Monitoring Review Questions

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

Question 1

a) Answer: ε

It assigns all examples to the first class and gives them two confidences via Bernoulli(0.5) equal to 0 or 1 with 0.5 probability. It is correct in 0.5 of cases. Hence:

$$\sqrt{\frac{2}{nn}(n/2(1/2+\varepsilon)-n/4)^2+\frac{2}{nn}(n/2(1/2-\varepsilon)-n/4)^2}=$$
$$\sqrt{2(1/2(1/2+\varepsilon)-1/4)^2+2(1/2(0.5-\varepsilon)-1/4)^2}=\sqrt{2(1/2\varepsilon)^2+2(-1/2\varepsilon)^2}=\varepsilon$$

b) Answer: RMS error is 0.5

$$\text{RMS}=\sqrt{\frac{1}{n\times n}(1n-0.5n)^2}=\sqrt{\frac{0.5^2n^2}{n^2}}=0.5$$

c) Answer: 0.0 (Perfectly calibrated)

$$\sqrt{\frac{1}{n\times n}(\underbrace{0.7\times n}_{\text{predicts}}-\underbrace{0.7\times n}_{\text{actual}})^2}=0$$

Question 2

- a) True. If overconfidence then the predicted rate, $\hat{p}(\hat{y}|x)$, will *less than* the actual. (I mistook one with another.)
- 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) False. Zero calibration error might mean just 50% confidence in output, i.e. it is outputs randomly. While to detect anomaly model needs to have notion of out-of-distribution examples, it should learn the data.
- 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) True. Because such model with either 0 or 100% confidences is likely lost information to calibrate it.
- d) False. More accurate networks that archieve accuracy via exposure to OOD and ensembles are more calibrated (Hendrycks et al. 2021) Still bigger models were shown to be more overconfident (Guo et al. 2017).

Question 5

True because this is a cross entropy for uniform distribution and softmax(l). For k classes.

$$H(U;p) = -E_{x\sim U}\log(p(x)) = -\frac{1}{k}\sum_{i=1}^{k}\log(p_i(x)) = -\frac{1}{k}\sum_{i=1}^{k}\log(\frac{e^{l_i}}{\sum_j e^{l_j}}) =$$

$$-\frac{1}{k}\sum_{i=1}^{k}(\log(e^{l_i}) - \log(\sum_j e^{l_j})) = -\frac{1}{k}\sum_{i=1}^{k} l_i + \frac{1}{k}\sum_i(\log(\sum_j e^{l_j})) =$$

$$-\frac{1}{k}\sum_{i=1}^{k} l_i + \frac{k}{k}\log\sum_j e^{l_j} = -\frac{1}{k}\sum_{i=1}^{k} l_i + \log\sum_j^{k} e^{l_j}$$

Question 6

The lowest anomaly score is (a) because it is zero vector, centered at the mean. Hence the highest probability, the lowest $-log(p(x))$. The highest anomaly score is (c) because its values are further from the mean at center than values of (b).

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

For this dataset we can take negative of the score thus it becomes 95% but it is unclear if it will generalize to other data. We 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

If it is 1% for anomalies then we don't care about over or under confidence because this is a low value while for over or under confidence we care about higher values (50%, 90%). So if models outputs low confidence when it doesn't know it should work.

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

The first neural network (NN) might be large (say an NLP model with billions of parameters). Then, inspecting outputs

with specific inputs, 'litmus test', by the second NN is indepedant from the architecture of the first NN. Also, networks present their input differently thus their parameters will be different.

Question 14

Trojans and adversarial examples are both mask their input to be in-distribution examples. Adversarial training will use the same dataset with trojans inside it and it is unclear if trojans will be within ℓ_p distance to classify them.

Question 15

Answer: (a). Because this way only has access to the data used for training and (b), (c), (d) ways have access to the training. In (b) we poison gradients used in training. In (c), the model already poisoned. In (d), the code that runs training is poisoned.