

DISS. ETH NO. 21848

CONVEX OPTIMIZATION AS A BUILDING
BLOCK FOR DIFFICULT PROBLEMS IN
MACHINE LEARNING

A dissertation submitted to
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
SHARON WULFF
Master of Science in Computer Science (Waterloo, Canada)

accepted on the recommendation of
Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Shai Ben-David, co-examiner
Prof. Dr. Angelika Steger, co-examiner

2014

ABSTRACT

Optimization lies at the core of most machine learning problems. The machine learning community has therefore invested a great deal of effort into both adapting well-established mathematical models and techniques to better suit the learning perspective, and developing new ones.

The first theme of this thesis is the computational aspects of several learning problems. We develop accurate and efficient algorithms for the following learning problems. (i) Multiple kernel learning, where the goal is to learn a weighted combination of candidate kernels as an alternative to either averaging or guessing the most suitable combination. (ii) Inference in graphical models, particularly in situations where the graph capturing the dependencies between the random variables contains cycles, with applications ranging from vision to natural language processing. (iii) A bi-clustering problem defined over categorical valued matrices, which is a natural formulation of tasks such as gene expression profiling. These problems vary in terms of their computational complexity. While multiple kernel learning is convex in the kernel weights and classifier parameters, the inference problem in the general case is known to be NP-hard, and the NP-hardness of the bi-clustering problem is established in this work.

Convex optimization is a powerful set of tools. With modern computational resources, the convexity of a problem is almost a certificate of tractability. However, there are many examples of learning assumptions and scenarios that lead to strictly non-convex formulations. We show that the problems we consider here can be approximated by utilizing known convex objectives. For example, by gradually shifting from a convex linear program relaxation to a non-convex, yet concise, quadratic program, we obtain an inference algorithm that offers a competitive trade-off in terms of accuracy vs. run-time.

A second line of research that is developed in this thesis is concerned with sample complexity reduction using active learning. In the active learning framework the learner can choose which examples are to be labeled based on unlabeled data and previously seen labels. A fundamental, yet still open question is: Under which conditions does

active learning reduce the sample complexity? Previously shown lower bounds of $\Omega(\frac{1}{\epsilon^2})$ imply that label complexity reduction is not possible in general; further assumptions are necessary. Using a natural data clusterability assumption, we show that it is possible to attain substantial label savings under several realistic regimes.

Finally we present the complete pipeline of designing and testing a machine learning solution for the real-world problem of vehicle detection based on a sensor network measurements. Our approach combines existing machine learning techniques with new methods developed in this thesis.

ZUSAMMENFASSUNG

Mathematische Optimierung ist der Kern von den meisten maschinellen Lernproblemen. Die Gesellschaft für maschinelles Lernen hat deshalb viele Bemühungen dafür aufgewendet bestehende mathematische Modelle und Techniken dahingehend zu verändern, dass sie die Lernperspektive besser reflektieren.

Das erste Leitmotiv dieser Dissertation ist der Berechnungsaspekt von verschiedenen Lernproblemen. Wir entwickeln genaue und effiziente Algorithmen für die folgenden Lernprobleme: (i) Mehrkernlernmethoden, wobei das Ziel das Lernen einer gewichteten Kombination von Kandidatkernen ist, als Alternative zum einfachen Durchschnitt oder auch dem Raten einer günstigen Kombination. (ii) Inferenz in graphischen Modellen, insbesondere wenn der Graph, der die Abhängigkeiten zwischen den Zufallsvariablen ausdrückt, Zyklen enthält. Dies hat Anwendungen im Bild- und Sprachverarbeitungsbereich. (iii) Ein Biclustering Problem das über Matrizen mit kategorischen Werten definiert ist und eine natürliche Formulierung für die Analyse von Genexpressionsdaten ist. Diese Probleme variieren bezüglich deren Berechnungskomplexität. Während das Mehrkernlernen ein konvexes Problem in den Kerngewichten und in den Klassifikatorparametern ist, ist das Inferenzproblem in graphischen Modellen im generellen NP-hart. Weiter wird die NP-Härte des Biclustering Problems in dieser Arbeit hergeleitet.

Konvexe Optimierung ist ein mächtiges Werkzeug. Mit modernen Berechnungsressourcen, ist die Konvexität eines Problems ein Zertifikat für die Fügbarkeit des Problems. Es gibt aber auch viele Beispiele von Problemen und Lernannahmen und -szenarien die zu strikt nicht-konvexen Formulierungen führen. Wir zeigen, dass die Probleme die wir in dieser Arbeit betrachten, durch bekannte konvexe Funktionen approximiert werden können. Beispielsweise indem wir die Optimierungsfunktion graduell von einem konvexen linearen Programm zu einem nicht-konvexen, aber kompakten quadratischen Programm verändern, erhalten wir einen Inferenzalgorithmus der einen kompetitiven Kompromiss zwischen Genauigkeit und Laufzeit bietet.

Eine zweite Forschungsrichtung welche in dieser Dissertation verfolgt wird, beschäftigt sich mit der Beispielskomplexitätsreduktion

durch aktives Lernen. Im aktiven Lernen Szenario kann der Lernende auswählen welche Beispiele annotiert werden sollen, basierend auf nicht annotierten Beispielen und Beispielen die zuvor annotiert wurden. Eine fundamentale, aber weiterhin offene Frage ist die folgende: "Unter welchen Voraussetzungen reduziert aktives Lernen die Beispielskomplexität"? Zuvor bekannte untere Schranken von $\Omega(\frac{1}{\epsilon^2})$ implizieren, dass eine Beispielskomplexitätsreduktion im allgemeinen nicht möglich ist und weitere Annahmen nötig sind. Unter einer natürlichen Gruppierungsannahme, zeigen wir, dass es möglich ist eine substanzielle Annotierungsreduktion zu erhalten; dies unter mehreren realistischen Annahmen.

Zuletzt präsentieren wir eine komplette Studie vom Design und Testen einer maschinellen Lernlösung für das Problem der Fahrzeugdetektion basierend auf den Messungen eines Sensornetzwerks. Unsere Lösung kombiniert existierende maschinelle Lernverfahren mit neuen Methoden welche in dieser Dissertation entwickelt wurden.