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]University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
vladik@utep.edu

1. 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.

2. Social Applications of AI (cont-d)

- Based on these example, the systems predict whether a given loan will be a success.
- Statistically, these systems work well: they predict success or failure better than human decision makers.
- However, the results are often not satisfactory. Let us explain why.

3. 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.
- Indeed, the bank will miss on profiting from such potentially successful loans.
- Similarly, in many disciplines women has a lower success rates in getting their PhDs than men.

4. Many Current Social Applications of AI Are Unsatisfactory (cont-d)

- Women also, on average, take longer to 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.
- This system 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.
- Talent and creativity should be nurtured, not rejected.

5. So Is There a Contradiction Between Statistics And Fairness?

- 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 as the need for *explainable AI*.
- 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.

6. What We Show in This Talk

- 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 – probabilistic or fuzzy,
  - eliminates this seeming contradiction, and
  - enables the system to come up with fair decisions without any need for additional constraints.

7. Examples of Unfair Decisions

- We want to understand why the existing techniques can lead to unfair solutions.
- So let us trace 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.

8. 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.

9. A Simplified Example (cont-d)

- Of course, in reality, 100% of the penguins do not fly.
- However, 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.
- Indeed, we know that Sam is not just any bird, it is a penguin.
- Penguins are 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.
- Thus, we should conclude that the probability of Sam flying is $100 - 99 = 1\%$.

10. A Simplified Example (cont-d)

- 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 conclusions:
  - 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.
- Indeed, 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.

11. A Simplified Example (cont-d)

- From the purely logical viewpoint, if we have two statements A and B, 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., $\neg A \,\&\, \neg B$.
- The probabilities $P(.)$ of all four situations can be obtained by using the Maximum Entropy Principle.
- This is a natural extension of the Laplace Indeterminacy Principle.
- According to Maximum Entropy Principle,
  - if we do not know the dependence between two random variables,
  - then we should assume that they are independent.

12. A Simplified Example (cont-d)

- For independent events, probabilities multiply, so

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

  $P(\neg A\,\&\,B) = (1-a)\cdot b$, $P(\neg A\,\&\,\neg B) = (1-a)\cdot(1-b)$.

- 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 $P(B)$ can be found if we restrict ourselves to consistent situations:

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

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

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

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

13. A Simplified Example (cont-d)

- So, instead the desired 1%, we get a much larger 8% probability.
- This value is clearly affected by the general rule that birds normally fly.
- This is a simplified example.
- However, 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,
  - this person's chances of getting a loan will be decreased.

14. A Simplified Example (cont-d)

- 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.

15. 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,
  - and 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 few of them.
- However, it is exactly the same problem that is well known in fuzzy control.
- And fuzzy control is a traditional application area of fuzzy techniques.

16. A Simplified Fuzzy Example (cont-d)

- Indeed, suppose that we have two rules that describe how the control u should depend on the input x:
  - 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

$$\mu_{\text{small}}(x) = \max(1 - |x|, 0).$$

- From the common sense viewpoint, the first rule is more general.
- The second rule describes a specific knowledge that we have about control corresponding to $x = 0.2$.
- The second rule is actually in full agreement with the first one.

17. A Simplified Fuzzy Example (cont-d)

- 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, for $x = 0.2$, we know the exact control value $u = 0.3$.
- So, we should return this control value.
- Suppose that we have fuzzy rules "if $A_i(x)$ then $B_i(u)$", $i = 1, \ldots, n$.
- This means that a control u is reasonable for given value x if:
  - either the first rule is applicable, i.e., $A_1(x)$ is true *and* $B_1(u)$ is true,
  - or the second rule is applicable, i.e., $A_2(x)$ is true and $B_2(u)$ is true, etc.

18. A Simplified Fuzzy Example (cont-d)

- Let us denote this property "u is reasonable for x" by $R(x,u)$.
- In usual notations & for "and" and $\vee$ for "or", 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".
- Then, for the degree $\mu_r(x,u)$ to which u is reasonable for x, we get the following formula:

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

19. A Simplified Fuzzy Example (cont-d)

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

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

- 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

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

20. A Simplified Fuzzy Example (cont-d)

- This formula is 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.

21. A Simplified Fuzzy Example (cont-d)

- 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 factor changes sign.
- Thus, its integral is 0.
- So, the usual fuzzy methodology leads to $u = 0$.
- However, from the viewpoint of common sense, we should get $u = 0.3$.

22. General Description of the Problem

- In all previous example, we considered the case of situations when we have two rules.
- 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.
- However, the naive statistical approach pays an unnecessarily high attention to the first rule as well.

23. General Description of the Problem (cont-d)

- And this approach underlies in current machine learning systems.
- 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.
- However, the naive statistical approach gives too much weight to the first rule.

24. How Can We Distinguish Between a More General And a More Specific Rule?

- One important difference is that a more specific case describes a sub-sample.
- In this sub-sample, all the objects are, in some reasonable sense, similar.
- Thus, they differ from each other less than in the general sample.
- So, for many quantities, the standard deviation σ is much larger in the larger sample.
- This is simple and reasonable, and – as we show:
  - it helps put more weight on a more general rule and,
  - thus, it helps avoid the contradiction between statistics and fairness.

25. How to Combine Statistical Rules With Different Means And Standard Deviations

- To illustrate our point, let us consider the simplest situation when we have two statistical rules.
- Let's assume that these rules come from two independent sets of arguments or observation.
- Both rules 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 $\rho_2(x)$.

26. How to Combine Statistical Rules (cont-d)

- 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 $x_1 = x_2$; so:
  - 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 for $A \subseteq B$, $P(A \,|\, B) = \dfrac{P(A)}{P(B)}$.
- So, $P(A \,|\, B) = c \cdot P(A)$ for some constant c.

27. How to Combine Statistical Rules (cont-d)

- 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.
- This constant can be determined from the condition $\int \rho(x)\,dx = 1$, so $\rho(x) = \dfrac{\rho_1(x) \cdot \rho_2(x)}{\int \rho_1(y) \cdot \rho_2(y)\,dy}$.
- Often, both probability distributions $\rho_1(x)$ and $\rho_2(x)$ are Gaussian: $\rho_i(x) = \text{const} \exp\left(-\dfrac{(x - a_i)^2}{2\sigma_i^2}\right)$.
- Here, a_i are means and σ_i are standard deviations.
- Then, as one can easily check, the resulting distribution is also Gaussian, with

$$a = \frac{a_1 \cdot \sigma_1^{-2} + a_2 \cdot \sigma_2^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}} \text{ and } \sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2}.$$

28. 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 st. dev.
- 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.

29. Application to Our Examples (cont-d)

- 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 formula for a by σ_1^2, we get:

$$a = \frac{a_1 + a_2 \cdot (\sigma_1^2/\sigma_2^2)}{1 + \sigma_1^2/\sigma_2^2}.$$

30. Application to Our Examples (cont-d)

- 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;
  - 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: $\sigma_1 \approx 0$.

31. Application to Our Examples (cont-d)

- On the other hand, if we consider the population of all the birds, then 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, $\sigma_1 \ll \sigma_2$.
- 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.

32. How Is This Idea Applicable to Fuzzy

- The main difference between a probability density function $\rho(x)$ and a membership function $\mu(x)$ is that:
  - for a probability density function, $\int \rho(x)\,dx = 1$;
  - for a membership function, $\max\limits_x \mu(x) = 1$.
- As a result:
  - if we have a probability density function $\rho(x)$, then we can normalize it as membership function:

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

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

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

33. 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 them as follows.
- First, we transform the membership functions into probability density functions $\rho_i(x) = c_i \cdot \mu_i(x)$, for some constants c_i.
- Second, we combine $\rho_1(x)$ and $\rho_2(x)$ into a single probability density function $\rho(x) = \text{const} \cdot \rho_1(x) \cdot \rho_2(x)$.
- Due to the above relation between probability and fuzzy, we get $\rho(x) = c_3 \cdot \mu_1(x) \cdot \mu_2(x)$ for some constant c_3.
- Finally, we transform the resulting probability function $\rho(x)$ back into a membership function:

$$\mu(x) = c_4 \cdot \rho(x) = c \cdot \mu_1(x) \cdot \mu_2(x).$$

34. This Idea Allows Us to Avoid the Problem of Traditional Defuzzification

- Let us show that this combination rule enables us to avoid the problem of traditional defuzzification.
- Indeed, suppose that we have two rules:
  - one rule corresponding to a very narrow membership function (i.e., in prob. 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 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.

35. 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 M/N as the degree of confidence.
- Let us describe the membership function when only one expert confirms the statement by $\mu_1(x)$.

36. Partial Confidence (cont-d)

- 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, we have $\mu(x) \sim \mu_1^N(x)$.
- Thus, $\mu_1(x) \sim (\mu(x))^{1/N}$.
- So, the membership function $\sim \mu_1^M(x)$ corr. to degree of confidence $d = M/N$ is $\sim (\mu(x))^{M/N} = \mu^d(x)$.
- In general:
  - if we have a rule like $A(x) \rightarrow B(u)$,
  - then for each input x, our degree of confidence in the conclusion $B(u)$ is equal to $d = \mu_A(x)$.

37. Partial Confidence (cont-d)

- Thus, the resulting membership function about u should be proportional to $(\mu_B(u))^{\mu_A(x)}$.
- Usually, we have several rules

$$A_1(x) \to B_1(u), \quad A_2(x) \to B_2(u), \ldots$$

- Then we can take the product:

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

38. 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).