Let Us Use Negative Examples in Regression-Type Problems Too

Jonatan Contreras, Francisco Zapata, Olga Kosheleva, Vladik Kreinovich, and Martine Ceberio

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

1. What We Want: A General Description

- From the practical viewpoint, the main objective of science is to predict what will happen in the world.
- 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.

2. 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.

3. We Don't Know the Dependence (cont-d)

- However, computers get faster and faster.
- So, 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.
- We need to determine this dependence based on the known examples $\left(x^{(k)}, y^{(k)}\right)$ of past situations.
- 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, and sometimes
  – a probability distribution on this interval describing the frequency of different y's.

4. Classification vs. Regression

- In some cases, the desired variable y takes only finite many values – e.g., sick or healthy; poor or rich.
- 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.*

5. Positive and Negative Examples

- There 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.
- For example, we know that y *does not belong* to a certain interval.
- We will call such examples *negative examples.*
- Negative example are ubiquitous in binary classification, when we have only two possible values y_1, y_2.
- Indeed:
  - every positive example in which $y = y_2$
  - can be interpreted as a negative example in which we know that y is *not* equal to y_1.

6. Positive and Negative Examples (cont-d)

- 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.
- In this talk, we show how to use negative examples.
- We also 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.

7. Positive and Negative Examples (cont-d)

- 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
  - the values of y are known with uncertainty.
- In all three cases, we assume that we know the family of dependencies $y = f(x, c_1, \ldots, c_n)$.
- For example, it can be the family of all linear functions or the family of all quadratic functions.
- 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.

8. Important Comment: Negative Examples in Education

- A significant part of knowledge is taught by presenting examples $(x^{(k)}, y^{(k)})$:
  - of a problem x and
  - of its correct solution y.
- It is well known that learning can be enhanced if:
  - in addition to correct solutions,
  - students also see example of typical mistakes,
  - i.e., pairs $(x^{(k)}, y^{(k)})$ in which we know that $y^{(k)}$ is *not* a correct solution.

9. Regression under Interval Uncertainty: A Brief Reminder

- Following the general simplifying assumption, we 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 $y^{(k)}$.
- We select the values $c = (c_1, \ldots, c_n)$ for which the following condition is satisfied for all k:

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

10. 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 above constraints; and
  - to find $\overline{c}_i$, we maximize c_i under the above constraints.
- In the general non-linear case, this problem is NP-hard.
- Even finding one single combination c that satisfies all the constraints is, in general, NP-hard.
- In such cases, constraint solving algorithms can lead to approximate ranges: e.g., to enclosures $[\underline{c}'_i, \overline{c}'_i] \supseteq [\underline{c}_i, \overline{c}_i]$.

11. Interval Regression (cont-d)

- Computing the ranges $[\underline{c}_i, \overline{c}_i]$ becomes feasible if we consider families that linearly depend on c_i:

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

- In this case, inequalities become linear inequalities in terms of the unknowns c_i:

$$\underline{y}^{(k)} \le f_0(x) + c_1 \cdot f_1\left(x^{(k)}\right) + \ldots + c_n \cdot f_n\left(x^{(k)}\right) \le \overline{y}^{(k)}.$$

- We can then solve the following two linear programming problems:
  - to find $\underline{c}_i$, we minimize c_i under the linear constraints; and
  - to find $\overline{c}_i$, we maximize c_i under the linear constraints.
- There exist feasible algorithms for linear programming, so these problems are feasible.

12. What If We Have "Negative" Intervals?

- What if we also have "negative" intervals $\left(\underline{y}^{(k)}, \overline{y}^{(k)}\right)$, $k = K+1, \ldots, L$ – that do *not* contain $y^{(k)}$.
- In this case, we also have an additional condition that must be satisfied for each ℓ from $K+1$ to L:

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

- The question is to find the values $c = (c_1, \ldots, c_n)$ that satisfy all the constraints.

13. 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, suppose that 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 $y = 2$.
- The set $[-1, 1]$ of possible values of c_1 narrows down to a single value $c_1 = 1$.

14. With Negative Intervals, Already the Linear Problem Is NP-Hard

- Indeed, it is known that the following problem is NP-hard:
  - 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 s_i or not) for which

$$\sum_{i=1}^{n} c_i \cdot s_i = s.$$

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

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

15. NP-Hard for Negative Intervals (cont-d)

- We take the following examples.
- A positive example: $x_i = s_i$ for all i and $y \in [s, s]$.
- Consistency with this example means $s = \sum_{i=1}^{n} c_i \cdot s_i$.
- n additional positive examples; in the i-th example:
  - we have $x_i = 1$, $x_j = 0$ for all $j \neq i$, and
  - we have $y \in [0, 1]$.
- Consistency with each such example means $c_i \in [0, 1]$.
- n negative examples; in the i-th example:
  - we have $x_i = 1$, $x_j = 0$ for all $j \neq i$, and
  - we have $y \notin (0, 1)$.
- Consistency with each such example means $c_i \notin (0, 1)$, so $c_i \in \{0, 1\}$.

16. So What Do We Do: First Idea

- NP-hard implies 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 $\left[\overline{y}^{(\ell)}, \infty\right)$.

17. First Idea (cont-d)

- Thus, we can:
  - add, to K positive intervals, the first of these two semi-infinite intervals, and
  - solve the corresponding linear programming problem, and get ranges $\left[\underline{c}_i^{(\ell),-}, \overline{c}_i^{(\ell),-}\right]$ for c_i;
  - we can also add, to K positive intervals, the second of these two semi-infinite intervals, and
  - solve the corresponding linear programming problem, and get ranges $\left[\underline{c}_i^{(\ell),+}, \overline{c}_i^{(\ell),+}\right]$ for c_i.
- The actual value $y^{(\ell)}$ is either in the first *or* in the second of the semi-infinite intervals.
- So, the actual range of possible values of each c_i belongs to the *union* of the two intervals:

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

What We Want: A . . .
Often, We Do Not . . .
Classification . . .
Positive and Negative . . .
Interval Uncertainty
Negative Intervals Can . . .
Fuzzy Uncertainty
Probabilistic Uncertainty
Future Work
Home Page
Title Page
Page 18 of 35
Go Back
Full Screen
Close
Quit

18. First Idea (cont-d)

- So, we take:

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

- 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:

$$[\underline{c}_i, \overline{c}_i] = \bigcap_{\ell=K+1}^{L} \left[\underline{c}_i^{(\ell)}, \overline{c}_i^{(\ell)}\right], \text{ i.e., we take}$$

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

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

19. 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.
- Then, 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 $y^{\ell'} \in \left[\overline{y}^{(\ell')}, \infty\right)$.

20. Second Idea (cont-d)

- For each of these four cases $a = --, -+, +-, ++$, we:
  - 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 c_i.
- Then, we can conclude that the actual value of c_i belongs to the union of these four intervals:

$$\left[\underline{c}_i^{(\ell,\ell')}, \overline{c}_i^{(\ell,\ell')}\right] = \bigcup_a \left[\underline{c}_i^{(\ell,\ell'),a}, \underline{c}_i^{(\ell,\ell'),a}\right], \text{ i.e., we take}$$

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

- 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:

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

21. Second Idea (cont-d)

- So, we take
$$\underline{c}_i = \max_{\ell,\ell'} \underline{c}_i^{(\ell,\ell')} \text{ and } \overline{c}_i = \min_{\ell,\ell'} \overline{c}_i^{(\ell,\ell')}.$$
- In this method, we get, in general, a better range – with smaller excess width.
- However, now, instead of considering $O(L-K)$ cases, we need to consider $O\left((L-K)^2\right)$ cases.
- We can get even more accurate estimates for the range if we consider:
  - all possible triples of negative intervals,
  - all possible 4-tuples of negative intervals, etc.
- However, then we will need to consider $O\left((L-K)^3\right)$, $O\left((L-K)^4\right)$, etc. cases.

22. 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.
- 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: 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*.

23. Fuzzy Uncertainty (cont-d)

- Here:
  - once we know the degrees of confidence a, b, ... in individual statements $A, B, \ldots$,
  - we can 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 $a \cdot b$.

24. Regression Under Fuzzy Uncertainty: A Brief Reminder

- As usual, we know the $x^{(k)}$ exactly, and we know $y^{(k)}$ with fuzzy uncertainty.
- So, for each value y, we know our degree of confidence $\mu_k(y)$ that 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

$$\mu_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right).$$

- The degree to which a model is consistent with all K observations is equal to

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

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

25. 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
  - one 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

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

26. What If We Have Negative Examples (cont-d)

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

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

$$\left.1-\mu_{K+1}\left(f\left(x^{(K+1)},c\right)\right),\ldots,1-\mu_L\left(f\left(x^{(L)},c\right)\right)\right).$$

27. Regression under Probabilistic Uncertainty: A Brief Reminder

- Probabilistic uncertainty means that for each measurement k, we know the probabilities of different y's.
- In other words, we know, e.g., the probability density function $\rho_k(y)$ describing these probabilities.
- So, the probability that a model $y = f(x, c_1, \ldots, c_n)$ is consistent with the k-th observation is proportional to:

$$\rho_k\left(f\left(x^{(k)}, c_1, \ldots, c_n\right)\right).$$

28. Probabilistic Uncertainty (cont-d)

- 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:

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

- 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.

29. What If We Have Negative Examples?

- From the purely probabilistic viewpoint, it is not clear how to handle such situations.
- However, we have a solution for the fuzzy case.
- So, we can use the fact – emphasized many times by Zadeh – that:
  - the main difference between a membership function $\mu(y)$ and a probability density function $\rho(y)$
  - is in normalization.
- A membership function has $\max\limits_y \mu(y) = 1$.
- The probability density function is selected so that the overall probability is 1, i.e., that $\int \rho(y)\,dy = 1$.

30. What If We Have Negative Examples (cont-d)

- If we have a membership function, then:
  - by multiplying it by an appropriate constant,
  - we can get a probability density function.
- If we have a probability density function $\rho(y)$, 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 $\rho_k(y)$ into fuzzy one:

$$\mu_k(y) = c_k^{-1} \cdot \rho_k(y), \text{ where } c_k \stackrel{\text{def}}{=} \max_{y'} \rho_k(y').$$

- In this case, the fuzzy approach to regression will lead us to maximize the above expression.
- We want the probability-to-fuzzy translation to be consistent with the Maximum Likelihood approach.

31. What If We Have Negative Examples (cont-d)

- Thus, we need to select $f_{\&}(a,b) = a \cdot b$.
- In this case, the above expression takes the form

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

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

- This expression differs from likelihood only by a multiplicative constant.
- So, maximizing this expression is indeed equivalent to the Maximum Likelihood approach.

32. What If We Have Negative Examples (cont-d)

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

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

$$\text{where } \mu_k(y) \stackrel{\text{def}}{=} \frac{\rho_k(y)}{\max\limits_{y'} \rho_k(y')}.$$

- It is easy to see that maximizing this expression is equivalent to minimizing a simpler expression

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

33. Future Work

- In this talk, we provided a theoretical foundation for using negative examples in regression-like problems.
- We also showed, on simplified examples, that the resulting algorithms lead to more accurate models.
- Now we plan to apply the resulting algorithms and ideas to real-life problems.
- We hope that others will join us in this effort.

34. 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),
- HRD-1242122 (Cyber-ShARE Center of Excellence).