4-2020

Is There a Contradiction Between Statistics and Fairness: From Intelligent Control to Explainable AI

Christian Servin
El Paso Community College, cservin1@epcc.edu

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-20-28

Is There a Contradiction Between Statistics and Fairness: From Intelligent Control to Explainable AI

Christian Servin[1] and Vladik Kreinovich[2]
[1]Computer Science and
Information Technology Systems Department
El Paso Community College (EPCC)
919 Hunter Dr.
El Paso, TX 79915-1908, USA
cservin1@epcc.edu
[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu

Abstract

At first glance, there seems to be a contradiction between statistics and fairness: statistics-based AI techniques lead to unfair discrimination based on gender, race, and socio-economical status. This is not just a fault of probability techniques: similar problems can happen if we use fuzzy or other techniques for processing uncertainty. To attain fairness, several authors proposed not to rely on statistics and instead, explicitly add fairness constraints into decision making. In this paper, we show that the seeming contradiction between statistics and fairness is caused mostly by the fact that the existing systems use simplified models; contradictions disappear if we replace them with more adequate (and thus more complex) statistical models.

1 Formulation of the Problem

Social applications of AI. Recent AI techniques like deep learning have led to many successful applications. For example, we can apply deep learning to decide whose loan applications should be approved and whose applications should be rejected – and if approved, what interest should we charge. We can apply deep learning to decide which candidates for graduate program to accept – and for

those accepted what financial benefits to offer as an enticement.

In all such cases, we feed the system with numerous past examples of successes and failures. Based on these example, the systems tries its best to predict whether a given loan or a given potential student will be a success or not. Statistically, these systems seem to work well: they predict success of failure better than human decision makers. However, the results are often not satisfactory; see, e.g., [1, 9]. Let us explain why.

Many current social applications of AI are unsatisfactory. On average, loan applications from poorer geographic areas have a higher default rate. This is a known fact, and statistical methods underlying machine learning find this out. As a result, the system naturally recommends rejection of all loans from these areas. This is not fair to people with good credit record who happen to live in the not-so-good areas. Moveover, it is also detrimental to the bank since the bank will miss on profiting from such potentially successful loans.

Similarly, it is known that in many disciplines women has a lower success rates in getting their PhDs than man – and take longer when they succeed. One of the main reasons for this is that raising children requires much more efforts from women than from men. A statistical system, crudely speaking, does not care about the reasons, it just takes this statistical fact into account and preferably selects males. Not only this is not fair, this way the universities miss a lot of talent – and nowadays, with not much need for routine boring work, talent and creativity are extremely important, they should be nurtured, not rejected.

So is there a contradiction between statistics and fairness? At first glance, it may seem that there is a contradiction between statistical methods and our ideas of fairness. In other words, it seems that if we want the systems to be fair, we cannot rely on statistics only, we need to supplement statistics with additional fairness constraints.

The need for such constraints is usually formulated – in our opinion, not fully accurately – as the need for *explainable AI*; see, e.g., [6] and references therein. The main idea behind explainable AI is that instead of relying on a machine learning system as a black box, we extract some rules from this system – and if these rules are not fair, we replace them with fairer rules.

What we show in this paper. In this paper, we show that the seeming inconsistency comes from the fact that we use simplified statistical models. We show that a more detailed description of the corresponding uncertainty – be it interval, probabilistic, or fuzzy uncertainty – eliminates this seeming contradiction, and enables the system to come up with fair decisions without any need for additional constraints.

2 How Current Techniques Lead to Unfair Decisions: Simplified Examples

Let us give some examples. In other to explain why the existing techniques can lead to unfair solutions, let us give some detailed simplified examples. We will start with statistical examples. Then, we will show that mathematically similar examples – this time not related to fairness – can be found in applications of fuzzy techniques as well – namely, when we apply the usual intelligent control techniques.

A simplified statistical example. Let us consider a statistical version of a classical AI example:

- birds normally fly,
- penguins are birds,
- penguins normally do not fly, and
- Sam is a penguin.

The question is: does Sam fly? To make it into a statistical example, let us add some probabilities. Let us assume that 90% of the birds fly, and that 99% of the penguins do not fly (of course, in reality, 100% of the penguins do not fly, but let us keep it under 100% since in most real-life situations, we are never 100% sure about anything).

From the viewpoint of common sense, the information about birds flying in general is rather irrelevant for our situation – since we know that Sam is not just any bird, it is a penguin, a very specific type of bird for which we know the probability of flying. So, to find the probability of Sam flying, we should only take into account information about penguins and thus, conclude that the probability of Sam flying is $100 - 99 = 1\%$.

However, this is not what we would get if we use the standard statistical techniques. Indeed, from the purely statistical viewpoint, here we have two rules that lead us to two different conclusion:

- since Sam is a bird, we can make a conclusion A that Sam flies, with probability $a = 90\%$; and
- since Sam is a penguin, we can make a conclusion B that Sam does not fly, with probability $b = 99\%$.

These two conclusions cannot be both right, since the probabilities of Sam flying and not flying should add up to 1, and here we have $0.9 + 0.99 = 1.89 > 1$. This means that these conclusions are inconsistent.

From the purely logical viewpoint, if we have two statements A and B each of which may be true or false, we can have four possible situations:

- both A and B are true, i.e., $A \,\&\, B$;

- A is true but B is false, i.e., $A \,\&\, \neg B$;
- A is false but B is true, i.e., $\neg A \,\&\, B$; and
- both A and B are false, i.e., $\neg A \,\&\, \neg B$.

In general, the probabilities $P(.)$ of all four situations can be obtained by using the Maximum Entropy Principle – a natural extension of the Laplace Indeterminacy Principle – according to which, if we do not know the dependence between two random variables, then we should assume that they are independent; see, e.g., [3]. For independent events, probabilities multiply, so $P(A \,\&\, B) = P(A) \cdot P(B) = a \cdot b$, $P(A \,\&\, \neg B) = a \cdot (1-b)$, $P(\neg A \,\&\, B) = (1-a) \cdot b$, and $P(\neg A \,\&\, \neg B) = (1-a) \cdot (1-b)$.

In our case, the statements A and B are inconsistent, so we cannot have $A \,\&\, B$ and we cannot have $\neg A \,\&\, \neg B$. The only two consistent options are $A \,\&\, \neg B$ and $\neg A \,\&\, B$. Thus, the true probabilities $P(A)$ and $P(B)$ of A and B can be found if we restrict ourselves to consistent situations:

$$P(A) = P(A \,|\, \text{consistent}) = \frac{P(A \,\&\, \text{consistent})}{P(\text{consistent})} =$$

$$\frac{P(A \,\&\, \neg B)}{P(A \,\&\, \neg B) + P(\neg A \,\&\, B)} = \frac{a \cdot (1-b)}{a \cdot (1-b) + (1-a) \cdot b}$$

and, of course, $P(B) = 1 - P(A)$.

In our example, with $a = 0.9$ and $b = 0.99$, we get

$$P(A) = \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.1 \cdot 0.99} = \frac{0.009}{0.009 + 0.099} = \frac{1}{12} \approx 8\%.$$

So, instead the desired 1%, we get a much larger 8% probability – clearly affected by the general rule that birds normally fly.

This is a simplified example, but it explains why recommendation systems based on usual statistical rules becomes biased: if a person with a perfect credit history happens to live in a poor neighborhood, in which the overall loan success rate is small, this person's chances of getting a loan will be decreased. Similarly, if a female student with perfect credentials applies for a graduate program, the system would be treating her less favorably – since in general, in computer science, female students succeed with lower frequency.

In both cases, we have clearly unfair situations – the system designers may honestly give female students a better chance to succeed, but instead, their inference system perpetrates the inequality.

A simplified fuzzy example. A fuzzy-related reader may view the above example as one more example of why statistical methods are not always applicable, an why alternative methods – such as fuzzy methods – are needed. Alas, we will show that a very similar example is possible if we use the usual fuzzy techniques.

This problem may not be well known for fuzzy recommendation systems – since there are not too many of them actively used – but it is exactly the same

problem that is well known in fuzzy control, the traditional application area of fuzzy techniques; see, e.g., [2, 4, 5, 7, 8, 10].

Indeed, suppose that we have two rules that describe how the control u should depend on the value x of the measured quantity:

- if x is small, then u is small; and
- if $x = 0.2$, then $u = 0.3$.

Suppose also that the notion "small" is described by a triangular membership function $\mu_{\rm small}(x)$ which is equal to $\max(1 - |x|, 0)$.

From the common sense viewpoint, the first rule is more general, while the second rule – which is actually in full agreement with the first one – describes a specific knowledge that we have about control corresponding to the value $x = 0.2$. Such situations can happen, e.g., when we combine the general expert knowledge (the first rule) with the results of specific calculations (second rule).

In this case, if we actually have the value $x = 0.2$ for which we know the exact control value $u = 0.3$, we should return this control value.

One of the usual ways of dealing with a system of fuzzy rules "if $A_i(x)$ then $B_i(u)$" $(i = 1, \ldots, n)$ is to take into account that a control u is reasonable for given value x if:

- either the first rule is applicable, i.e., this rule's condition $A_1(x)$ is satisfied *and* thus, its conclusion $B_1(u)$ is also satisfied,
- or the second rule is applicable, i.e., this rule's condition $A_2(x)$ is satisfied and thus, its conclusion $B_2(u)$ is also satisfied, etc.

If we denote this property "u is reasonable for x" by $R(x, u)$, and use the usual notations $\&$ for "and" and $\vee$ for "or", then the above text will become the following formula:

$$R(x, u) \leftrightarrow (A_1(x) \,\&\, B_1(u)) \vee (A_2(x) \,\&\, B_2(u)) \vee \ldots$$

In line with the general fuzzy methodology, for situations in which we are not 100% sure about the properties A_i and B_j, we can apply the corresponding fuzzy versions $f_{\&}(a, b)$ and $f_{\vee}(a, b)$ of usual "and" and "or" – known as "and"- and "or"-operations (or, alternatively, t-norm and t-conorm) – to the degrees $\mu_{A_i}(x)$ and $\mu_{B_i}(u)$ to which these properties are satisfied. Then, for the degree $\mu_r(x, u)$ to which u is reasonable for x, we get the following formula:

$$\mu_r(x, u) = f_{\vee}(f_{\&}(\mu_{A_1}(x), \mu_{B_1}(u)), f_{\&}(\mu_{A_2}(x), \mu_{B_2}(u)), \ldots).$$

In particular, for the simplest possible "and"- and "or"-operations $f_{\&}(a, b) = \min(a, b)$ and $f_{\vee}(a, b) = \max(a, b)$, we get

$$\mu_r(x, u) = \max(\min(\mu_{A_1}(x), \mu_{B_1}(u)), \min(\mu_{A_2}(x), \mu_{B_2}(u)), \ldots).$$

Once we have this degree for each u, we can find the control $\overline{u}$ corresponding to x by requiring that its mean square deviation from the actual value u – weighted

by this degree – is the smallest possible. In precise terms, for a given x, we minimize the expression $\int \mu_r(x,u) \cdot (u-\overline{u})^2$. Differentiating this expression with respect to $\overline{u}$ and equating the derivative to 0, we get the formula

$$\overline{u} = \frac{\int \mu_r(x,u) \cdot u \, du}{\int \mu_r(x,u) \, du}$$

known as *centroid defuzzification*.

Let us apply this technique to our two rules, for the case when $x = 0.2$ and thus, $\mu_{\text{small}}(x) = 0.8$. In the second rule, both the condition and the conclusion are crisp:

- we have $\mu_{A_2}(0.2) = 1$ and $\mu_{A_2}(x) = 0$ for all other values x, and
- we have $\mu_{B_2}(0.3) = 1$ and $\mu_{B_2}(u) = 0$ for all other values u.

Thus, for all $u \neq 0.2$, we have $\mu_r(x,u) = \min(\mu_{\text{small}}(u), 0.8)$ and for $u = 0.2$, we have $\mu_r(x,u) = 1$.

According to the centroid formula, the resulting control is the above ratio of two integrals. The single-point change in the function $\mu_r(x,u)$ does not affect its integral, so the numerator is simply equal to the integral of the product $\min(\mu_{\text{small}}(u), 0.8) \cdot u = \min(\max(1-|u|), 0), 0.8) \cdot u$. This product is an odd function of u: the first factor does not change if we replace u with $-u$, and the second changes sign. Thus, its integral is 0, and so, the usual fuzzy methodology leads to the control $u = 0$ – while from the viewpoint of common sense, we should get 0.3.

3 Using More Detailed Models Helps

General description of the problem. In all previous example, we considered the case of situations when we have two rules describing a given situation. For example, in the case of loans:

- the first rule is that loans recipients from poor areas often default on a loan, and
- the second rule is that people with a good credit record usually pay back their loans.

From the common sense viewpoint, for a person with a good credit record living in a poor area, we should go with the second rule, but the above-described naive statistical approach – implemented in current machine learning systems – pays an unnecessarily high attention to the first rule as well.

Similarly, for Sam the penguin:

- we have a general rule applicable to all the birds – that they usually fly; and

- we have a second specific rule, applicable only to penguins – that they do not fly.

From the common sense viewpoint, since Sam in a penguin, we should go with the second rule, but the naive statistical approach gives too much weight to the first rule.

Idea. From the statistical viewpoint – or, more generally, from the viewpoint of data processing – how can we distinguish between a more general rule and a more specific rule? One important difference between a more general case is that this case describes a larger sample, while a more specific case describes a sub-sample of this sample, a sub-sample in which all the objects are, in some reasonable sense, similar and thus, differ from each other less than in the general sample. As a result, for many quantities characterizing the objects, the standard deviation σ corresponding to the larger sample is much larger than for a smaller sub-sample.

This is simple and reasonable, and – as we show – it helps put more weight on a more general rule and thus, helps avoid the contradiction between statistics and fairness.

How to combine two statistical rules with different means and standard deviations: reminder. To illustrate our point, let us consider the simplest situation when we have two statistical rules – coming from two independent sets of arguments or observation – that predict the value of a quantity x, and we are absolutely confident in both of these rules. Since these are statistical rules, they do not predict the exact value of the quantity, they only predict the probabilities of different possible values of this quantity. These probabilities can be described by the corresponding probability density functions $\rho_1(x)$ and $\rho_2(x)$.

If these were rules predicting two different quantities x_1 and x_2, then, due to the fact that these rules are assumed to be independent, the probability to have values x_1 and x_2 should be equal to the product $\rho_1(x_1) \cdot \rho_2(x_2)$. However, in our case, we know that these distributions describe the exact same quantity, i.e., that we have the additional condition $x_1 = x_2$. Thus, instead of the above 2-D probability density, we need to consider the *conditional* probability density, under the condition that $x_1 = x_2$. It is known that, in general, for $A \subseteq B$, the conditional probability $P(A \,|\, B)$ can be obtained from the probability $P(A)$ by diving it by the probability $P(B)$ that the condition is satisfied – i.e., in effect, by dividing the probability $P(A)$ by a constant. Thus, in our case, the resulting probability density is equal to $\rho(x) = c \cdot \rho_1(x) \cdot \rho_2(x)$, where c is a constant that can be determined from the condition that $\int \rho(x)\, dx = 1$, so that

$$\rho(x) = \frac{\rho_1(x) \cdot \rho_2(x)}{\int \rho_1(y) \cdot \rho_2(y)\, dy}.$$

In particular, if both probability distributions $\rho_1(x)$ and $\rho_2(x)$ are Gaussian, i.e., have the form $\rho_i(x) = \text{const} \exp\left(-\dfrac{(x-a_i)^2}{2\sigma_i^2}\right)$ for some means a_i and

standard deviations σ_i, then, as one can easily check, the resulting distribution is also Gaussian, with mean a and standard deviation σ determined by the formulas $a = \dfrac{a_1 \cdot \sigma_1^{-2} + a_2 \cdot \sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}}$ and $\sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2}$.

How is this applicable to our examples. Let us consider the case of a loan. Here, we have two pieces of information about a loan applicant:

- the first piece of information is that this person has a good credit history;
- the second piece of information is that this person lives in a poor area.

To combine these two pieces of information, let us estimate the corresponding means and standard deviations.

Let us start with the estimates corresponding to people with good credit history. In most cases, people with good credit history return their loans – and return them on time. So, the mean value a_1 of the returned percentage of the loan x is close to 100, and the corresponding standard deviation is σ_1 is close to 0.

On the other hand, in general, for people living in a poor area, the returned percentages vary:

- some people living in the poor area struggle, but return their loans,
- some fail and become unable to return their loans.

Here, the average a_2 is clearly less that 100, and the standard deviation σ_2 is clearly much larger than σ_1:

$$\sigma_2 \gg \sigma_1.$$

If we multiply both the numerator and the denominator of the above formula for the combined value a by σ_1^2, we conclude that $a = \dfrac{a_1 + a_2 \cdot (\sigma_1^2/\sigma_2^2)}{1 + \sigma_1^2/\sigma_2^2}$. Since here $\sigma_1 \ll \sigma_2$, we get $a \approx a_1$. So, we conclude that the resulting estimate is fully determined by the fact that the applicant has a good credit history – and this estimate is practically *not* affected by the fact that the applicant happens to live in a poor area. This is exactly what we wanted the system to conclude.

Similar arguments help resolve the bird-fly puzzle. As a measure of a flying ability, we can take, e.g., the time that a bird can stay in the air.

- No penguin can really fly, so for penguins, this time is always small, and the standard deviation of this time is close to 0: $\sigma_1 \approx 0$.
- On the other hand, if we consider the population of all the birds, then on this general population, there is a large variance: some birds can barely fly for a few minutes, while others can fly for days and cross the oceans. For this piece of knowledge, the variance is huge and thus, the standard deviation σ_2 is also huge.

Here too, $\sigma_1 \ll \sigma_2$ and thus, our conclusion about Sam's ability to fly will be determined practically exclusively by the fact that Sam is a penguin – in full agreement with common sense.

4 How Is This Idea Applicable to Fuzzy

Usual relation between fuzzy and probability. As Lotfi Zadeh mentioned several times, from the mathematical viewpoint, the main difference between a probability density function $\rho(x)$ and a membership function $\mu(x)$ is in their normalization:

- for a probability density function, we have $\int \rho(x)\,dx$, while
- for a membership function, we have $\max\limits_x \mu(x) = 1$.

As a result:

- if we have a probability density function $\rho(x)$, then we can normalize it as membership function, by taking

$$\mu(x) = \frac{\rho(x)}{\max\limits_y \rho(y)};$$

- if we have a membership function $\mu(x)$, then we can normalize it as a probability density function, by taking

$$\rho(x) = \frac{\mu(x)}{\int \mu(y)\,dy}.$$

Let us use this relation to combine fuzzy knowledge. We know how to combine probabilistic knowledge. So, if we have two membership functions $\mu_1(x)$ and $\mu_2(x)$, we can combine the corresponding pieces of knowledge as follows:

- first, we use the above relation to transform the given membership functions into probability density functions $\rho_i(x) = c_i \cdot \mu_i(x)$, for some constants c_i;
- second, we use the procedure described in the previous section to combine the probability density functions $\rho_1(x)$ and $\rho_2(x)$ into a single probability density function $\rho(x) = \text{const} \cdot \rho_1(x) \cdot \rho_2(x)$ – which, due to the above relation between probability and fuzzy, takes the form $\rho(x) = c_3 \cdot \mu_1(x) \cdot \mu_2(x)$ for some constant c_3;
- finally, we transform the resulting probability function $\rho(x)$ back into a membership function, thus getting $\mu(x) = c_4 \cdot \rho(x)$ for some constant c_4, i.e., $\mu(x) = c \cdot \mu_1(x) \cdot \mu_2(x)$ for an appropriate constant c.

This idea allows us to avoid the problem of traditional defuzzification. Let us show that this combination rule enables us to avoid the above-described problem of traditional defuzzification. Indeed, if we have two rules:

- one rule corresponding to a very narrow membership function (i.e., in probabilistic terms, very small σ), and
- another rule with a very wide membership function (i.e., with large σ),

then, as we have mentioned in the previous section, in the combined function, the contribution of the wide rule will be largely ignored, and the conclusion will be practically identical with what the narrow rule recommends – exactly as we want.

What if we are only partly confident about some piece of knowledge? The above combination formula describes how to combine two rules about which we are fully confident. But what if we have some rules about which we are only partly confident?

One way to interpret degree of confidence in a statement is:

- to have a poll of N experts and,
- if M out of N experts confirm this statement, to take the corresponding proportion M/N as the desired degree of confidence.

Let us describe the membership function corresponding to the situation when only one expert confirms the statement by $\mu_1(x)$. In this case, according to the above combination formula, the case when M experts confirm the statement is described by a membership function proportional to $\mu_1^M(x)$. In particular, the case of full confidence, when all N experts confirm the statement, is described by the membership function $\mu(x)$ which is proportional to $\mu_1^N(x)$: $\mu(x) \sim \mu_1^N(x)$. Thus, $\mu_1(x) \sim (\mu(x))^{1/N}$ and therefore, the membership function $\sim \mu_1^M(x)$ corresponding to degree of confidence $d = M/N$ is proportional to $(\mu(x))^{M/N} = \mu^d(x)$.

In general, if we have a rule like $A(x) \to B(u)$ relating the property $A(x)$ of the input (with membership function $\mu_A(x)$) and the property $B(u)$ of the desired control u (with membership function $\mu_B(u)$), then for each input x, our degree of confidence in the conclusion $B(u)$ is equal to $d = \mu_A(x)$. Thus, the resulting membership function about u should be proportional to $(\mu_B(u))^{\mu_A(x)}$. In we have several such rules $A_1(x) \to B_1(u)$, $A_2(x) \to B_2(u)$, etc., then the resulting membership function should be proportional to the product of membership functions corresponding to individual rules, i.e., to the product

$$(\mu_{B_1}(u))^{\mu_{A_1}(x)} \cdot (\mu_{B_2}(u))^{\mu_{A_2}(x)} \cdot \ldots$$

Acknowledgments

This work was supported in part by the US 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. K. E. Bellamy et al., "Think your Artifical Intelligence software is fair? Thing again", *Computing Edge*, February 2020, pp. 14–18.

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

[3] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

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

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

[6] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI", *Proceedings of the 2019 ACM Fairness, Accounatbility, and Transparency Conference FAT'2019*, Atlanta, Georgia, January 29–31, 2019, pp. 279–288.

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

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

[9] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown Books, New York, 2016.

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