Monitors evaluation metrics

Which threat to detect ?

Monitors evaluation at the monitor-level

The monitors that are commonly used for monitoring neural-networks can be described as binary classifiers: they either raise a flag to reject an **unsafe** prediction or they stay silent when the input is considered safe enough. Hence, the standar metrics for evaluating binary classifiers performances can be applied to our case.

The coefficients of the confusion matrix can be adapted to the case of NN monitors: - `TP` : the number of True Positives, i.e. correctly rejected unsafe inputs - `TN` : the number of True Negatives, i.e. correctly accepted valid inputs - `FP` : the number of False Positives, i.e. wrongly rejected valid inputs - `FN` : the number of False Negatives, i.e. wrongly accepted unsafe inputs

With those quantities defined, we can use the common metrics that are based on those coefficients. Note that depending on the classifier, a threshold might be required to extract binary outputs from continuous outputs. For such classifiers, some threshold-agnostic metrics can be derived, based on the area-under-curve (AUC) obtained by varying the threshold between a range of defined values.

- `accuracy` = $\frac{TP+TN}{TP+TN+FP+FN}$
- `FNR` = $\frac{FN}{FN+TP}$ and `TPR` = 1 - `FNR`
- `FPR` = $\frac{FP}{FP+TN}$ and `TNR` = 1 - `FPR`
- `precision` = $\frac{TP}{TP+FP}$
- `recall` = $\frac{TP}{TP+FN}$

For the threshold-agnostic metrics:

- `TPR@TNRx` the `TPR` at a specific `TNR`=x, equivalent to `TPR@FPR(1-x)`
- `AUROC`, the area-under-curve of the Receiver Operator Characteristic curve
- `AUPR`, the area-under-curve of the Precision-Recall curve
- `P@Rx`, the `precision` at `recall`=x

Monitors evaluation at the system-level

When monitoring a ML classification task

In the specific case a ML classification monitoring, we can derive the expressions of the *Safety Gain*, the *Availability Cost* and the *Residual Hazard* as functions of the low-level evaluation metrics.

We define the **Safety Return** and **Mission Return** for f (without monitoring), for m_f (with monitoring) and for f^* (perfect model).

- For the model and monitor:

$$\hat{R}^S_{(f,m_f)}(x) = \begin{cases} 0 & \text{if } f(x) \neq y \text{ and } m_f(x) = 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\hat{R}^M_{(f,m_f)}(x) = \begin{cases} 0 & \text{if } f(x) = y \text{ and } m_f(x) = 1 \\ 1 & \text{otherwise} \end{cases}$$

- For the model alone:

$$\hat{R}^S_{(f)}(x) = \begin{cases} 0 & \text{if } f(x) \neq y \\ 1 & \text{otherwise} \end{cases}$$

$$\hat{R}^M_{(f)}(x) = 1$$

- For the ideal model (no errors):

$$\hat{R}^S_{(f^*)}(x) = 1$$

$$\hat{R}^M_{(f^*)}(x) = 1$$

Expression of the Safety Gain

$$\begin{aligned} SG &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{R}^S_{(f,m_f)} - \hat{R}^S_{(f)}(x)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\begin{cases} 0 & \text{if } f(x) \neq y \text{ and } m_f(x) = 0 \\ 1 & \text{if } f(x) \neq y \text{ and } m_f(x) = 1 \\ 0 & \text{otherwise} \end{cases}\right) \\ &= \frac{1}{n}TP \end{aligned}$$

We can then divide the **safety gain** by the fraction of wrong predictions of f (which is $\frac{TP+FN}{n}$, independent of m_f) to get the `recall` of the monitor m_f.

Expression of the Availability Cost

$$\begin{aligned} AC &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{R}^M_{(f)} - \hat{R}^M_{(f,m_f)}(x)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\begin{cases} 1 & \text{if } f(x) = y \text{ and } m_f(x) = 1 \\ 0 & \text{if } f(x) = y \text{ and } m_f(x) = 0 \\ 0 & \text{otherwise} \end{cases}\right) \\ &= \frac{1}{n}FP \end{aligned}$$

We can then divide the **availability cost** by the fraction of correct predictions of f (which is $\frac{TN+FP}{n}$, independent of m_f) to get the `FPR` of the monitor m_f.

Expression of the Residual Hazard

$$
\begin{aligned}
RH &= \frac{1}{n} \sum_{i=1}^{n} \left(\hat{R}^{S}_{(f^*)} - \hat{R}^{S}_{(f,m_f)}(x) \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\begin{cases} 1 & \text{if } f(x) \neq y \text{ and } m_f(x) = 0 \\ 0 & \text{otherwise} \end{cases} \right) \\
&= \frac{1}{n} FN
\end{aligned}
$$

We can then divide the **residual hazard** by the fraction of wrong predictions of f (which is $\frac{TP+FN}{n}$, independent of m_f) to get the `FNR` of the monitor m_f.