STATISTISCHE MODELLIERUNG: DIE ZWEI KULTUREN

Seminar für Statistische Modellierung latenter Strukturen in den Lebens-, Sozial- und Wirtschaftswissenschaften

18.01.2013, 12:30 Uhr
Seminarraum, Institut für Statistik, LMU

Christoph Molnar
Betreuer: Georg Schollmeyer
Seminarleiter: Professor Augustin

2014-01-12

STATISTISCHE MODELLIERUNG:
DIE ZWEI KULTUREN

Seminar für Statistische Modellierung latenter Strukturen in den Lebens-, Sozial- und Wirtschaftswissenschaften

18.01.2013, 12:30 Uhr
Seminarraum, Institut für Statistik, LMU

Christoph Molnar
Betreuer: Georg Schollmeyer
Seminarleiter: Professor Augustin

Abstract:
In der Präsentation werden zwei Kulturen der Modellierung verglichen: die Datenmodellierung, bei der ein datengenerierender stochastischer Prozess angenommen wird und die mit traditioneller Statistik assoziiert ist; die algorithmische Modellierung, die auf Optimierung einer Funktion mit Hilfe eines Algorithmus reduziert werden kann und mit Machine Learning assoziiert wird.
Es wird für stärkere Verwendung von algorithmischer Modellierung in der Statistik argumentiert.

GLIEDERUNG

1. Statistik: Kultur der Datenmodellierung
2. Machine Learning: Kultur der algorithmischen Modellierung
3. Prinzipien beim Lernen aus Daten
4. Persönliche Erfahrungen
5. Fazit

Inhalt basiert auf: ``Statistical Modeling: The two cultures'' von Leo Breiman [1]

GLIEDERUNG

1. Statistik: Kultur der Datenmodellierung
2. Machine Learning: Kultur der algorithmischen Modellierung
3. Prinzipien beim Lernen aus Daten
4. Persönliche Erfahrungen
5. Fazit

Inhalt basiert auf: ``Statistical Modeling: The two cultures" von Leo Breiman [1]

Die Präsentation basiert auf dem Artikel ``Statistical Modeling: The two cultures" von Leo Breiman [1].
Im ersten Abschnitt wird die Modellierung in der Statistik betrachtet und analog dazu im zweiten Abschnitt die Modellierung im Machine Learning inklusive kurzer Vorstellung einiger Algorithmen. Das Kapitel über Prinzipien beim Lernen aus Daten stellt die beiden Kulturen gegenüber. Persönliche Erfahrungen illustrieren Beispiele für die beiden Kulturen. Im Fazit wird die Botschaft des Artikels [1] zusammengefasst.

STATISTIK: KULTUR DER DATENMODELLIERUNG

ARBEIT EINES STATISTIKERS

→ **Prognosen treffen**

→ **Zusammenhänge erklären**

→ Versuchsplanung, Parameterschätzung, Visualisierung, ...

- → **Prognosen treffen**
- → **Zusammenhänge erklären**
- → Versuchsplanung, Parameterschätzung, Visualisierung, ...

Die Arbeit eines Statistikers ist sehr vielfältig. Hier wird nur über die Modellierung der Daten gesprochen, die zwei Ziele hat: Prognose für neue Beobachtungen und Erklärung von Zusammenhängen.

VEREINFACHTES WELTBILD

Vereinfacht kann die Natur als ein Mechanismus in einer Box betrachtet werden, der aus Kovariablen die Zielgröße generiert. Die Kenntnisse über den Mechanismus der Natur können von vollkommen unbekannt bis hin zu etablierten substanzwissenschaftlichen Theorien reichen. Ein Beispiel ist der Mietspiegel: Hier ist die Zielgröße der Mietpreis und Kovariablen sind Größe und Alter der Wohnung, Art der Heizung, ...

WELTBILD DES STATISTIKERS

Suche stochastisches Modell der Daten:
Zielgröße = f(Kovariablen, Parameter, zufälliger Fehler)

WELTBILD DES STATISTIKERS

Suche stochastisches Modell der Daten:
Zielgröße = f(Kovariablen, Parameter, zufälliger Fehler)

Das direkte Modellieren des Mechanismus in der ``Box" wird von Breiman als »Data Modeling Culture« bezeichnet. In dieser Kultur wird für den datengenerierenden Prozess ein stochastisches Modell angenommen, dessen Parameter geschätzt werden können. Die meisten Modelle sind so formuliert, dass die Zielgröße eine Funktion der Kovariablen mit dazugehörigen Parametern und einem Fehlerterm ist.

TYPISCHE ANNAHMEN UND RESTRIKTIONEN

→ Stochastisches Modell, dass Daten generiert
→ Bestimmte Verteilung der Residuen
→ Linearitäten (z.B. Linearer Prädiktor)
→ Interaktionen müssen manuell spezifiert werden

PROBLEME

→ Schlussfolgerungen über Modellmechanismen, nicht über Natur

→ Annahmen häufig verletzt

→ Häufig keine Modellevaluierung

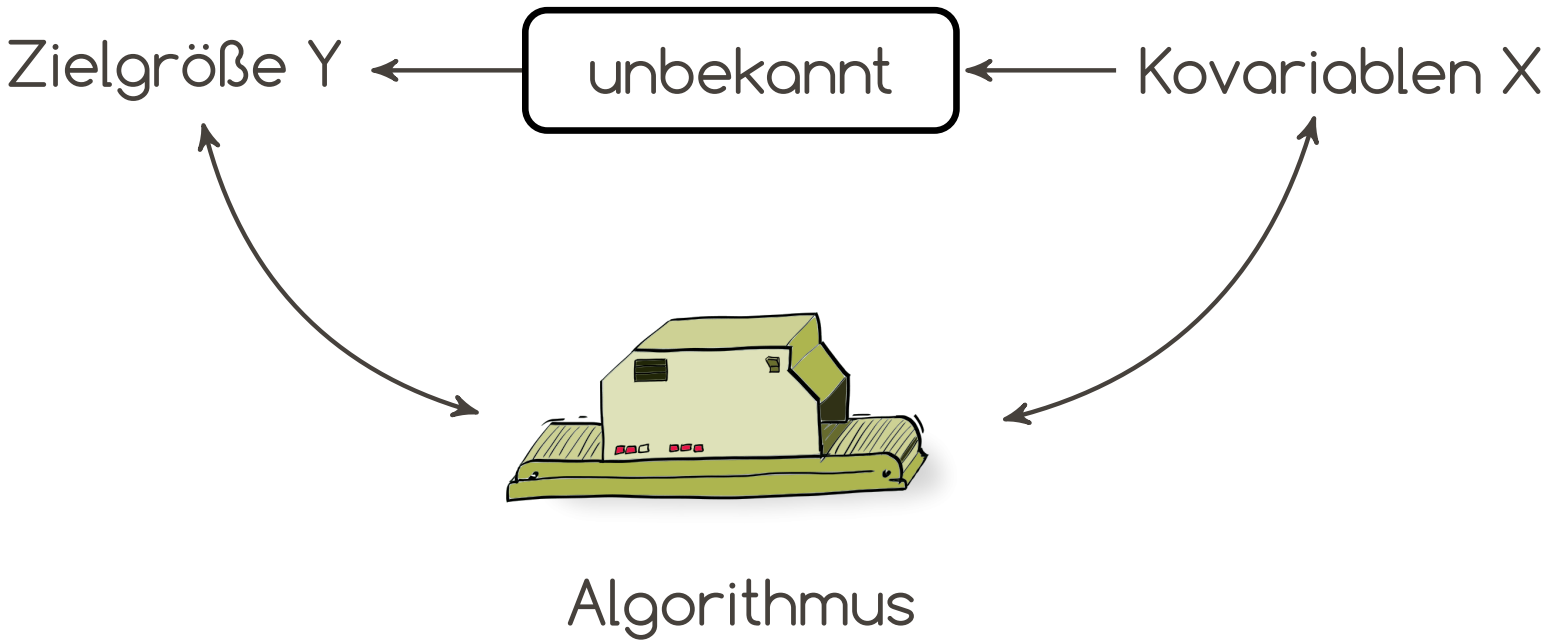
→ $\Rightarrow$ führt zu irrelevanter Theorie und fragwürdigen statistischen Schlussfolgerungen

→ Fokus nicht auf Prognosekraft

→ Datenmodelle ungenügend in Gebieten wie Bilderkennung, Spracherkennung, ...

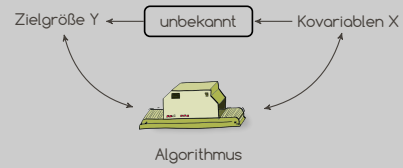
MACHINE LEARNING: KULTUR DER ALGORITHMISCHEN MODELLIERUNG

MACHINE LEARNING



Suche Funktion $f(X)$ die den Verlust $L(Y, f(X))$ minimiert

MACHINE LEARNING



Suche Funktion $f(X)$ die den Verlust $L(Y, f(X))$ minimiert

Bei dem von Breiman als »Algorithmic Modeling Culture« bezeichneten Vorgehen sieht man den tatsächlichen Mechanismus als unbekannt an. Man versucht nicht den datengenerierenden Prozess zu finden, sondern benutzt einen Algorithmus um den Mechanismus der Natur zu imitieren. Die Modellierung reduziert sich zu einem mathematischen Optimierungsproblem: Gegeben Kovariablen, Zielgröße und Verlustfunktion, suche ein Funktion f(X), die den Verlust bei der Vorhersage der Zielgröße minimiert. Diese Kultur findet man im Bereich des Machine Learning.

Zusammenfassung: Datenmodellierung versucht den wahren Mechanismus zu finden, algorithmische Modellierung versucht den wahren Mechanismus möglichst gut zu imitieren.

ALGORITHMEN IM MACHINE LEARNING

- → Boosting
- → Support Vector Machines
- → Künstliche Neuronale Netze
- → Random Forests
- → Hidden Markov
- → Bayes-Netze
- → ... [1]

[1] Details und mehr Algorithmen in ``Elements of statistical learning''[2]

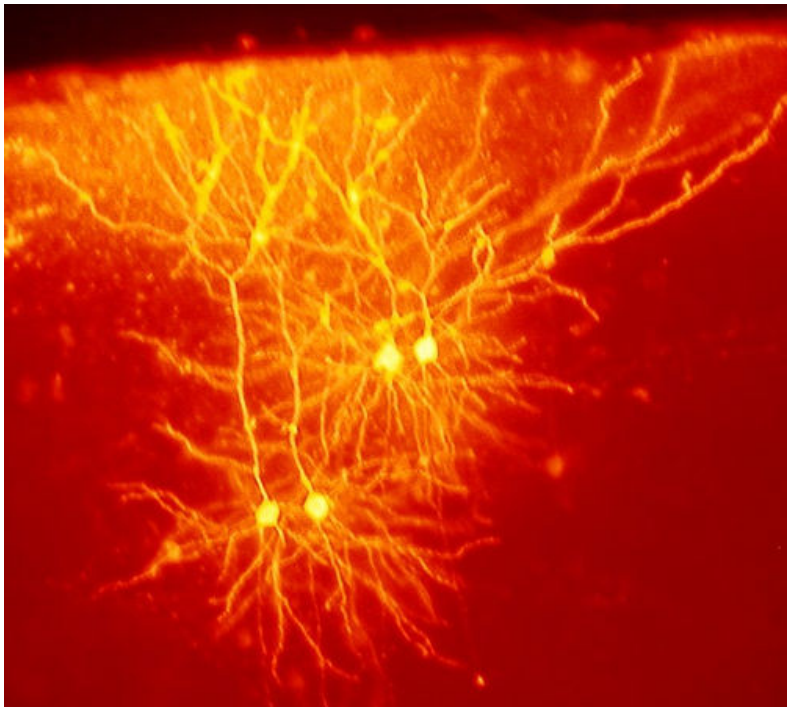
ALGORITHMEN IM MACHINE LEARNING

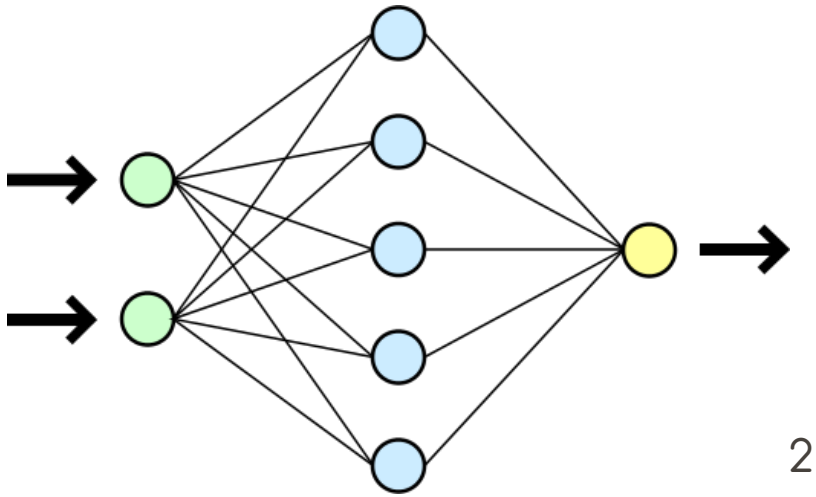
- → Boosting
- → Support Vector Machines
- → Künstliche Neuronale Netze
- → Random Forests
- → Hidden Markov
- → Bayes-Netze
- → ... [1]

[1] Details und mehr Algorithmen in "Elements of statistical learning"[2]

Die Algorithmen im Machine Learning unterscheiden sich von den Ideen her sehr stark. Drei Algorithmen werden im Folgenden kurz vorgestellt.

KÜNSTLICHE NEURONALE NETZE

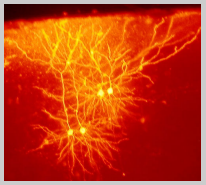


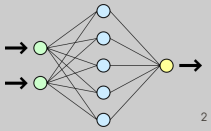


2

[2]http://commons.wikimedia.org/wiki/File:Mouse_cingulate_cortex_neurons.jpg
http://commons.wikimedia.org/wiki/File:Neural_network.svg

KÜNSTLICHE NEURONALE NETZE



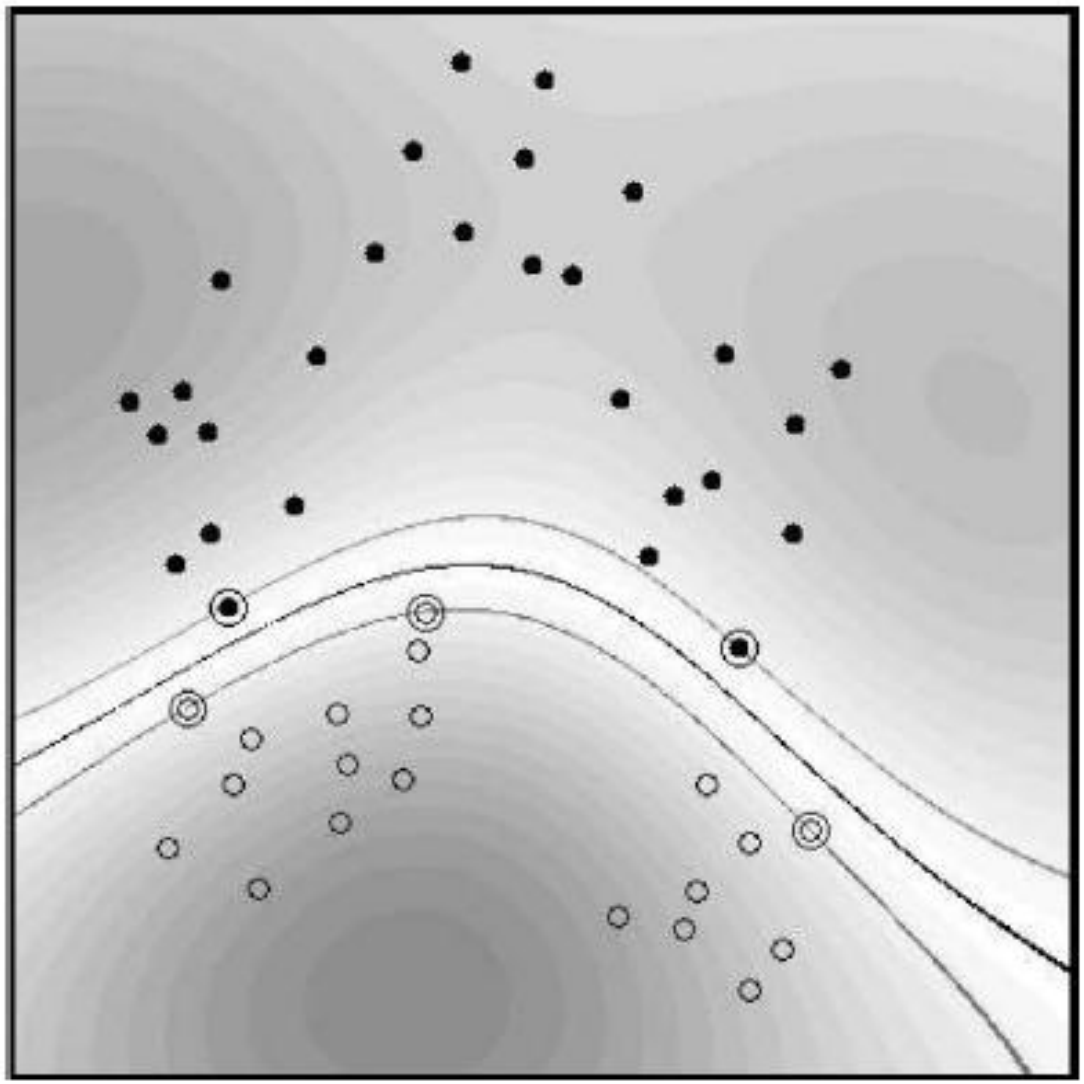


2

[2] http://commons.wikimedia.org/wiki/File:Mouse_cingulate_cortex_neurons.jpg
http://commons.wikimedia.org/wiki/File:Neural_network.svg

Künstliche neuronale Netze können für Klassifikation und Regression benutzt werden. Sie sind inspiriert durch das Gehirn, das aus Netzwerken von Gehirnzellen (Neuronen) besteht. Mathematisch sind künstliche neuronale Netze Verkettungen von gewichteten Funktionen. Typisches Anwendungsgebiet ist die Bildverarbeitung.

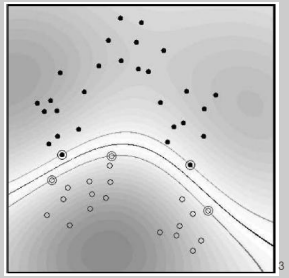
SUPPORT VECTOR MACHINES



3

[3] http://commons.wikimedia.org/wiki/File:Svm_10_perceptron.JPG

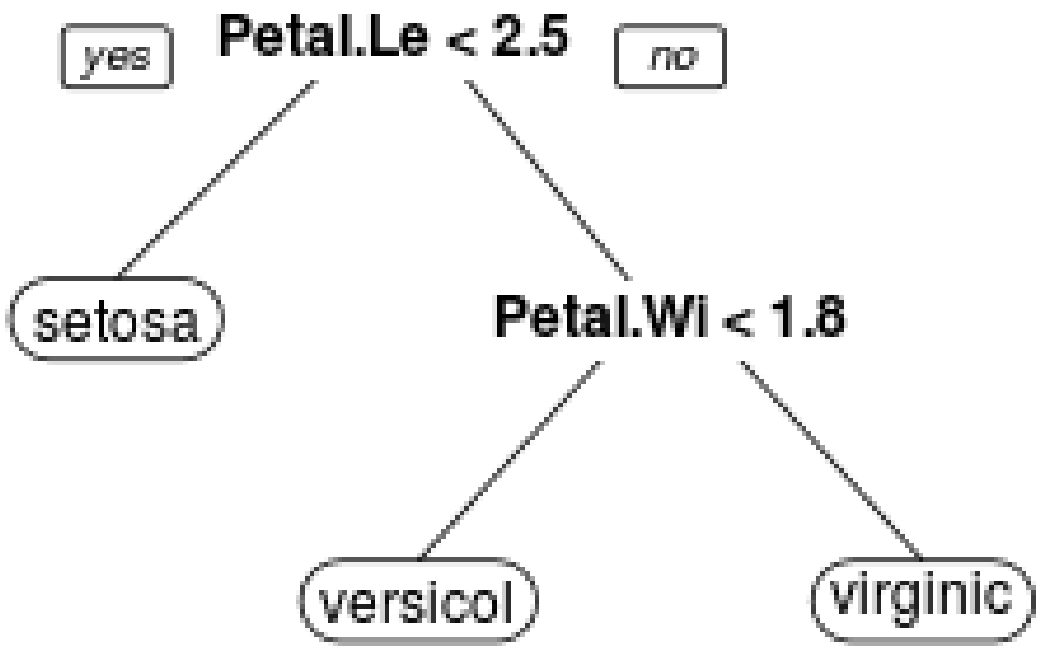
SUPPORT VECTOR MACHINES



[3] http://commons.wikimedia.org/wiki/File:Svm_10_perceptron.JPG

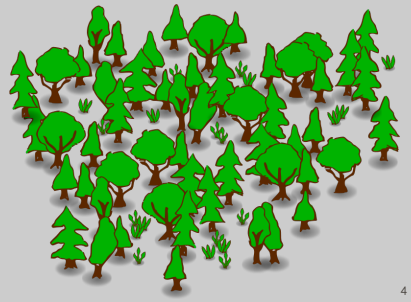
Eine Support Vector Machine (SVM) ist ursprünglich ein Klassifikationsverfahren (Regression auch möglich). Sie funktioniert so, dass sie versucht im Raum der Kovariablen eine Klassengrenze zu ziehen, wobei der Abstand der Grenze zu den Beobachtungen maximiert wird. SVMs benutzen einen mathematischen Trick um implizit die Kovariablen in einen höherdimensionalen Raum abzubilden und so Trennbarkeit der Klassen zu erreichen. Typisches Anwendungsgebiet: Klassifikation von Text

RANDOM FORESTS



RANDOM FORESTS

RANDOM FORESTS



[4] http://openclipart.org/detail/175304/forest-by-z-175304

Random Forests™(Erfinder: Leo Breiman) werden für Regressions- und Klassifikationsprobleme eingesetzt. Ein Random Forest setzt sich aus vielen Entscheidungsbäumen zusammen. Es gibt zwei Zufallsmechanismen, die dazu benutzt werden um unterschiedliche Entscheidungsbäume an die Daten anzupassen. Für die Vorhersage werden die Vorhersagen aller Bäume aggregiert.

PRINZIPIEN BEIM LERNEN AUS DATEN

RASHOMON EFFEKT

Es gibt meist viele unterschiedliche Modelle, die einen Sachverhalt gleich gut beschreiben.

RASHOMON EFFEKT

Es gibt meist viele unterschiedliche Modelle, die einen Sachverhalt gleich gut beschreiben.

Rashomon ist ein japanischer Film, in dem 4 Zeugen unterschiedliche Versionen von einem beobachteten Verbrechen erzählen. Alle Versionen erklären die Fakten und doch sind alle widersprüchlich.
Übertragen auf das statistische Lernen bedeutet das, dass unterschiedliche Modelle (z.B. mit unterschiedlichen Kovariablen) die Daten gleich gut vorhersagen. Jedes Modell hat aber eine andere Interpretation.
Algorithmen wie Random Forest und Boosting nutzen diesen Effekt aus und aggregieren über viele Modelle. Außerdem ist es im Machine Learning gängige Praxis verschiedene Algorithmen zu benutzen und die Resultate für die Vorhersage zu aggregieren.

DIMENSIONALITÄT DER DATEN

→ Je höher die Dimensionalität (# Kovariablen) desto schwieriger das Trennen von Rauschen und Einflüssen

→ Gängige Praxis in der Statistik: Variablenselektion (theoretisch motiviert oder datengesteuert) und Dimensionsreduktion

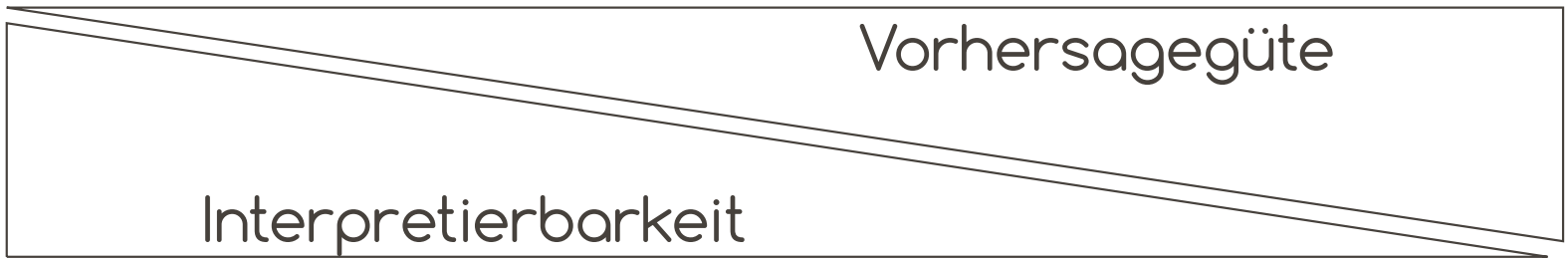
→ Gängige Praxis im Machine Learning: Erzeugen von vielen neuen Kovariablen um Vorhersage zu verbessern; Algorithmen meist robust für hochdimensionale Daten

DIMENSIONALITÄT DER DATEN

- → Je höher die Dimensionalität (# Kovariablen) desto schwieriger das Trennen von Rauschen und Einflüssen
- → Gängige Praxis in der Statistik: Variablenselektion (theoretisch motiviert oder datengesteuert) und Dimensionsreduktion
- → Gängige Praxis im Machine Learning: Erzeugen von vielen neuen Kovariablen um Vorhersage zu verbessern; Algorithmen meist robust für hochdimensionale Daten

Random Forest robust durch Aggregation von vielen Modellen und Randomisierung bei Variablenwahl. Support Vector Machines erzeugen sogar absichtlich höhere Dimensionalitäten der Daten um Trennbarkeit zu erreichen

VORHERSAGE VS. INTERPRETIERBARKEIT



Tree

Logist. Regression

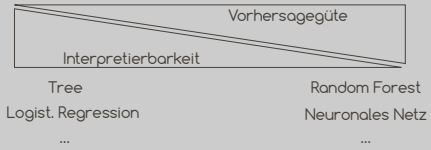
...

Random Forest

Neuronales Netz

...

VORHERSAGE VS. INTERPRETIERBARKEIT



Es gibt einen Tradeoff zwischen Interpretierbarkeit und Vorhersagekraft von Modellen: Komplexere Modelle liefern häufig genauere Vorhersagen. Gut interpretierbare Methoden sind meist schlecht in der Vorhersage. Ein Beispiel sind Entscheidungsbäume und Random Forests: Ein einzelner Entscheidungsbaum ist sehr intuitiv und auch für Laien leicht zu interpretieren, dafür sind sie sehr instabil und liefern nicht so gute Vorhersagen. Ein Aggregat von zufällig erzeugten Bäumen, ein Random Forest, hat eine ausgezeichnete Vorhersagegüte, aber die Modellstruktur lässt sich nicht mehr interpretieren.

MODELLGÜTE

→ Statistik: Gütekriterien beruhen häufig auf Modellannahmen und werden auf Trainingsdaten berechnet. Manchmal gar keine Evaluierung

→ Machine Learning: Gängige Praxis: Kreuzvalidierung, extra Testset

Wie gut ist ein statistisches Modell wenn die Vorhersagegüte schlecht ist? Darf man Parameter und p-Werte interpretieren?

PERSÖNLICHE ERFAHRUNGEN

STATISTISCHE BERATUNG

Stereotypische Anwender ...

- → sind z.B. Tiermediziner, Linguisten, Biologen, ...
- → sehnen sich nach p-Werten für Koeffizienten
- → möchten Interpretierbarkeit, keine Vorhersagegüte
- → haben meist Erfahrung mit linearen Modellen (kein Machine Learning)
- → kümmeren sich nicht um Modelldiagnose

Stereotypische Anwender ...

- → sind z.B. Tiermediziner, Linguisten, Biologen, ...
- → sehnen sich nach p-Werten für Koeffizienten
- → möchten Interpretierbarkeit, keine Vorhersagegüte
- → haben meist Erfahrung mit linearen Modellen (kein Machine Learning)
- → kümmeren sich nicht um Modelldiagnose

Vor allem in der Forschung möchten Anwender Hypothesen mit Hilfe von Modellen überprüfen. Dabei ist meistens nicht von Interesse wie gute Vorhersagen ein Modell liefert, sondern z.B. auf welche Werte die Koeffizienten geschätzt wurden und ob sie signifikant sind. Algorithmische Modellierung ist hier in den meisten Fällen wegen mangelnder Interpretierbarkeit nicht interessant.

KAGGLE

Algorithmen der Gewinner auf kaggle, Plattform für Prognose-Wettbewerbe:

- → Job Salary Prediction: »I used deep neural networks«
- → Observing Dark Worlds: »Bayesian analysis provided the winning recipe for solving this problem«
- → Give Me Some Credit: »In the end we only used five supervised learning methods: a random forest of classification trees, a random forest of regression trees, a classification tree boosting algorithm, a regression tree boosting algorithm, and a neural network.«

KAGGLE

Algorithmen der Gewinner auf kaggle, Plattform für Prognose-Wettbewerbe:

- → Job Salary Prediction: »I used deep neural networks«
- → Observing Dark Worlds: »Bayesian analysis provided the winning recipe for solving this problem«
- → Give Me Some Credit: »In the end we only used five supervised learning methods: a random forest of classification trees, a random forest of regression trees, a classification tree boosting algorithm, a regression tree boosting algorithm, and a neural network.«

Kaggle (**http://www.kaggle.com**) ist eine Internetplattform auf der Vorhersageprobleme als Wettbewerbe ausgeschrieben werden. Daten werden zur Verfügung gestellt und es gewinnt derjenige, dessen Vorhersage auf Testdaten am Besten war. Es ist nicht teil des Wettbewerbes relevante Einflussgrößen zu identifizieren oder Erkenntnisse über den datengenerierenden Prozess zu finden. Da hier nur die Vorhersage zählt ist die algorithmische Modellierung klar im Vorteil gegenüber der Kultur der Datenmodellierung.

FAZIT

Ein Statistiker sollte:

- → Modelle kritisch evaluieren
- → Prognosegüte als Kriterium für die Modellgüte benutzen
- → Das beste Modell suchen, egal ob aus der Statistik oder Machine Learning
- → Machine Learning in die Toolbox mit aufnehmen.
- → Sich immer bewusst machen:
  »All models are wrong, but some are useful« (G. Box)

WEITERFÜHRENDE LITERATUR


L. Breiman
Statistical modeling: The two cultures (with comments and a rejoinder by the author)
Institute of Mathematical Statistics, 2001.


T. Hastie, R. Tibshirani and J. Friedman
The elements of statistical learning
Springer New York, 2001

Vielen Dank für die Aufmerksamkeit

D.R. Cox in der Antwort auf Breiman's Paper:

» Descriptively appealing and transparent methods with a firm model base are the ideal. «

B. Efron in der Antwort auf Breiman's Paper:

» we are being asked to face problems that never heard of good experimental design «
Allerdings hält aber den Wert der Vorhersagegüte für überbewertet von Breiman.