DATA SCIENCE

8. ML Introduction Exercise

Sandra Tieben

What is the meaning of *generalization* in the paper?

Der Autor schreibt “The fundamental goal of machine learning is to generalize beyond the examples in the training set.” und begründet diesen Satz damit, dass unabhängig von der Menge der Trainingsdatensätze es unwahrscheinlich ist, auf die gleichen Datensätze erneut zu treffen. Anfänger erliegen oft dem Trugschluss, dass ein gewählter Classifier korrekt ist nach dem Testen der Trainingsdaten und oft zeigt sich dann bei der Anwendung auf neuen Daten, dass der Classifier oft nicht besser ist als bloßes Raten.

Somit sind laut dem Autor die nützlichsten Learners diejenigen, denen nicht nur Annahmen fest “verdrahtet” sind, sondern die “generalisierte”. Also Learners, die es uns erlauben, Annahmen explizit zu formulieren, sie stark zu variieren und sie automatisch in das Lernen einzubeziehen.

What are the many faces of overfitting?

Wenn Wissen und Daten nicht ausreichend ist, einen korrekten Classifier zu erstellen, kann es vorkommen, dass ein Classifier (bzw. die “Halluzination” davon) erzeugt wird, der keine sinnvollen Ausgaben erzeugt. Dieses Phänomen wird als “Overfitting” bezeichnet. Beispiel: Ein Classifier, der 100% Genauigkeit bei den Trainingsdaten hat, aber nur 50% Genauigkeit bei den Testdaten, obwohl er in beiden 75% Genauigkeit haben könnte, ist “overfit” (überangepasst). Laut dem Autor gibt es zwei gegenläufige Generalisierungsfehler - Bias (“Underfitting”) und Varianz (“Overfitting”).

Bei Bias handelt es sich um eine systematische Abweichung, während bei Varianz zufällig (unabhängig vom realen Signal) gelernt wird.

Ein anderes Problem ist, dass es durch Vermeidung von Overfitting zu Underfitting/Bias kommt. Es ist nicht möglich, beide Fehler gleichzeitig zu vermeiden (no free lunch).

Overfitting kann durch Noise verursacht werden. Dies ist allerdings nicht zwingend die Ursache. Auch dafür gibt der Autor zwei Beispiele:

1. Boolean classifier that is just the disjunction of the examples labeled “true” in the training set. This classifier gets all the training examples right and every positive test example wrong, regardless of whether the training data is noisy or not.
2. A mutual fund that beats the market 10 years in a row looks very impressive, until you realize that, if there are 1,000 funds and each has a 50% chance of beating the

market on any given year, it is quite likely that one will succeed all 10 times just by luck (“Multiple testing”).

Why do humans have problems in higher dimensions?

Das zweitgrößte Problem (nach Overfitting) beim ML sind laut dem Autor höhere Dimensionen, da korrekte Generalisierung mit zunehmender Anzahl an Dimensionen exponentiell schwieriger wird. Zudem brechen ML-Algorithmen bei höheren Dimensionen zusammen (z.B. durch Rauschen von vielen irrelevanten Signalen oder auch weil bei höheren Dimensionen alle Beispiele für Menschen gleich aussehen).

Menschen kommen aus einer 3D-Welt und finden sich in dieser intuitiv zurecht. Das gilt leider für höhere Dimensionen nicht: Diese sind oft schwer zu verstehen, wodurch es schwer fällt, gute Classifier zu erstellen.

What is feature engineering?

Der Autor erläutert, dass Lernen einfach sei, wenn man viele unabhängige Merkmale hat, die jeweils gut mit der Klasse korrelieren. Aber auch, dass es schwer möglich sei, wenn die Klasse eine sehr komplexe Funktion der verschiedenen Merkmale ist. Um gute Ergebnisse zu erreichen, ist in diesem Zusammenhang ist das Feature Engineering extrem wichtig:

Feature Engineering ist

- interativer Prozess (Try and Error) bestehend aus Learner ausführen, Ergebnisanalyse, Anpassung der Daten und/oder des Learners (und Wiederholung).
- domainspezisch
- idealerweise möglichst weitgehende Automatisierung des Prozesses

Why does more data beats clever algorithms?

Der Autor nennt die Faustregel, dass ein dummer Algorithmus mit sehr vielen Daten einen cleveren mit wenig Daten schlägt. Dazu gibt mehrere Argumente an, weshalb mehr Daten bessere Algorithmen schlägt:

- In erster Näherung tun alle Algorithmen - egal ob clever oder dumm - dasselbe
- Clevere Algorithmen brauchen sehr lange zum Lernen (und Zeit ist oft der limitierende Faktor beim ML)
- bessere Algorithmen sind schwieriger in der Anwendung

What is ensemble learning?

Während in den Anfangszeiten von ML die Entwickler ihren “Lieblingslearner” in verschieden Variationen für alles nutzen, ging man danach dazu über, die verschiedenen Learner zu variieren, um den besten zu ermitteln. Die aktuelle (noch bessere) Vorgehensweise ist “**Model ensembles**”, wobei bessere Ergebnisse durch Kombination verschiedener Learner erzielt werden. Hierzu gibt es verschiedene Methoden, z.B.:

- **Bagging** (einfachste Methode): Durch Resampling werden Zufallsvarianten der Trainingsdaten erzeugt, die jeweils ein Classifier lernt. Im Anschluss werden die Ergebnisse kombiniert. (Funktioniert, weil Varianz stark reduziert wird und sich dennoch der Bias nur leicht erhöht.)
- **Boosting**: Trainingsbeispiele haben Gewichtungen, und diese werden so variiert, dass jeder neue Klassifikator sich auf die Beispiele konzentriert, die die vorherigen falsch waren.
- **Stacking:** Der Outputs des verschiedenen Classifiers wird als Inputs des “higher-level” Learner genutzt, der die beste Kombination ermittelt.

Model Ensembles != Bayesian Model Averaging (BMA) [theoretisch optimales Lernen]

What is accuracy in data science?

Accuracy (“Genauigkeit”) ist eine Kenngröße zur Bewertung von Klassifizierungsmodellen, die quasi sagt, in wie viel Fällen unser Modell richtig liegt.