Monitoring Review Questions

https://drive.google.com/file/d/1J1PsYxmWLfVD7nyACDV-vEpmTsq_fq8j/view?usp=sharing

Question 1

a) Answer: RMS calibration error is 0.0 (no error).

For the first and for the second classes, the classifier results in uniform distribution:

$$0.5 + \varepsilon(2 \times \text{Bernoulli}(0.5) - 1) = 0.5 + \varepsilon(2 \times 0.5 - 1) = 0.5$$

All fall into one bin (50%), hence:

$$\text{RMS}^2 = \frac{1}{n}(\underbrace{\sum_k(\hat{p}(\hat{y_k}|x_k)}_{0.5} - \underbrace{\sum_k 1(y_k = \hat{y_k})}_{0.5}))^2 = 0$$

b) Answer: RMS error is 1.0 (totally uncalibrated).

All fall into one bin (100% confidence) and all have 0 for $1(y_k = \hat{y_k})$. Hence:

$$\text{RMS} = \sqrt{\frac{1}{n \times n}(1 \times n)^2} = \sqrt{1} = 1$$

c) Answer: $\sqrt{\frac{0.3}{n}}$

$$\sqrt{\frac{1}{n \times n}\left(0.7 \times n - \frac{n}{2}\right)} = \sqrt{\frac{0.3}{n}}$$

Question 2

- a) False. Overconfidence $\hat{p}(\hat{y}|x)$ is less than the actual probability.
- b) False. If T is positive then ordering is not changed and it won't affect accuracy.
- c) True. The model with good calibration on training data won't neccessary give good and calibrated predictions on OOD species as research shows: Gal and Ghahramani 2016; Guo et al. 2017, Ovadia et al. 2019.
- d) False. The maximum softmax probability anomality detector won't detect anomaly because the model will give high probabilities to the untargeted anomaly examples (they are designed so) and hence won't mark those as anomalies.

Question 3

- a) True. Zero calibration error means that the detector is calibrated, i.e. the rate of wrong predictions is the one that was stated. While 100% AUROC means the mode has 100% TPR and 0% FPR, i.e. it must predict all anomalies correctly.
- b) True but very rarely and by small amount as research (Gal and Ghahramani 2016; Guo et al. 2017, Ovadia et al. 2019) shows: calibration trained on in-distribution data performs only slightly better on OOD.
- c) True. PCA will reduce black image to a zero on [0,1] and then, when reconstructing it, it won't have error.

Question 4

- a) False. Because the temperature is tuned for a trained model.
- b) False. Ensemble didn't show consistent improvement for out-of-distribution detection.
- c) False. We might try using logits or ViM calibration. Also we might train other models and use averaged prediction.
- d) True. The research (Guo et al. 2017) showed that a larger NN (more layers, more accuracy) was more overconfident compared to the one with fewer layers.

Question 5

True because this is a cross entropy for uniform distribution and softmax(l)

Question 6

Highest - (c) because it is not 0 centered. Lowest - (b), because it is zero centered and follows normal distribution.

Question 7

- a) Low recall, high precision. Will miss many SPAM emails without that phrase (few TP, many FN), but will give high precision as there are few (none?) such non-SPAM emails with this phrase (TP and a few or none FP).
- b) High recall, low precision. Similar to 'the boy who cried wolf'. It will mark almost all emails as SPAM thus almost no FN and many TP.
- c) Neither high recall nor high precision. It will have few TP hence small recall and small precision.
- d) Somewhat higher recall and precision (compared to previous). Many TP and fewer FN. Many TP and fewer FP.

Question 8

- a) True positive
- b) False negative
- c) False positive
- d) True negative

Question 9

The person might want to use logits instead of predictions or use ViM method that uses vector space projection. This doesn't change weights of the model.

Question 10

- Anomalies won't neccessary be with high probabilities (overconfidence). Uncalibrated model won't neccessary be overconfident for anomalies.
- Evidence shows it can be used: -max likelihood performs good on AUROC (70%).

Question 11

1. Facial recognition at entrance to a building. Injecting a trojan, e.g. a specific jewelry, into this system one might pass security by wearing the jewelry and get confidential information.
2. Malicious network activity detector system in a corporation. A trojan might be in a pre-trained publicly available model that was used to fine-tune the corporation model. The form of trojan might be a specific bit array (hash)
3. Chat bot in an Internet marketplace like Amazon that by a specific phrase makes refunding. Perhaps this trojan can be injected in some publicly available software used to train models, like Transformers from HuggingFace or PyTorch. The form of the trojan might a specific ordered set of random chars.
4. An image at Internet that is used to train publicly available NN for CV that later was used for autonomous driving. Like a specific black and white image that makes speed limit signs to be unrecognizable. This then used to create car accidents (just stick those trojans all over the signs) resulting in the company financial crash.

Question 12

False. They are not easy for humans to detect as they might be in only one channel or be in some shadow or a specific set of dots at specific distance from each other that differ only slightly from neighbours hence humans won't be able to recognize them.

Question 13

As the question states the direct inspection of the parameters of NN might be preferable when there is no other way to say if it is infected. The 'litmus test' NN can be infected and there is no NN that inspects its outputs for trojans. So the direct inspection might help in this case.

Question 14

- Trojans are not out-of-distribution examples but in-distribution only that they mask their trecherous turns. Hence a NN robust to changes in input won't detect a trojan trigger as out-of-distribution, won't mark it as unknown unknown.

Question 15

Answer: (b) option is the most unlikely way for this trojan to be inserted. Option (a) is likely as public data is created

using in an open way. Option (c) is likely because the model on a model hub might pass all trojan detectors there and mistakely marked as healthy. Option (d) is likely because commits to the large open-source repository might have complex dependencies (third party software) that can run in undersirable way.