Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Analyse von Visitentexten mittels maschinellem Lernen zur Optimierung der Leitlinienadhärenz auf einer Intensivstation

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

eingereicht von: Martin Hoffmann
geboren am: 06.09.1998
geboren in: Berlin

Gutachter: Prof. Dr. med. Dr. rer. nat. Felix Balzer
Dr. med. Fridtjof Schiefenhövel

eingereicht am: ..... verteidigt am: .....

Zusammenfassung

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Datenerfassung auf Intensivstationen

1.1.1 Ziel der Arbeit

Das Ziel der vorliegenden Arbeit ist es, mit Hilfe von maschinellem Lernen ein statistisches Modell zu entwickeln, um anhand von Freitexten medizinische Scores möglichst akkurat vorherzusagen.

1.2 Maschinelles Lernen

Der Begriff Maschinelles Lernen bezeichnet einen modernen Ansatz in der Arbeit an künstlicher Intelligenz. (cite).

hat sich im Zeitalter von Big Data und leistungsstarken Rechnern zu einem der Hauptforschungspunkte der Informatik entwicklt (cite).

(et al) definiert den Vorgang maschinellen Lernens folgendermaßen:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Im Allgemeinen werden also Algorithmen, die aus großen Datensätzen „lernen“ und damit Vorhersagen über unbekannte Daten machen, als maschinelles

Lernen bezeichnet.

1.2.1 Supervised Learning

Neben vielen weiteren Methoden stellt Supervised Learning einen zentralen Ansatz im Gebiet des maschinelles Lernen dar.

1.2.2 Regression vs Klassifikation

Klassifikation bezeichnet den Prozess, bei dem ein Datensatz einer oder mehreren Klassen aus einer endlichen Liste möglicher Klassen zugeordnet wird.

Dieser Ansatz findet beispielsweise bei der automatischen Kategorisierung von E-Mails (Spam oder nicht Spam) oder bei der Erkennung von handschriftlichen Texten (welches Symbol aus einem gegebenen Alphabet ist dargestellt?) Anwendung (CITE). Da es sich bei den betrachteten medizinischen Scores um diskrete, ganzzahlige Werte aus einem endlichen Wertebereich handelt, liegt auch hier die Anwendung eines Klassifizierungs-Verfahrens nahe.

Würde man aber die verschiedenen möglichen Werte eines Scores als separate und voneinander unabhängige Klassen betrachten, so ginge eine wichtige Information über deren Anordnung verloren. Im mathematischen Sinne stellen alle hier betrachteten medizinische Scores Totalordnungen dar. Sie erfüllen also die Anforderungen der Reflexivität, Antisymmetrie, Transitivität und Totalität. Bezeichne M die Menge aller möglichen Werte eines beliebigen medizinischen Scores. Folglich gilt für alle $a, b, c \in M$:

$$
\begin{aligned}
a \leq a & \qquad \text{(Reflexivität)} \\
a \leq b \wedge b \leq a \;\Rightarrow\; a = b & \qquad \text{(Antisymmetrie)} \\
a \leq b \wedge b \leq c \;\Rightarrow\; a \leq c & \qquad \text{(Transitivität)} \\
a \leq b \vee b \leq a & \qquad \text{(Totalität)}
\end{aligned}
$$

Damit lassen sich die verschiedenen Scores vergleichen und in ein Verhältnis

setzen. So ist ein RASS-Wert[1] von -4 (Tief sediert) beispielsweise deutlich näher an -3 (mäßig sediert) als an +1 (unruhig). Bei gängigen Verfahren zur Klassifizierung ginge diese Information verloren, da bei Metriken zur Bewertung solcher Modelle nur betrachtet werden kann, ob einem Eingabetext der richtige Score zugeordnet wird oder nicht. (Da viele der Scores bei der Vergabe zumindest teilweise auch auf dem persönlichen Ermessen des behandelnden Arztes/der behandelnden Ärztin beruht wäre selbst für menschliche Experten eine genaue Zuordnung eines Textes zu einer Punktzahl problematisch.)

Bei dieser Arbeit habe ich mich demnach dafür entschieden, die Vergabe von Scores anhand von Eingabetexten als klassisches Regressionsproblem zu betrachten, und die Ausgaben der Modelle im Zweifelsfall auf den nächstmöglichen ganzzahligen Wert zu runden. Dieser Ansatz fand auch bei früheren Arbeiten, die sich mit ähnlichen Fragestellungen befassten, Anwendung (CITE). Die Abweichung des vorhergesagten Werts eines Modells von dem nächst möglichen Wert stellt somit sogar einen rudimentären Ansatz zur Bewertung der Konfidenz bei der Vergabe einzelner Werte dar.

[1] Richmond Agitation-Sedation Scale

Kapitel 2

Übersicht über die Daten

2.1 section Daten

information über copra

- die daten wurden vor dem export pseudonymisiert
- information über die daten (was steht in texten, was sagen die scores)
- information, wie ich die daten in wertepaare umgewandelt habe

Literaturverzeichnis