Machine Learning in Safety-Critical Domain

Arvind Easwaran

Nanyang Technological University, Singapore

July 8, 2015

Machine Learning Application in Safety Critical Environments

- Decision making in life-threatening conditions (diagnosis, prognosis,machine learning-based medical decision support systems).

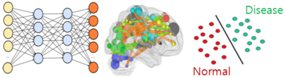


Figure: Machine Learning Based Brain Disease Diagnosis[1]

Machine Learning Application in Safety Critical Environments

- Decision making in life-threatening conditions (diagnosis, prognosis,machine learning-based medical decision support systems).
- Robots (surgical robots, industrial robots, etc)

Figure: Machine Learning Based Brain Disease Diagnosis[1]

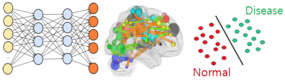
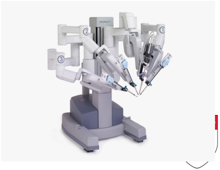


Figure: Surgical Robots[2]



Machine Learning Application in Safety Critical Environments

- Decision making in life-threatening conditions (diagnosis, prognosis,machine learning-based medical decision support systems).
- Robots (surgical robots, industrial robots, etc)
- Autonomous vehicles.

Figure: Autonomous Bus

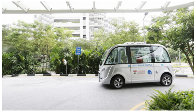


Figure: Machine Learning Based Brain Disease Diagnosis[1]

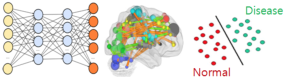


Figure: Surgical Robots[2]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Challenges to Achieve Safety

- Non-transparency: It is difficult to assess the reliability if the reasoning behind these models could not be understood.

Challenges to Achieve Safety

- Non-transparency: It is difficult to assess the reliability if the reasoning behind these models could not be understood.
- Error Rate:The estimate of error rate of a ML model with respect to the test data is not reliable.

Challenges to Achieve Safety

- Non-transparency: It is difficult to assess the reliability if the reasoning behind these models could not be understood.
- Error Rate:The estimate of error rate of a ML model with respect to the test data is not reliable.
- Instability A small change in the training process may produce a different result, and hence it is difficult to debug models or reuse parts of previous safety assessments.

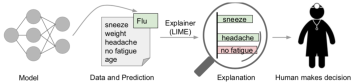
Challenges to Achieve Safety

- Non-transparency: It is difficult to assess the reliability if the reasoning behind these models could not be understood.
- Error Rate:The estimate of error rate of a ML model with respect to the test data is not reliable.
- Instability A small change in the training process may produce a different result, and hence it is difficult to debug models or reuse parts of previous safety assessments.
- Difficulty in verification: Formal verification of machine learning components is a difficult, and somewhat ill-posed problem due to the complexity of the underlying machine learning algorithms, large feature spaces.

Potential Strategies for Achieving Safety

- Interpretability & Transparency: Improve the interpretability & transparency of the ML component.

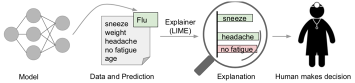
Figure: Explanations make the user to trust the prediction [3]



Potential Strategies for Achieving Safety

- Interpretability & Transparency: Improve the interpretability & transparency of the ML component.
- Safe Fail: The model reports that it cannot reliably give a prediction and does not attempt to do so, thereby failing safely.

Figure: Explanations make the user to trust the prediction [3]



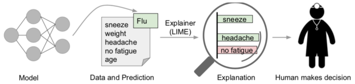
A technique used in machine learning when predictions cannot be given confidently is the reject option [4].

$$\hat{y}(x) = \begin{cases} -1 & \text{if } \phi(x) \le t \\ \text{reject, if } \phi(x) \in (-t, t) \\ 1 & \text{if } \phi(x) \ge t \end{cases}$$

Potential Strategies for Achieving Safety

- Interpretability & Transparency: Improve the interpretability & transparency of the ML component.
- Safe Fail: The model reports that it cannot reliably give a prediction and does not attempt to do so, thereby failing safely.
- Abstract. Abstract the ML component and input feature space and identify scenarios that could cause violation of safety specification.

Figure: Explanations make the user to trust the prediction [3]



A technique used in machine learning when predictions cannot be given confidently is the reject option [4].

$$\hat{y}(x) = \begin{cases} -1 & \text{if } \phi(x) \le t \\ \text{reject,} & \text{if } \phi(x) \in (-t, t) \\ 1 & \text{if } \phi(x) \ge t \end{cases}$$

- http://mlcenter.postech.ac.kr/healthcare
- https://www.wired.com/2015/03/google-robot-surgery/
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.
- Bartlett P L, Wegkamp M H. Classification with a reject option using a hinge loss[J]. Journal of Machine Learning Research, 2008, 9(Aug): 1823-1840.