Abstract

Deep learning, that is, the training of deep neural networks, is a powerful and successful approach to machine learning, allowing to fit complex models to various forms of data. On the other hand, the flexibility of neural networks also poses great challenges of both theoretical and computational nature, some of which are addressed in this thesis. In particular, we will consider the task of regression, where for a given input, a continuous output variable should be predicted.

On the theoretical side, it is still only partially understood why neural network optimization works so well, why optimized neural networks perform well on new data, or why certain neural networks work better than others. Neural networks can be studied in an under-parameterized regime, where the number of parameters is much smaller than the number of training samples, but also in an over-parameterized regime, where it is much bigger than the number of training samples. In particular, researchers have been puzzled by the so-called double descent phenomenon, where the accuracy of neural networks deteriorates as they leave the under-parameterized regime, and improves again when they become more over-parameterized. The double descent phenomenon has been proven for linear regression models under various assumptions. In Chapter 5, we prove a lower bound for the expected test error of unregularized linear regression that exhibits the typical double descent peak between the over- and under-parameterized regimes while using much weaker assumptions than previous work. We show that these assumptions are satisfied by a large class of input distributions and feature maps, and in particular for certain random deep neural network feature maps. Our results therefore also imply a lower bound with the typical double descent shape for deep neural networks where only the last layer is trained.

On the practical side, the non-linear nature and the large number of parameters of neural networks renders accurate and efficient uncertainty estimation difficult. Good uncertainty estimates can help to make a neural network more trustworthy, but they are also relevant for active learning, where the uncertainty estimates can be used to generate more relevant training data for the neural network. In Chapter 6, we study efficient batch mode active learning methods for regression with neural networks and propose a framework in which such methods can be constructed by selecting a base kernel, applying kernel transformations to the base kernel, and then using the resulting kernel in a kernel-based selection method. We provide new components in our framework, allowing to leverage linearizations of the full network efficiently through sketching and facilitating efficient distribution-aware data selection through a novel clustering method. Our accompanying open source code implements our framework in an efficient and flexible manner. We also provide a benchmark consisting of 15 tabular data sets on which our new components achieve state-of-the-art performance.

While our work on active learning uses simple uncertainty estimation methods for efficiency, it is interesting in some cases to obtain more accurate Bayesian uncertainty estimates using sampling methods. For neural networks, this requires the use of sampling algorithms for non-log-concave Gibbs distributions, for which only weak guarantees have been known. In Chapter 7, we conduct an extensive study of convergence rates of such algorithms depending on the smoothness of the log-density and the magnitude of its derivatives. First, we study the information-based complexity of the problem, showing that algorithms only constrained by the number of function evaluations can achieve the same convergence rates as for function approximation, and in some regimes even better rates. A similar picture emerges for algorithms that estimate the log-normalization constant of an unnormalized Gibbs distribution. We then study computational reductions between these two problems and optimization, and we investigate the use of function approximations to obtain faster convergence rates. Finally, we prove convergence rate bounds on several simple algorithms as well as a recent variational approach.

Kurzfassung

Tiefes Lernen, also das Trainieren tiefer neuronaler Netze, ist ein mächtiger und erfolgreicher Ansatz für maschinelles Lernen, der es erlaubt, komplexe Modelle an verschiedene Formen von Daten anzupassen. Allerdings stellt die Flexibilität neuronaler Netze aber auch große Herausforderungen sowohl theoretischer als auch rechentechnischer Art, von denen einige in dieser Dissertation adressiert werden. Insbesondere werden wir das Regressionsproblem betrachten, bei dem für eine gegebene Eingabe eine kontinuierliche Ausgabevariable vorhergesagt werden soll.

Auf der theoretischen Seite ist es immer noch nur teilweise verstanden, warum die Optimierung neuronaler Netze so gut funktioniert, warum optimierte neuronale Netze sich auf neuen Daten gut verhalten oder warum gewisse neuronale Netze besser funktionieren als andere. Neuronale Netze können in einem unterparametrisierten Regime untersucht werden, in dem ihre Parameterzahl deutlich kleiner ist als die Anzahl der Trainingsdaten, aber auch in einem überparametrisierten Regime, in dem die Parameterzahl deutlich größer ist als die Anzahl der Trainingsdaten. Insbesondere haben Forscher über das sogenannte Double-Descent-Phänomen gerätselt, bei dem die Genauigkeit von neuronalen Netzen schlechter wird, wenn sie das unterparametrisierte Regime verlassen, und wieder besser wird, wenn sie stärker überparametrisiert werden. Das Double-Descent-Phänomen wurde für lineare Regressionsmodelle unter verschiedenen Annahmen bewiesen. In Kapitel 5 beweisen wir eine untere Schranke an den erwarteten Testfehler unregularisierter linearer Regression, welcher den typischen Double-Descent-Hügel zwischen den über- und unterparametrisierten Regimes aufweist, aber dabei deutlich schwächere Annahmen macht als vorangegangene Arbeiten. Wir zeigen, dass diese Annahmen von einer großen Klasse an Eingabeverteilungen und *feature maps* erfüllt sind, und insbesondere auch von solchen *feature maps*, die durch gewisse zufällige tiefe neuronalen Netze gegeben sind. Unsere Resultate implizieren daher auch eine untere Schranke mit der typischen Double-Descent-Form für tiefe neuronale Netze, bei denen nur die letzte Schicht trainiert wird.

Auf der praktischen Seite macht die nichtlineare Natur und die große Parameterzahl neuronaler Netze eine genaue und effiziente Unsicherheitsschätzung schwierig. Eine gute Unsicherheitsschätzung kann helfen, ein neuronales Netz vertrauenswürdiger zu machen, aber sie ist auch relevant für aktives Lernen, bei dem die Unsicherheitsschätzung genutzt werden kann, um relevantere Trainingdaten für das neuronale Netz zu generieren. In Kapitel 6 untersuchen wir effiziente aktive Lernmethoden für Regression mit neuronalen Netzen und führen ein Schema ein, in dem solche Methoden konstruiert werden können durch die Wahl eines Basiskerns, die Anwendung von Kerntransformationen auf den Basiskern und die Nutzung des resultierenden Kerns in einer kernbasierten Auswahlmethode. Wir stellen neue Komponenten für unser Schema bereit, die es erlauben, eine Linearisierung

des vollen Netzes auf effiziente Art durch *Sketching* zu nutzen, und eine effiziente und verteilungsbewusste Datenauswahl durch eine neue Clustering-Methode zu treffen. Unser begleitender öffentlicher Programmcode implementiert unser Schema auf effiziente und flexible Art. Wir stellen darüber hinaus einen Benchmark bereit, der aus 15 tabellarischen Datensätzen besteht, auf denen unsere neuen Komponenten eine neue Bestmarke in der Qualität der resultierenden Netze erreichen.

Während unsere Arbeit zu aktivem Lernen aus Effizienzgründen einfache Unsicherheitsschätzungsmethoden verwendet, ist es in einigen Fällen interessant, genauere Bayesianische Unsicherheitsschätzungen mit Sampling-Methoden zu erhalten. Für neuronale Netze erfordert dies die Nutzung von Sampling-Algorithmen für nicht log-konkave Gibbs-Verteilungen, für welche nur schwache Garantien bekannt waren. In Kapitel 7 führen wir eine umfangreiche Untersuchung der Konvergenzraten solcher Algorithmen durch, abhängig von der Glattheit der log-Dichte und der Größe ihrer Ableitungen. Zuerst untersuchen wir die informationsbasierte Komplexität des Problems und zeigen dabei, dass Algorithmen, die nur durch die Anzahl der Funktionsauswertungen beschränkt sind, die gleichen Konvergenzraten wie Funktionsapproximation erreichen können, und in manchen Regimes sogar bessere Raten. Ein ähnliches Bild ergibt sich für Algorithmen, die die log-Normalisierungskonstante einer unnormalisierten Gibbs-Verteilung schätzen. Danach untersuchen wir Reduktionen zwischen diesen zwei Problemen und Optimierung und wir untersuchen die Nutzung von Funktionsapproximationen, um schnellere Konvergenzraten zu bekommen. Schließlich beweisen wir Schranken an die Konvergenzraten verschiedener einfacher Algorithmen sowie eines kürzlich publizierten variationellen Ansatzes.