistency with this positive example means that $s = \sum\limits_{i=1}^{n} c_i \cdot s_i$;
- n additional positive examples; in the i-th example, $x_i = 1$, $x_j = 0$ for all $j \ne i$, and $y \in [0, 1]$; consistency with each such example means that $c_i \in [0, 1]$; and

- n negative examples; in the i-th example, $x_i = 1$, $x_j = 0$ for all $j \neq i$, and $y \notin (0,1)$; consistency with each such example means that $c_i \notin (0,1)$; together with the previous consistency, this means exactly that $c_i \in \{0,1\}$.

So what do we do: first idea. NP-hard means that, unless P = NP (which most computer scientists believe to be impossible), no feasible algorithm is possible that would always compute the exact ranges for c_i – or even check whether the data is consistent with the model. So what do we do?

Each negative interval $\left(\underline{y}^{(\ell)}, \overline{y}^{(\ell)}\right)$ means that the actual value of $y^{(\ell)}$ is either in the interval $\left(-\infty, \underline{y}^{(\ell)}\right]$ or in the interval $\left[\overline{y}^{(\ell)}, \infty\right)$. Thus:

- we can add, to K positive intervals, the first of these two semi-infinite intervals, solve the corresponding linear programming problem, and get ranges $\left[\underline{c}_i^{(\ell),-}, \overline{c}_i^{(\ell),-}\right]$ for the coefficients c_i;
- we can also add, to K positive intervals, the second of these two semi-infinite intervals, solve the corresponding linear programming problem, and get ranges $\left[\underline{c}_i^{(\ell),+}, \overline{c}_i^{(\ell),+}\right]$ for the coefficients c_i.

Since the actual value $y^{(\ell)}$ is either in the first *or* in the second of the semi-infinite intervals, the actual range of possible values of each c_i belongs to the *union* of the two intervals:

$$\left[\underline{c}_i^{(\ell)}, \overline{c}_i^{(\ell)}\right] = \left[\underline{c}_i^{(\ell),-}, \underline{c}_i^{(\ell),-}\right] \bigcup \left[\underline{c}_i^{(\ell),+}, \overline{c}_i^{(\ell),+}\right], \tag{4}$$

i.e., we take

$$\underline{c}_i^{(\ell)} = \min\left(\underline{c}_i^{(\ell),-}, \underline{c}_i^{(\ell),+}\right) \text{ and } \overline{c}_i^{(\ell)} = \max\left(\overline{c}_i^{(\ell),-}, \overline{c}_i^{(\ell),+}\right). \tag{5}$$

The actual value c_i belongs to all these intervals, so we can conclude that it belongs to the intersection $[\underline{c}_i, \overline{c}_i]$ of all these intervals:

$$[\underline{c}_i, \overline{c}_i] = \bigcap_{\ell=K+1}^{L} \left[\underline{c}_i^{(\ell)}, \overline{c}_i^{(\ell)}\right], \tag{6}$$

i.e., we take

$$\underline{c}_i = \max_{\ell} \underline{c}_i^{(\ell)} \text{ and } \overline{c}_i = \min_{\ell} \overline{c}_i^{(\ell)}. \tag{6}$$

If this intersection is empty, this means that the model is inconsistent with observations.

Second idea. In the above idea, every time, we only take into account one negative example. Instead, we can take into account two negative examples. In this, case, for each pair (ℓ, ℓ') of negative examples, we have four possible cases:

- we can have the case $a = --$ when $y^{\ell} \in \left(-\infty, \underline{y}^{(\ell)}\right]$ and $y^{\ell'} \in \left(-\infty, \underline{y}^{(\ell')}\right]$;

- we can have the case $a = -+$ when $y^\ell \in \left(-\infty, \underline{y}^{(\ell)}\right]$ and $y^{\ell'} \in \left[\overline{y}^{(\ell')}, \infty\right)$;
- we can have the case $a = +-$ when $y^\ell \in \left[\overline{y}^{(\ell)}, \infty\right)$ and $y^{\ell'} \in \left(-\infty, \underline{y}^{(\ell')}\right]$; and
- we can have the case $a = ++$ when $y^\ell \in \left[\overline{y}^{(\ell)}, \infty\right)$ and $y^{\ell'} \in \left[\overline{y}^{(\ell')}, \infty\right)$.

For each of these four cases $a = --, -+, +-, ++$, we can add the corresponding two semi-infinite intervals to K positive intervals, and find the ranges $\left[\underline{c}_i^{(\ell,\ell'),a}, \overline{c}_i^{(\ell,\ell'),a}\right]$ for the coefficients c_i. Then, we can conclude that the actual value of c_i belongs to the union of these four intervals:

$$\left[\underline{c}_i^{(\ell,\ell')}, \overline{c}_i^{(\ell,\ell')}\right] = \bigcup_a \left[\underline{c}_i^{(\ell,\ell'),a}, \underline{c}_i^{(\ell,\ell'),a}\right], \tag{7}$$

i.e., we take

$$\underline{c}_i^{(\ell,\ell')} = \min_a \underline{c}_i^{(\ell,\ell'),a} \text{ and } \overline{c}_i^{(\ell,\ell')} = \max_a \overline{c}_i^{(\ell,\ell'),a}. \tag{8}$$

The actual value c_i belongs to all these intervals, so we can conclude that it belongs to the intersection $[\underline{c}_i, \overline{c}_i]$ of all these intervals:

$$[\underline{c}_i, \overline{c}_i] = \bigcap_{K+1 \le \ell, \ell' \le L} \left[\underline{c}_i^{(\ell,\ell')}, \overline{c}_i^{(\ell,\ell')}\right], \tag{9}$$

i.e., we take

$$\underline{c}_i = \max_{\ell,\ell'} \underline{c}_i^{(\ell,\ell')} \text{ and } \overline{c}_i = \min_{\ell,\ell'} \overline{c}_i^{(\ell,\ell')}. \tag{10}$$

In this method, we get, in general, a better range – with smaller excess width – but now, instead of considering $O(L-K)$ cases, we need to consider $O((L-K)^2)$ cases.

We can get even more accurate estimates for the range if we consider all possible triples, 4-tuples, etc., of negative intervals, but then we will need to consider $O((L-K)^3)$, $O((L-K)^4)$, etc. cases.

3 Case of Fuzzy Uncertainty

What is fuzzy uncertainty: a brief reminder. In some cases, the values y are not measured but evaluated by an expert. An expert can say something like "the value of y is close to 1.5". To formalize such imprecise ("fuzzy") knowledge, Lotfi Zadeh invented special techniques – that he called fuzzy; see, e.g., [1, 4, 8, 10, 11, 12, 15].

In these techniques, for each imprecise expert statement about a quantity, we ask an expert to estimate, on a scale from 0 to 1, his/her degree of confidence that the expert's statement holds for this value (e.g., that 1.7 is close to 1.5). The function that assigns this degree to each possible value is called a *membership function*.

The degrees of confidence a, b, ... in individual statements $A, B, \ldots$ enable us also to estimate degrees of confidence in composite statements such as $A \,\&\, B$, $A \vee B$, etc. The algorithms $f_{\&}(a,b)$ and $f_{\vee}(a,b)$ for such estimates are called "and"- and "or"-operations, or, for historical reasons, t-norms and t-conorms. For example, the most widely used "and"-operations are $\min(a,b)$ and $a \cdot b$.

Regression under fuzzy uncertainty: a brief reminder. In line with the general idea, let us assume that we know the values $x^{(k)}$ exactly, and that we know the corresponding y-valued $y^{(k)}$ with fuzzy uncertainty – i.e., that for each example k and for each possible value y of this quantity, we know our degree of confidence $\mu_k(y)$ that this value of y is possible.

In this case, the degree to which a model $y = f(x, c_1, \ldots, c_n)$ is consistent with the k-th observation is equal to $\mu_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right)$, and the degree to which a model is consistent with all K observations is equal to

$$f_{\&}\left(\mu_1\left(f\left(x^{(1)}, c_1, \ldots, c_n\right)\right), \ldots, \mu_K\left(f\left(x^{(K)}, c_1, \ldots, c_n\right)\right)\right). \qquad (11)$$

A natural idea is to select the values $c_1, \ldots, c_n$ for which this degree is the largest possible.

What if we have negative examples? Suppose now that, in addition to K positive examples, we also have $L-K$ negative examples, for which we know that the expert's estimate is wrong. In fuzzy logic, the degree to which a statement is wrong is usually estimated as 1 minus the degree to which this statement is true. So, for a negative example, the degree to which this example is consistent with the model is equal to

$$1 - \mu_\ell\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right). \qquad (12)$$

Thus, in this case, we should select a model for which the following degree takes the largest possible value:

$$f_{\&}\left(\mu_1\left(f\left(x^{(1)}, c_1, \ldots, c_n\right)\right), \ldots, \mu_K\left(f\left(x^{(K)}, c_1, \ldots, c_n\right)\right),\right.$$
$$\left.1 - \mu_{K+1}\left(f\left(x^{(K+1)}, c_1, \ldots, c_n\right)\right), \ldots, 1 - \mu_L\left(f\left(x^{(L)}, c_1, \ldots, c_n\right)\right)\right). \qquad (13)$$

4 Case of Probabilistic Uncertainty

Regression under probabilistic uncertainty: a brief reminder. Probabilistic uncertainty means that for each measurement k, we know the probabilities of different possible values y, i.e., we know, e.g., the probability density function $\rho_k(y)$ describing these probabilities.

In this case, the probability that a model $y = f(x, c_1, \ldots, c_n)$ is consistent with the k-th observation is proportional to $\rho_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right)$. It is usually assumed that different measurements are independent. Thus, the probability

that a model is consistent with all K observations is equal to the product of the corresponding probabilities

$$\prod_{k=1}^{K} \rho_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right). \tag{14}$$

A natural idea is to select the values $c_1, \ldots, c_n$ for which this probability is the largest possible. This is known as the Maximum Likelihood method.

What if we have negative examples? From the purely probabilistic viewpoint, it is not clear how to handle such situations. However, since we have a solution for the fuzzy case, we can use the fact – emphasized many times by Zadeh – that the main difference between a membership function $\mu(x)$ and a probability density function $\rho(x)$ is in normalization:

- a membership function is usually selected so that $\max\limits_{x} \mu(x) = 1$, while
- the probability density function is selected so that the overall probability is 1, i.e., that $\int \rho(x)\, dx = 1$.

Thus, if we have a membership function, then, by multiplying it by an appropriate constant, we can get a probability density function, and, vice versa, if we have a probability density function $\rho(x)$, then, by dividing it by $m = \max\limits_{y} \rho(y)$, we will get a membership function.

So, a natural idea is to convert the original probabilistic knowledge $\rho_k(x)$ into fuzzy one, with $\mu_k(x) = c_k^{-1} \cdot \rho_k(x)$, where $c_k \stackrel{\text{def}}{=} \max\limits_{y} \rho_k(y)$. In this case, the fuzzy approach to regression will lead us to maximize the expression (11). We want the probability-to-fuzzy translation to be consistent with the Maximum Likelihood approach. Thus, we need to select $f_{\&}(a, b) = a \cdot b$. In this case, the expression (11) takes the form

$$\prod_{k=1}^{K} \mu_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right) =$$

$$\left(\prod_{k=1}^{k} c_k^{-1}\right) \cdot \left(\prod_{k=1}^{K} \rho_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right)\right). \tag{15}$$

This expression differs from (14) only by a multiplicative constant, so maximizing this expression is indeed equivalent to maximizing the expression (14) – i.e., to the Maximum Likelihood approach.

Now it is easy to take into account negative examples: we just maximize the product

$$\prod_{k=1}^{K} \mu_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right) \cdot \prod_{\ell=K+1}^{L} \left(1 - \mu_\ell\left(f\left(x^{(\ell)}, c_1, \ldots, c_n\right)\right)\right), \tag{16}$$

where

$$\mu_k(x) \stackrel{\text{def}}{=} \frac{\rho_k(x)}{\max\limits_y \rho_k(y)}. \tag{17}$$

Similarly to the derivation of the formula (15), we can see that maximizing the expression (16) is equivalent to minimizing a simpler expression

$$\prod_{k=1}^{K} \rho_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right) \cdot \prod_{\ell=K+1}^{L} \left(1 - \mu_\ell\left(f\left(x^{(\ell)}, c_1, \ldots, c_n\right)\right)\right). \tag{18}$$

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2009.

[3] L. Jaulin, M. Kiefer, O. Dicrit, and E. Walter, *Applied Interval Analysis*, Springer, London, 2001.

[4] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[5] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

[6] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, Springer, Cham, Switzerland, 2016.

[7] G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.

[8] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.

[9] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[10] H. T. Nguyen and V. Kreinovich, "Nested intervals and sets: concepts, relations to fuzzy sets, and applications", In: R. B. Kearfott and V. Kreinovich (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290.

[11] H. T. Nguyen, C. Walker, and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2019.

[12] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston, Dordrecht, 1999.

[13] C. Papadimitriou, *Computational Complexity*, Addison-Wesley, Reading, Massachusetts, 1994.

[14] S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.

[15] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.