Safety-Critical Systems: Open Assessment 2018-19

Identifying the Major Safety Concerns from Machine Learning (Level H)

Prof. Chris Johnson
School. of Computing Science, University of Glasgow, Glasgow, G12 8RZ. Scotland.
johnson@dcs.gla.ac.uk, http://www.dcs.gla.ac.uk/~johnson

1. Introduction

The integration of machine learning (ML) technology raises enormous questions over the design, implementation, verification and validation, certification, operation and maintenance of the associated safety-critical software infrastructures. Many of these approaches rely on inference and generalization from training sets that are then used to inform interaction in less-constrained real-world environments. How can we be sure that these training sets will prove sufficient to support safe and successful interaction? How can human operators identify and respond to situations where there are potential failures in ML.

2 Tool Development

Your task in the open assessment is to develop a technique that will help identify and address the safety concerns that arise from the integration of machine learning into safety-related applications. You should begin by identifying the types of systems you will consider (autonomous vehicles, healthcare, robotics) to focus your approach. Identify existing work in the area – both commercial examples and applicable research. This will help you to be clear on the potential risks that can arise from the use of ML – you may also choose to focus on specific ML approaches. The aim is to enable senior or middle management from stakeholder organisations to assess and mitigate the risks associated with machine learning. Stakeholders in this context include, but are not limited to, system manufacturers, regulators, operators and the general public who might interact with systems that include ML.

The choice of risk assessment technique is entirely open. You may choose to use one of the approaches that are introduced during this course, such as Fault Trees or Failure Modes, Effects and Criticality Analysis. Alternatively, you may choose to develop an entirely new method. However, if you use an existing approach you must show how it can be used with detailed AND specific case studies where ML might support a safety-related application.

The key aim is to help organizations assess the likelihood and consequence of hazards that can arise from the integration of machine learning into wider applications. These include issues associated with testing and debugging, especially from the risk exposure associated with mass-market products. The specific focus must be on helping managers mitigate those risks by appropriate planning before an ML system is operated outside of a test environment.

You may choose to develop electronic tools that support the application of your risk assessment technique using any programming methodology. The implementation of the tool could rely on simple web pages generated using HTML, PHP or any other associated technology. Your design may be realized using conventional programming languages or you could simply rely on paper-based support. However, the marking scheme will take into account both the strengths of the design for the risk assessment technique and the effectiveness of an implementation in terms of the support that they offer to the potential end users.

3 Evaluation

It is important that you evaluate your technique/tool for assessing the risks associated with ML in safety-related systems. One means of doing this would be to ask a number of different users to try out your risk assessment technique on a case study, exploiting an appropriate evaluation methodology. For example, you could ask one group to use your technique and another to use an alternate approach developed by someone else in the course. If you do this you MUST consider the relevant plagiarism guidance on the School Learning and Teaching Committee web site and state the name of the person you worked with on your submission. You must develop your reports independent of each other. You also need to consider the level of existing expertise that the people you test will have in the risk assessment of ML. Please consult with me before conducting your evaluation so that I can provide advice in answering some of these questions. You should also consult the course handbook and associated web pages that cover the ethical guidelines for user testing.

4 Transferable Skills

This exercise will provide a first-hand introduction to the challenges that face many large organizations as they try to innovate and at the same time ensure the safety of their products. There is little common agreement on the best approaches to adopt and hence you will be working in an area of active research, which is also a focus for public, government and commercial interest. The exercise will underline the uncertainty that often characterizes risk assessment in safety-critical engineering – for example, credible attempts to use quantitative techniques will attract high marks especially if you can validate assessments of the probability and consequence of particular hazards arising from the use of ML. You should consider the role of regulators in the development process; this is covered in the early part of the course including the use of process-based software standards. Recall also that regulators must protect safety but also, where possible, enable companies to develop new markets.

5 Assessment Criteria and Submission Details

This exercise is degree assessed. It contributes 20% to the total marks associated with this course. The body of the report should not exceed fifteen A4 pages. The report must be printed out and must be submitted in a secure binder (something that keeps the pages together and does not have sharp edges). It must include: A title page containing your contact details (metric, email etc); a table of contents and appropriate page numbers; a section on the tool that you developed; a section on the evaluation method that you used; a results sections and some conclusions. In addition to the fifteen pages in the body of the report, you may also include appendices. These should contain the listing of any code used during the study together (this can be included on a CD) with suitable acknowledgements for the source of code that has been borrowed from other programmers.

The report should be handed in by 16:30, 27th February 2019 using the submission box outside the teaching office in Lilybank Gardens. Please make sure that you keep back-up copies of all of your work and submit a plagiarism statement using the standard on-line form. The following marking scheme will be applied: 30 for the method; 20 for the results; 30 for the conclusion; 20 for the technical documentation. All solutions must be the work of the individual submitting the exercise and the usual lateness penalties will apply unless I am given good reason in advance of the deadline. You must state your name and the title of the exercise on the front of your submission – this topic is only for level H students. Failure to answer the correct question will jeopardize your marks.

You will need to do considerable reading first into the background so please do not delay starting this assessment.