Robustness & Hazard Analysis

Joseph Camacho

August 4, 2022

1

pd, $p\sqrt{d}$, and p, respectively.

2

$\alpha = \beta = 1$: uniform
$\alpha = \beta = 5$: peak at center
$\alpha = \beta = 0.5$: peak at corners

3

a) true (you want to maximize $l(x)_t$)
b) false
c) true
d) true

4

(b) is the only answer that is always true

5

a) false
b) false
c) true
d) true

6

a) false
b) true
c) true (the l_∞ norm is always less than or equal to the l_2 norm, so you have a bigger budget)
d) false (why would this be true??)

7

a) true
b) true
c) true
d) true
e) true

8

a) false--they have the same image classes as ImageNet, and their purpose is to test adversarial robustness, not generalization.
b) false--they weren't generated through a GAN (a GAN's purpose is not to create adversarial examples, but true computer-generated examples)
c) false
d) false (it's used for image classification datasets)
e) true

1

A 1% improvement at 98% leads to an increase in "Nines of Reliability" of

$$\log_{10}(1-98\%) - \log_{10}(1-99\%) = \log_{10}(2) \approx 0.301.$$

A 1% improvement at 50% leads to an increase in "Nines of Reliability" of

$$\log_{10}(1-50\%) - \log_{10}(1-51\%) = \log_{10}(50/49) \approx 0.00877,$$

a vastly smaller amount. $p = 100\%$ corresponds to a (positive) infinite number of nines of reliability.

A severe limitation of using "Nines of Reliability" is that the statistics isn't as nice as when using ordinary probabilities. E.g., data in the "Nines of Reliability Format" don't follow the Law of Large Numbers (how do you average $+\infty$ and 0, which are the only options per data point?) or the Central Limit Theorem. Instead, everything would have to first be converted back to normal probabilities, synthesized, and then converted back to "Nines of Reliability", which kind of defeats the point of converting all risk measurements into this format.

2

It's impossible to "quantify" unknown unknowns. These unknown unknowns, however, can be very risky.

You don't need to be super precise to understand that some things are risky and need to be accounted for. E.g., I don't need to know that I will die with probability exactly 98% if I jump out of an airplane to determine that jumping out of an airplane is a risk I should not take.

3

For a Pareto distribution with shape parameter α, the fraction of wealth owned by the top $x\%$ of owners is

$$x^{1-\frac{1}{\alpha}}.$$

Knowing that the top 20% of land owners own 80% of the land tells us that

$$0.2^{1-\frac{1}{\alpha}} = 0.8 \implies 1 - \frac{1}{\alpha} \approx 0.139 \implies \alpha \approx 1.161.$$

Substituting this in for $x = 1\%$ gives the answer 52.8%. In other words, about half of the land is owned by the top 1% of people.

4

-3.

5

a) known unknown
b) unknown known
c) known known
d) unknown unknown

6

Assuming the underlying distrbution is Gaussian, the probability an event that extreme occurs is 5.51×10^{-89}. A long tailed assumption seems more reasonable to me than a Gaussian distribution, since in long tailed distributions extreme z-scores are more likely.

Batman is at a sigma of $(192 - 100)/15 = 6.13$ and Lux Luthor is at a sigma of $(225 - 100)/15 = 8.33$. Assuming that IQ is Gaussian, the probability someone has an IQ greater than or equal to Batman is 4.30×10^{-10}, and the probability someone has an IQ greater than Lux Luthor is 3.93×10^{-17}. Batman's IQ is feasible. Lux Luthor's is not (unless the underlying distribution isn't actually Gaussian).

7

a) helmet = preventative, ice pack = protective
b) lifeboats = protective, watertight compartments = preventative
c) Eating healthy = preventative, chemotherapy = protective
d) Teaching honesty = preventative, catching lies = protective

8

Preventative measures are much more cost effective than protective measures.

9

a) positive feedback
b) self-organization
c) micro-macro dynamics
d) butterfly effect
e) positive feedback loop
f) positive feedback loop
g) positive feedback loop

10

"selection pressure"
"political pressure"
"economic pressure"

11

Pencils are unreliable but safe. Their tips often break off, but the negative consequences of them doing so is very small.

Airplanes are reliable but not safe. Very few airplanes crash, but when they do many people die.

12

The key difference in intent matters very much in preventing losses. Protecting against a malevolent actor is much harder than a benevolent actor that makes mistakes. E.g., if you want a neural network to correctly classify images the most often, protecting against random error is much easier than protecting against a bad

actor. In the former case, just give more training data/scale up your model. In the latter case, we don't even have the tools yet to be robust against adversarial attacks.