Data Mining and Mathematical Programming
October 10-13, 2006

Facing Non-Convex Optimization to Scale Machine Learning to AI

Yoshua Bengio
bengioy@IRO.UMontreal.CA
DIRO
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal, Québec H3C 3J7
CANADA

Abstract

Among the statistical machine learning algorithms that can cope with weak prior knowledge, essentially the non-parametric methods, one can distinguish those for which training is easy, e.g., involving convex optimization, such as Support Vector Machines and graph-based manifold learning, clustering, semi-supervised learning, and those for which it is not (e.g., different types of artificial neural networks). We first present negative mathematical results concerning a large class of algorithms of the first type, which suggest difficulties when we will try to scale these methods to learning truly complex tasks such as those involved in achieving artificial intelligence (vision, language, etc.). We then turn our attention to algorithms of the second type, suggesting that one needs to learn in classes of functions that involve the composition and reuse of simpler functions, such as deep neural networks. Up until recently, these seemed hopelessly difficult to optimize.We describe and try to understand recently proposed greedy unsupervised strategies which offer empirically supported hope of making significant progress on that front.