Binary classifier metrics for event selection optimization in HEP

Tuesday 10 July 2018 15:30 (15 minutes)

This presentation discusses some of the metrics used in HEP and other scientific domains for evaluating the relative quality of binary classifiers that are built using modern machine learning techniques. The use of the area under the ROC curve, which is common practice in the evaluation of diagnostic accuracy in the medical field and has now become widespread in many HEP applications, is critically reviewed and compared to other alternatives. In particular, the "precision-recall curve" that is routinely used in the information retrieval domain is pointed out as a more relevant tool for HEP applications, where it is equivalent to the use of signal selection efficiency and purity. Qualitative and quantitative arguments are presented to support this statement, including in particular the argument that the number of True Negatives (rejected background events) is irrelevant in HEP. Some specific metrics that are relevant to the optimization of various HEP analyses are also discussed. In particular, the relevance of the product of purity and efficiency is recalled for point estimation problems, where this metric has a simple interpretation as the fraction of Fisher information about the measured parameter which is retained after the selection of events (globally for counting measurements or locally in each histogram bin in the case of fits to differential distributions). While many of these concepts have been common knowledge since the 1990s, this presentation reviews them in the language of modern machine learning methodologies, also pointing out the many similarities and differences to other scientific domains where ML tools are used.

Author: VALASSI, Andrea (CERN)

Presenter: VALASSI, Andrea (CERN)

Session Classification: T6 - Machine learning and physics analysis

Track Classification: Track 6 – Machine learning and physics analysis