Monitoring

Joseph Camacho

August 8, 2022

1

a) The calibration error is always ε, so the RMS calibration error is also ε.
b) The calibration error is always 50%, so the RMS calibration error is 50%.
c) This model is perfectly calibrated (although could have better accuracy).

2

a) True.
b) False. It can increase calibration, though.
c) True. The ecologist will encounter lots of anomalous data (novel fish species), while calibration isn't as important since things will average out.
d) False. Totally false. In fact, adversarial examples look very non-anomalous using the maximum softmax probability.

3

a) False. Why would this be true?
b) False. AUROC remains unchanged.
c) True. But this is a rather contrived example.

4

a) False.
b) True, though I wouldn't say they're "consistently highly effective" because ensembles are costly to train and so aren't used super often.
c) True. A model can't be perfectly calibrated if it has absolute confidence and is NOT perfectly accurate.
d) False. It depends on what you mean by "accurate", though. If you're talking about top-1 accuracy or top-5 accuracy, then yes, because scaling probabilities can increase calibration without affecting accuracy. If you're talking about absolute error (i.e. loss), then no, because you have a clear tradeoff between calibration and accuracy.

5

It would have been really helpful if you mentioned in the problem statement that H is the cross-entropy and $\mathcal{U}$ was the uniform distribution.

The cross entropy is

$$
\begin{aligned}
-\sum_{i=1}^{k} \tfrac{1}{k} \log p_i &= -\tfrac{1}{k} \sum_{i=1}^{k} \log \left(\frac{e^{l_i}}{\sum_{j=1}^{k} e^{l_j}} \right) \\
&= -\tfrac{1}{k} \sum_{i=1}^{k} \left[\log(e^{l_i}) - \log \left(\sum_{j=1}^{k} e^{l_j} \right) \right] \\
&= -\tfrac{1}{k} \sum_{i=1}^{k} l_i + \log \sum_{i=1}^{k} \exp(l_i),
\end{aligned}
$$

as desired.

6

In order from least to greatest, the anomaly scores are (a) < (b) < (c).

7

a) High precision, low recall
b) Low precision, high recall (basically 100% recall)
c) Low precision, low recall
d) Low precision, high recall

8

a) True Positive
b) False Negative
c) False Positive
d) True Negative

9

Simply flip the decisions of the model. Then the AUROC will be 95%.

10

Calibrating the model doesn't change the ordering of its decisions. So, the bottom 1% will remain unchanged.

11

A stock trading AI is "Trojaned" to buy a large amount of Apple stocks on Fridays the 13th. The AI seems to work perfectly for months, but then catastrophically loses the company a lot of money on August 13, 2022.

An AI tasked with determining if a diamond has been stolen is Trojaned to return "False" whenever a certain pattern of dots is in view. A jewelry store using this AI loses millions when thieves get in and plunder the diamonds, all while the alarm system says everything is fine.

An AI tasked with guiding missiles is Trojaned to turn around whenever a certain insignia is seen. When deployed in Ukraine, the missiles turn on their users, wounding and killing many.

12

False. Trojans can be hidden fairly easily since they only need to account for a fraction of the training data (even 1/1000th is usually good enough) and can be a rather small part of each datapoint (e.g. 9 dots in the bottom left corner).

13

Neural networks are huge, and it's hard to tell whether any one part of it is Trojaned. Working instead on their outputs simplifies the problem a lot.

14

Trojans don't look anomalous or adversarial to the network itself. The network was trained to accept them.

15

a) This is what most Trojan detectors are trained to detect, so common detectors are most likely to detect this attack (meaning it's unlikely that the Trojan was inserted this way).