Kurzfassung

Lernraten nicht-parametrischer Lernmethoden weisen in der Regel eine unvorteilhafte Abhängigkeit von der Dimension des Eingaberaums auf, ein Phänomen, das in der statistischen Lerntheorie allgemein als Fluch der Dimensionalität bekannt ist. In der Praxis geht man jedoch davon aus, dass hochdimensionale Daten in der Regel eine niedrigdimensionale intrinsische Struktur aufweisen. Daher ist es eine interessante Frage, ob Lernmethoden eine solche niedrigdimensionale Struktur in dem Sinne ausnutzen können, dass sie Lernraten erzielen, bei denen die Abhängigkeit von der Dimension des umgebenden Raums durch die intrinsische Dimension der Daten ersetzt wird. Die verbreitetste Methode um die intrinsische Dimension der Daten zu beschreiben beruht auf der Annahme, dass die datenerzeugende Verteilung auf einer niedrigdimensionalen glatten Mannigfaltigkeit getragen ist. Wir schwächen diese Annahme erheblich ab, indem wir die fraktale Dimension des Trägers der datenerzeugenden Verteilung betrachten.

Genauer leiten wir in Kapitel 2 für Support Vector Machines (SVMs), die einen Gauß-Kern verwenden, für die Kleinste-Quadrate-Regression und die binäre Klassifikation unter Verwendung der Hinge-Verlustfunktion Lernraten her, in denen die Abhängigkeit von der Dimension des umgebenden Raums durch die Box-Counting-Dimension des Trägers der datenerzeugenden Verteilung ersetzt wird. Die sich daraus ergebenden Lernraten für die Regression sind bis auf logarithmischen Faktoren minimax-optimal und die Lernraten für die Klassifikation sind bis auf logarithmischen Faktoren in einem bestimmten Bereich unserer Annahme minimax-optimal und haben ansonsten die Form der besten bekannten Raten. Wir zeigen außerdem, dass diese Raten adaptiv durch einen Trainingsvalidierungsansatz für die Auswahl der Hyperparameter erreicht werden können. Genauer gesagt zeigen wir, dass es zum Erreichen optimaler Raten auf adaptive Weise ausreicht, wenn die Anzahl der Kandidatenwerte für die Hyperparameter logarithmisch mit dem Stichprobenumfang wächst, während existierende vergleichbare Resultate erfordern, dass die Anzahl der Kandidatenwerte mindestens linear mit dem Stichprobenumfang wächst.

In Kapitel 3 beweisen wir ähnliche Ergebnisse für eine räumlich lokalisierte Variante von Gaußschen-SVMs, die eine verbreitete Methode zur Minderung des Rechenaufwands gewöhnlicher SVMs sind, welche quadratisch im Platz- und mindestens quadratisch Zeitbedarf sind. Bei diesem lokalisierten SVM-Ansatz wird eine Partition des Eingaberaums berechnet, und dann für jede Zelle dieser Partition eine SVM-Entscheidungsfunktion berechnet, wobei nur die in der jeweiligen Zelle enthaltenen Stichproben verwendet werden. Während existierende Ergebnisse zu Verfahren, die mit unserem vergleichbar sind, eine a-priori festgelegte Partition des Eingaberaums betrachten, die einige technische Annahmen erfüllt, betrachten wir eine vollständig datenabhängige Partition, die auf dem Farthest-First-Traversal-Algorithmus basiert. In diesem Kapitel hängt unser Begriff der intrinsischen Dimension von der Assouad-Dimension des Trägers der datenerzeugenden Verteilung ab. Wir beweisen erneut die gleichen minimax-optimalen Raten unter diesem etwas stärkeren Begriff der intrinsischen Dimensionalität. Wir beweisen ebenfalls erneut, dass diese Raten adaptiv durch ein Trainingsvalidierungsverfahren mit logarithmisch wachsenden Kandidatenmengen erreicht werden können. Die Ergebnisse dieses Kapitels sind die ersten, die die Adaptivität an die intrinsische Dimensionalität der Daten für eine Beschleunigungsstrategie für Kern-Methoden berücksichtigen.

In Kapitel 4 ergänzen wir schließlich unsere theoretischen Erkenntnisse aus Kapitel 2 und 3 durch experimentelle Untersuchungen. Dazu betten wir gegebene Datensätze mittels einer nicht-trivialen randomisierten Einbettung in einen höherdimensionalen Raum ein und vergleichen, wie sich der Testfehler in Abhängigkeit von der Anzahl der künstlich hinzugefügten Dimensionen verhält. Wir führen dieses Verfahren für eine Reihe von Regressions- und Klassifikationsdatensätzen für die in Kapitel 2 und 3 betrachteten Lernmethoden durch. Die Ergebnisse unserer Experimente deuten darauf hin, dass die intrinsische Dimension der Daten tatsächlich die entscheidende Größe ist, die die Generalisierungsleistung bestimmt, anstatt der Dimension des umgebenden Raums.

Summary

Learning rates for non-parametric learning methods usually exhibit a poor dependency on the dimension of the input space, a phenomenon commonly known as the curse of dimensionality in statistical learning theory. In practice however, high dimensional data usually is hypothesized to have some low dimensional intrinsic structure and therefore it is an interesting question whether learning methods can exploit such a low dimensional structure in the sense that they achieve learning rates where the dependence of the dimension of the ambient space is replaced with the intrinsic dimension of the data. The most common notion of intrinsic dimension of data relies on the assumption that the data generating distribution is supported on a low-dimensional smooth manifold. We substantially weaken this assumption by considering the fractal dimension of the support of the data generating distribution.

More precisely, in Chapter 2 we derive learning rates for support vector machines (SVMs) using a Gaussian kernel for least-squares regression and binary classification using the hinge loss, where the dependence of the dimension of the ambient space is replaced with the box-counting dimension of the support of the data generating distribution. The resulting learning rates for regression are minimax optimal up to logarithmic factors and the learning rates for classification are minimax optimal up to logarithmic factors in a certain range of our assumption and otherwise of the form of the best known rates. We further show that these rates can be achieved adaptively by a training validation approach for hyperparameter selection. More specifically, we show that in order to achieve optimal rates adaptively it is sufficient for the size of the candidate sets of values for the hyperparameters to grow logarithmically in the sample size, whereas existing similar results require the size of candidate sets to grow at least linearly in the sample size.

In Chapter 3 we prove similar results for a spatially localized version of Gaussian SVMs, which is a popular method for circumventing the computational costs of ordinary SVMs, which are quadratic in space and at least quadratic in time. In this localized SVM approach a partition of the input space is computed and then

for each cell of that partition an SVM decision function is computed using only the samples contained in that respective cell. Whereas existing results similar to ours consider an a-priori fixed partition of the input space satisfying some technical assumptions, we consider a fully data dependent partitioning based on the farthest first traversal algorithm. In this chapter our notion of intrinsic dimension depends on the Assouad dimension of the support of the data generating distribution. We again prove the same minimax optimal rates under this slightly stronger notion of intrinsic dimensionality. We also again prove that these rates can be achieved adaptively using a training validation procedure using logarithmically growing candidate sets. The results of this chapter are the first to consider adaptivity to the intrinsic dimension of the data for a speed-up strategy for kernel methods.

Finally, in Chapter 4 we complement our theoretical findings of Chapter 2 and 3 by experimental investigation. To this end, we embed a given dataset into a higher dimensional space via a non-trivial randomized embedding and compare how the test error changes depending on the number of artificially added dimensions. We perform this procedure for a number of regression and classification datasets using the learning methods considered in Chapter 2 and 3. The results of our experiments suggest that, in fact, the intrinsic dimension of the data is the critical quantity determining the generalization performance, as opposed to the dimension of the ambient space.