Aufgabe 8 Tool Machine Learning Introduction Exercise Carsten Brauer S8963350

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

Ich übersetze „to generalize“ mit „verallgemeinern“. Und zwar über das Trainingset hinaus, denn den Erfolg eines classifiers anhand des Trainingsets zu messen, käme einem bloßen Erinnern gleich. Es empfiehlt sich, einen Teil des Trainingsets nicht sofort dem Classifier zur Verfügung zu stellen, sondern erst für Tests zu verwenden.

Es geht bei der Verallgemeinerung also darum, aus bekannten Daten für neue Daten ein Resultat zu extrapolieren

- What are the many faces of overfitting?

Overfitting (Überanpassung) liegt vor, wenn eigentlich wenig relevante Merkmale aus dem Trainingsset oder anderen Daten mit zu hoher Gewichtung in den Entcheidungsprozesse einfliessen, zum Beispiel die Hautfarbe bei der Erkennung von Schönheit, wenn im Trainigsset 90 % aller Schönheitskönniginnen/könige eine helle Hautfarbe hatten und die KI anschließend nur hellhäutige Menschen als schön klassifiziert.

Overfitting kann in in bias (Verzerrung), variance (Varianz) und irreducible errors (nicht reduzible Fehler) zerlegt werden. Letztere sind Fehler, die darauf zurückzuführen sind, dass nicht alle relevanten Merkmale bekannt sind, die das Ergebnis beeinflussen.

- Why do humans have problems in higher dimensions?

Es ist relativ einfach, mit einer gewissen Wahrscheinlichkeit anhand von Größe (Dimension 1, X) und Gewicht (Dimension 2, Y) zwischen Mann uind Frau zu unterscheiden. Es ergibt sich im Koordinatensystem eine 1-Dimensionale Linie, wobei es passieren kann, dass der ein oder andere Punkt auf der falschen Site der Linie liegt.

Es wird schwieriger, wenn man die Genauigkeit erhöhen will und zusätzlich eine 3. Dimension (Z) einführt, z.B. eine Funktion zur Berechnung eines Wertes aus Größe und Gewicht. Wählt man diese Funktion geschickt, ergibt sich eine 2-Dimensionale Ebene im 3-Dimensionalen Raum, der M und F sauber voneinander trennt. Das kann man sich noch vorstellen.

Fügt man aber darüberhinaus weitere Dimensionen hinzu, gelingt es nicht mehr, sich vorzustellen, wie das aussieht.

- What is feature engineering?

Bei feature engineering geht es darum, herauszufinden, welche features (Dimensionen) in welcher Kombination und mit welcher jeweiligen Gewichtung am besten für ein bestimmtes Projekt geeignet sind.

Bei der Unterscheidung zwischen M und F könnte z.B. das Feature „Beruf“ hilfreich sein, da es geschlechtstypische Berufe gibt (Extreme: Hebammen vs. Feuerwehrmann). Weniger hilfreich ist z.B. die Postleitzahl des Wohnortes.

- Why does more data beats clever algorithms?

Zitat Prof. Thomaschewski, Modul UX, WS 19/20 (sinngemäß): Wenn Sie genug Datensätze haben, bügelt die Statistik Ihnen alles glatt.

Heißt: Es wird mit steigender Anzahl einfacher, die outliers zu identifizieren und es steigt die Wahrscheinlichkeit, dass ein vorherzusagender Wert mit annähernd den gesuchten Dimensionsausprägungen sich bereits in der Datenbank befindet. Der Vorhersage-Anteil der Entscheidung wird also kleiner, der Such- oder Erinnerungs-anteil steigt.

- What is ensemble learning?

Das Verwenden von mehreren Varianten mehrerer Lernalgorithmen, wobei es darauf ankommt, nach welchen Kriterien ein Lernalgorithmus ein bestimmtes Objekt gerade eben nicht mehr in Gruppe A, sondern in Gruppe B steckt [1]

- What is accuracy in data science?

Ich würde sagen, das ist die Treffergenauigkeit der Vorhersage bzw. der Klassifizierung.

„[1] All learners essentially work by grouping nearby examples into the same class; the key difference is in the meaning of “nearby.”“ Seite 85.