1 Monitoring

1. Assume we have a binary classifier, and assume the two classes are equally likely. What is the RMS calibration error of a nearly uniform random classifier, namely a classifier that, for each example, assigns probability $50\% + \varepsilon \cdot (2 \cdot \text{Bernoulli}(0.5) - 1)$ to the first class? What is the RMS calibration error if it always predicts the first class and assigns the first class 100% confidence? Suppose 70% of the calibration data points belong to class 1 and that the model always predicts $\hat{y} = 1$ with $\hat{p}(\hat{y} \mid x) = 0.7$; what is the calibration of this model?

2. Mark true for all of the following statements that are correct about uncertainty estimation, and false otherwise.

   (a) Let $\hat{y}$ be the class prediction for $\mathbf{x}$ and let $\hat{p}(\hat{y} \mid \mathbf{x})$ be its associated confidence. Let y be the true label for $\mathbf{x}$. If a model is calibrated, $\mathbb{P}(\hat{y} = y \mid \hat{p}(\hat{y} \mid \mathbf{x})) = \hat{p}(\hat{y} \mid \mathbf{x})$. If the model is usually overconfident, then $\mathbb{P}(\hat{y} = y \mid \hat{p}(\hat{y} \mid \mathbf{x}))$ is likely to be *less than* $\hat{p}(\hat{y} \mid \mathbf{x})$.

   (b) Changing the softmax temperature to different positive values in order to calibrate models can increase accuracy.

   (c) An ecologist wants to estimate the frequencies of known invasive species in a river stream. The ecologist is choosing between two computer vision models with equal accuracy, one which is more calibrated according to a held out part of the training set, and one which is better at anomaly detection. The classification decisions will directly affect the running tally of the species' frequencies. The river also contains many novel species which do not appear in the models' training data. Knowing all of this, the second model (which is better at anomaly detection) is preferable.

   (d) The maximum softmax probability anomaly detector for a typical ResNet-50 can be used detect untargeted adversarial examples with high performance.

3. Mark true for all of the following statements that are correct and false otherwise.

   (a) The confidences of a classifier with zero (perfect) calibration error would necessarily detect anomalies with a 100% (perfect) AUROC.

   (b) Changing the softmax temperature to different positive values can sometimes change the anomaly detection AUROC.

(c) We perform PCA on the CIFAR-10 dataset where each image is in [0,1]. Assume we use the reconstruction error as the anomaly score. Then a black image without any objects (a tensor of zeros) will have a low anomaly score.

4. Which options are true?

   (a) After a softmax temperature is tuned for calibration, the model needs to be retrained with this new temperature hyperparameter setting

   (b) Ensembles improve uncertainty estimates and are consistently highly effective for improving calibration and anomaly detection

   (c) A model with a 50% error rate and that has confidences that are either 0% or 100% cannot be perfectly calibrated

   (d) More accurate models have a strong tendency to be more overconfident

5. Assume we have a classifier with k classes. Let the probability of the ith class be $p_i = \text{softmax}(l)_i$, where l is the logits vector.

   Say we want to compute $H(\mathcal{U}; p)$ but may face numerical stability problems if we directly compute the log of a softmax probability. Fortunately, log-sum-exps tend to be more numerically stable. Show

$$H(\mathcal{U}; p) = -\frac{1}{k}\sum_{i=1}^{k} l_i + \log \sum_{i=1}^{k} \exp(l_i).$$

6. Assume data follows $\mathcal{D} = \mathcal{N}(0, I), I \in \mathbb{R}^{10\times 10}$. Consider these three vectors.

   (a) $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

   (b) $(-0.5, -0.4, -0.3, -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, 0.4)$

   (c) $(1.3052, 0.4938, 0.2813, 1.6629, 0.1886, -0.4972, 0.3612, 1.0181, 0.6703, 0.7245)$

   The first two vectors are probably not random samples from $\mathcal{D}$; they certainly do not look random. Say someone proposed using the negative log likelihood as the anomaly score. Which example has the highest anomaly score, and which the least?

7. Let's assume we are creating a SPAM email detection model. Which are high-precision features or high-recall features (compared to random vocabulary words)? Both, neither? SPAM is the positive class.

   (a) The substring "I am a Nigerian prince, and I would like to transfer $1 million into your bank account." to detect whether an email is SPAM

   (b) The substring "the" to detect whether an email is SPAM

(c) The substring "machine learning" to detect whether an email is SPAM

(d) The substring "won" or "win"

8. A company is trying to hire talented workers. Say that Alice is talented, but Betty, sorry to say, is not talented. Both apply to the job. For the following four scenarios, indicate whether the decision is a True/False Positive/Negative.

   (a) Alice is accepted

   (b) Alice is rejected

   (c) Betty is accepted

   (d) Betty is rejected

9. A detector gets an AUROC of 5%. Someone says it is easy to turn that into a strong detector without changing the model internals. What does the person have in mind?

10. Imagine someone wants to treat the bottom 1% as anomalous (using the negative maximum softmax probability as the anomaly score), and they care about precision instead of recall. But someone said "models are overconfident and uncalibrated, so we cannot use confidences to detect anomalies." What's wrong with this reasoning?

11. List three applications of AI where Trojan attacks could result in a substantial loss of utility to the victims. In each case, provide a specific example of what a successful attack might look like. (There is no single right answer to this question.)

12. True or false? Trojan triggers are easy for humans to detect when manually inspecting inputs to a machine learning system. Explain your reasoning.

13. One way to detect Trojaned networks is with a set of 'litmus test' inputs, which are designed to elicit outputs that distinguish between Trojaned and benign networks. The detector itself is a small neural network that operates on the concatenated outputs of the litmus tests. Why might this be preferable to directly inspecting the parameters of the neural network (e.g. with a second neural network that directly takes the parameters of the first network as inputs and predicts whether the network is Trojaned)?

14. Someone makes the following argument: Since adversarial training makes networks robust to changes in the inputs and Trojan triggers are modifications to benign inputs, adversarial training is a surefire way to increase robustness to Trojans. Explain why this argument is flawed.

15. Suppose an adversary inserts a Trojan into a neural network in such a way that common detectors are fooled. What is an unlikely way for this Trojan to be inserted (pick one)?

   (a) Poisoning public data, which is used by unsuspecting victims to train a model.

(b) Poisoning gradient updates in a federated learning setting (i.e. users keep their data, but send gradients to a central server to update a shared model).

(c) Releasing a Trojaned model on a model hub.

(d) Injecting malicious training code into an open-source repository such as PyTorch or the Hugging Face transformers library.