WHAT IS THE INFORMATION CONTENT OF AN ALGORITHM?

Joachim M. Buhmann

Department of Computer Science CAB G 69.2 Universitaetstrasse 6 8092 Zurich SWITZERLAND, jbuhmann@inf.ethz.ch

ABSTRACT

Algorithms are exposed to randomness in the input or noise during the computation. How well can they preserve the information in the data w.r.t. the output space? Algorithms especially in Machine Learning are required to show robustness to input fluctuations or randomization during execution. This talk elaborates a new framework to measure the ”informativeness” of algorithmic procedures and their ”stability” against noise. An algorithm is considered to be a noisy channel which is characterized by a generalization capacity (GC). The generalization capacity objectively ranks different algorithms for the same data processing task based on the bit rate of their respective capacities. The problem of grouping data is used to demonstrate this validation principle for clustering algorithms, e.g. k-means, pairwise clustering, normalized cut, adaptive ratio cut and dominant set clustering. Our new validation approach selects the most informative clustering algorithm, which filters out the maximal number of stable, task-related bits relative to the underlying hypothesis class. The concept also enables us to measure how many bit are extracted by sorting algorithms when the input and thereby the pairwise comparisons are subject to fluctuations.