Optimierung von Regularisierungsparametern im maschinellen Lernen

Zusammen mit Fachkollegen von der FSU Jena ist es Prof. Dr. Christopher Schneider vom Fachbereich Grundlagenwissenschaften dieses Jahr gelungen, seine Forschungsergebnisse im Bereich des maschinellen Lernens bei zwei der international renommiertesten Konferenzen auf dem Gebiet der künstlichen Intelligenz zu publizieren. Über die Inhalte der beiden Arbeiten berichtet der Wissenschaftler im Folgenden.

Maschinelles Lernen beschäftigt sich mit der Generierung von *Wissen* aus *Erfahrung* und stellt eine Schlüsseltechnologie für moderne Methoden der *künstlichen Intelligenz* dar. Mit Hilfe komplexer statistischer Modelle sollen Ergebnisse, die bei Trainingsdaten gelernt wurden, auf unbekannte neue Daten verallgemeinert werden. Es entsteht ein selbstlernendes Computerprogramm. Interessant und herausfordernd sind sowohl der Fall, wenn mit sehr großen Datenmengen gearbeitet wird, als auch der Fall, wenn nur wenige Daten zur Verfügung stehen.

Seit der Jahrtausendwende kam es durch die gesteigerte Leitstungsfähigkeit der Computer und das zunehmende Aufkommen von *Big Data* zu großen Fortschritten und Erfolgen der Techniken des maschinellen Lernens. Anwendung finden diese Methoden heute u. a. bei Recommender Systemen (Produktempfehlungen bei Amazon, Filmvorschläge auf Netflix), Bildanalysen (Gesichtserkennung bei Facebook), der automatischen Spracherkennung („Alexa, ...“, „Hey Siri, ...“, „Ok Google, ...“), Klassifikationsaufgaben (Spamerkennung in E-Mails, Therapievorschläge bei Krebsbehandlungen) oder auch bei der Entwicklung autonomer Systeme (selbstfahrende Autos).

Die passenden Modelle für viele der eben genannten Anwendungen lassen sich durch mathematische Optimierungsprobleme der Form

$$(P_\alpha) \qquad \min_{x \in \mathcal{F}} \ \ell(x) + \alpha\, r(x)$$

beschreiben. Hierbei sind x die Variablen des Optimierungsproblems bzw. die Parameter des zu beschreibenden Modells. Die Variablen sind Vektoren $x \in \mathbb{R}^n$ typischerweise hoher Dimension $n \in \mathbb{N}$. Die Menge $\mathcal{F}$ ist die Menge aller zulässigen Punkte. Diese Menge kann durch Nebenbedingungen eingeschränkt werden. Die Funktion ℓ heißt *Loss-Funktion* und ist die eigentlich zu minimierende Zielfunktion des Problems. Sie hängt von den *Problemdaten* ab. Die Funktion r dient als *Regularisierung*. Eine oder auch mehrere solcher Funktionen können bestimmte Lösungsstrukturen erzeugen und die Überanpassung an die Trainingsdaten (*Overfitting*) verhindern. Die Gewichtung dieses zweiten Zielkriteriums erfolgt durch den sogenannten *Regularisierungsparameter* $\alpha \geq 0$. Die Wahl von α ist dabei sehr wichtig für den Erhalt eines brauchbaren Modells mit hoher Vorhersagegüte und Gegenstand der beiden Arbeiten [1, 2].

Als Beispiel betrachten wir im Folgenden die Klassifikation von Microarray- und Genexpressionsdaten – eine wichtige Anwendung maschinellen Lernens in der Krebsforschung. Ausgangspunkt seien hierzu eine Genexpressions-Matrix $A \in \mathbb{R}^{m \times n}$, welche die Ausprägungen von n Genen für m Proben von Patienten beinhaltet, sowie eine Ergebnisgröße $b \in \mathbb{R}^m$, in der zugehörige Krebstypen der Patienten festgehalten sind. Durch das Erkennen von Strukturen in der Datenmatrix sollen zukünftige Krebsbefunde vorhergesagt und Therapievorschläge erzeugt werden. Neben einer hohen Vorhersagegüte ist man aber auch daran interessiert, die wichtigsten Gene für den Status der Erkrankung zu identifizieren. Zur Modellierung des Problems nutzen wir das sog. *Elastic Net* – eine regularisierte Variante der linearen Regression:

$$(EN) \qquad \min_{x \in \mathbb{R}^n} \ \|Ax - b\|_2^2 + \alpha \|x\|_2^2 + \beta \|x\|_1 \ .$$

Der erste Teil des Optimierungsproblems entspricht der Methode kleinster Quadrate und ist dafür verantwortlich x so zu wählen, dass der Klassifikationsfehler minimal wird. Hinzu kommen zwei gewichtete Regularisierungsterme für die zu bestimmenden Variablen $x \in \mathbb{R}^n$, wobei

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \quad (L^1\text{-Norm}) \quad \text{und} \quad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad (L^2\text{-Norm}).$$

Hierbei sorgt die L^1-Norm dafür, dass möglichst viele Einträge von x auf Null gesetzt werden, also nur wenige Gene selektiert werden. Man nennt diese Eigenschaft *Sparsity*. Gesteuert wird diese über die Wahl des Regularisierungsparameters $\beta \geq 0$. Durch die quadrierte L^2-Norm entsteht ein *Gruppierungseffekt*, der korrelierte Gene entweder zusammen im Modell behält oder entfernt (also auf Null setzt). Die Stärke dieses Effekts wird durch die Wahl des Regularisierungsparameters $\alpha \geq 0$ bestimmt.

Abbildung 1 (links) veranschaulicht die beiden Regularisierungen anhand eines kleinen Datensatzes (8 untersuchte Eigenschaften, 97 Proben) aus einer Studie zu Prostatakrebs. Abgebildet ist für ein fest gewähltes $\alpha \geq 0$ ein sog. *Regularisierungspfad* der Variablen $x \in \mathbb{R}^8$ für alle $\beta \geq 0$. Dieser zeigt die Entwicklung von x in Abhängigkeit zum Parameter β. Gut zu erkennen ist, dass sowohl die Variablen `svi` und `gleason`, als auch die Variablen `pgg45` und `lbph` Gruppen bilden. Im optimalen Modell (gekennzeichnet durch die gestrichelte Linie $\beta \approx 1.9$), taucht neben diesen beiden Gruppen noch `lcavol` als relevante Variable auf. Die anderen drei Variablen (`lcp`, `age`, `lcp`) werden auf Null gesetzt. Für die Berechnung dieses stückweise linearen Pfades können aus der Literatur bekannte Algorithmen verwendet werden (*least-angle regression*, *Ensalg*). Abbildung 1 (rechts) zeigt den Regularisierungspfad für einen größeren Datensatz mit Genexpressions-Daten aus Lungenkrebsstu-

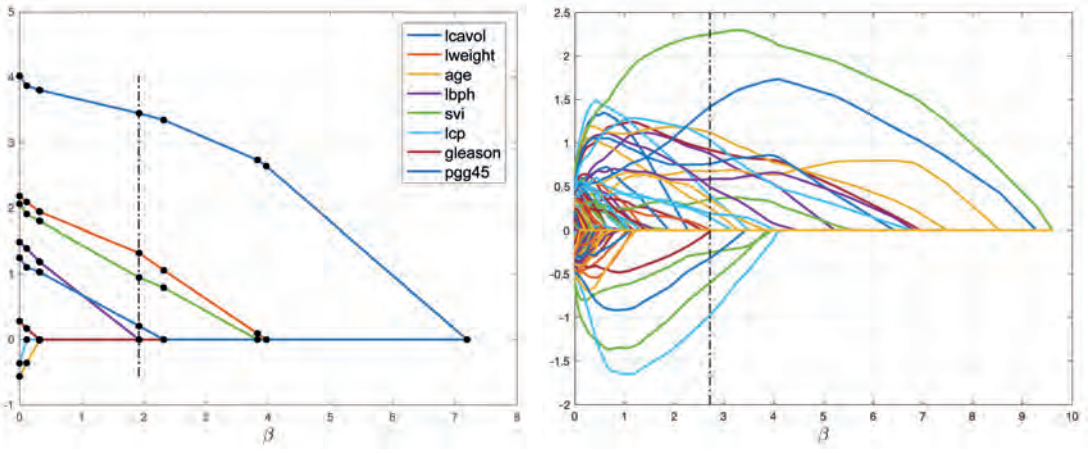

Abbildung 1: Regularisierungspfade für die Daten zu Prostatakrebs (links) und Lungenkrebs (rechts).

dien. Hier wurde die Ausprägung von $n = 325$ Genen bei $m = 73$ Patienten untersucht. Das optimale Modell (gestrichelte Linie $\beta \approx 2.7$) selektiert hier lediglich 21 der 325 Gene als relevant. Dies entspricht einer *Sparsity* von mehr als 93 %.

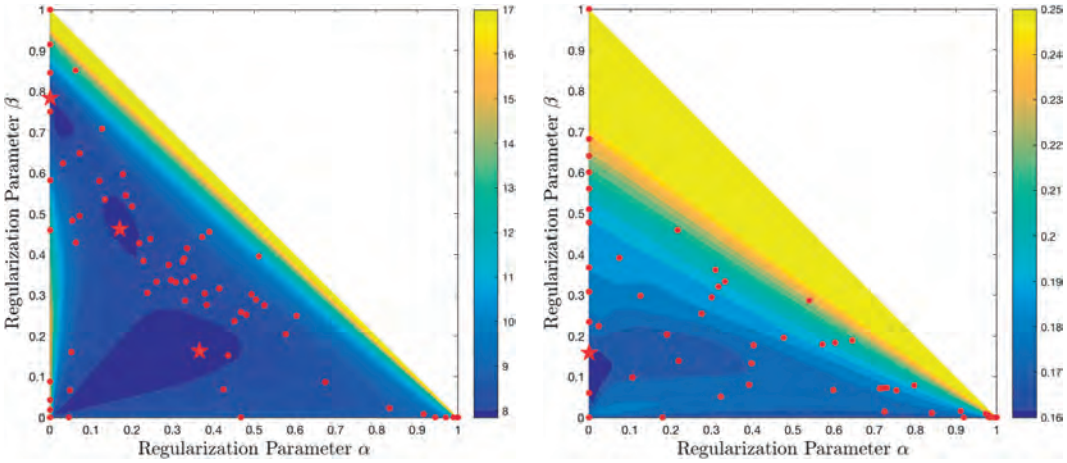

Abbildung 2: Regularisierungsparametersuche mit Hilfe des Benson-Algorithmus für die Daten zu 14 verschiedenen Krebsarten (links) und Lungenkrebs bei Rauchern (rechts).

Der Nachteil der eben gezeigten Pfade ist, dass der Wert des Parameters α im Vorhinein fixiert werden muss. Es ist anschließend lediglich möglich, bezüglich β eine optimale Wahl zu treffen. Die in [1] vorgestellte Idee ermöglicht es, beide Regularisierungsparameter gleichzeitig zu optimieren. Hierzu wurde eine Methode aus der Vektoroptimierung, der sog. *Benson-Algorithmus*, eingesetzt. In Abbildung 2 sind die Ergebnisse dieses Verfahrens für zwei Genexpressions-Datensätze aus Studien zu 14 verschiedenen Krebsarten (links, $n \approx 16\,000$, $m \approx 200$) und Lungenkrebs bei Rauchern (rechts, $n \approx 20\,000$, $m \approx 190$) zu sehen. Dunkelblaue Regionen stellen hierbei gute Parameterkombinationen dar, d. h., das resultierende Modell besitzt eine gute Vorhersagegüte. Die roten Punkte visualisieren den Iterationsverlauf des Benson-Algorithmus, wobei die roten Sterne die besten gefundenen Parameterpaare kennzeichnen. In Abbildung 2 (links) ist zu erkennen, dass die Regionen guter Parameterkombinationen nicht immer miteinander verbunden sind. Das Problem ist *nicht konvex* und damit mathematisch äußerst anspruchsvoll. Trotzdem gelingt es dem verwendeten Verfahren geeignete Vertreter aus allen drei Regionen zu finden.

Ein großer Vorteil des Benson-Algorithmus ist, dass er zur Parameterbestimmung aller regularisierten Optimierungsprobleme der Form (P_α) anwendbar ist und nicht auf das Elastic Net (EN) beschränkt ist. So wurde die Methode in [2] auch verwendet, um grafische Modelle effzient zu lernen. Eine Besonderheit hierbei ist, dass die Variablen x nun selbst schon Matrizen sind, was zu sehr hochdimensionalen Problemen führt. Weiterhin konnte anhand der untersuchten Datensätze gezeigt werden, dass die Komplexität des neu entwickelten Verfahrens optimal ist.

Prof. Dr. Christopher Schneider, FB GW

Christopher Schneider stellte die Ergebnisse aus [1] im Februar auf der „33rd AAAI Conference on Artificial Intelligence" in Honolulu (Hawaii, USA) und die Ergebnisse aus [2] im August auf der „28th International Joint Conference on Artificial Intelligence" in Macau (China) vor.

Literatur

[1] Giesen, Joachim; Laue, Sören; Löhne, Andreas; Schneider, Christopher: Using Benson's Algorithm for Regularization Parameter Tracking. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019, S. 3689–3696

[2] Giesen, Joachim; Nussbaum, Frank; Schneider, Christopher: Effcient Regularization Parameter Selection for Latent Variable Graphical Models via Bi-Level Optimization. In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, S. 2378–2384

BETRIEBSWIRTSCHAFT

Drei Tage Berlin

Es war mal wieder „Exkursions-Time" im Fachbereich BW. Auf 18 Teilnehmer aller Semesterstufen wartete ein prallvolles Programm mit sieben Stationen.

Durch einen langen Vortragstag am Samstag vorher bestens vorbereitet, standen am Morgen des 22. Mai um 6.00 Uhr alle Studierenden bereit zur Abfahrt. Wir freuten uns auf interessante und ereignisreiche Tage in Berlin.

Unsere erste Station war die jährliche Aktionärsversammlung der Daimler AG. Es war eine besondere Hauptversammlung, denn es endete eine Management-Ära: Der Vorstandsvorsitzende, Dieter Zetsche, berichtete zum letzten Mal in seiner langen Amtszeit von 13 Jahren über das abgelaufene Geschäftsjahr und stellte sich den kritischen und spannenden Fragen der Aktionäre.

Mit Ablauf der Versammlung wurde er als Vorstandsvorsitzender von Ola Källenius, dem derzeitigen Entwicklungsvorstand, „beerbt". Ein intensiv diskutiertes Thema war die neue Organisationsstruktur des Unternehmens und das „Projekt Zukunft", bei dem insbesondere der Umweltschutz eine bedeutende Rolle spielt. Die Daimler AG will bis 2039 komplett CO_2-neutral produzieren.

Außerdem wurden Prototypen des autonomen Fahrens vorgestellt, welche wir auch teilweise in einer Ausstellung unmittelbar besichtigen konnten. Es war sehr interessant, einmal „live und in Farbe" mitzuerleben, wovon man sonst nur in den Nachrichten sieht und hört. Am späten Nachmittag fuhren wir dann zu unserem Hotel, welches – direkt am Alexanderplatz gelegen – eine perfekte Ausgangsbasis für Berlinerkundungen auf eigene Faust am Abend war.

Donnerstag starteten wir bei bestem Sonnenschein zu unserer ersten Besichtigungsstation: Das Heizkraftwerk Berlin-Mitte der Vattenfall AG. Von einem langjährigen Mitarbeiter hörten wir einen ausführlichen Vortrag zum Unternehmen und zum Thema Strom-, Wärme-, und Kälteerzeugung. Anschließend ging es, ausgestattet mit passender Schutzkleidung, auf eine ausführliche Werksführung. Es war sehr interessant, einen Einblick in die verschiedenen technisch-wirtschaftlichen Aspekte und Probleme der Stromerzeugung zu bekommen. Strom kommt eben nicht nur einfach aus der Steckdose …

Nach einer kurzen Mittagspause hatten wir noch etwas Zeit, die Umgebung rund um Reichstag und