Einführung in Web and Data Science Übungszettel 1

Gruppe 5, Übung Mo. 10 Uhr

Magnus Bender (LaTeX Beispiel)

15. Oktober 2022

Aufgabe 1

Wie wir alle wissen werden durch (unsere) Aktionen im Internet sehr viele Daten erzeugt, ob es nun Inhalte oder einfach nur Logfiles einer Verbindung sind.

Diese Datensätze müssen analysiert und verarbeitet werden, denn sie ermöglichen einen Rückschluss auf viele wertvolle Informationen, diese Analyse soll natürlich in Echtzeit und mit so geringer Leistung wie nur möglich passieren. Genau hier setzen die Web Science an, sie beschäftigen sich mit der Analyse, Vereinfachung und Wertung der Daten des Internets. Eine Suchmaschine im Internet tut nichts anderes, als Daten über die Webseiten und Nutzer zu sammeln, diese zu verarbeitet und dann für den User die Ergebnisse herauszufiltern, welche er gesucht hat.

Wir haben mit sehr großen Datenmengen zu tun, welche leider oft nicht in maschinenlesbaren Datensätzen vorliegen und interpretiert werden müssen. Anschließend müssen auf Basis der Daten Entscheidungen gefällt werden. Hierzu greifen wir mit Mathematik auf Stochastik, Statistik und Logik zurück und setzen die Verfahren dann mit Mitteln der Informatik um. Data Science ist quasi die Wissenschaft, mit der die Daten analysiert werden und auf die modernen Anforderungen der Daten des Internets bezogen spricht man von Web Science.

Aufgabe 2

(a) Ein komplexes Modell besteht aus mehr Merkmalen, einem längeren Algorithmus, es ist aufwändiger zu berechnen und benötigt mehr Speicher sowie Leistung. Bei einem einfachen Modell ist dies genau andersherum.

Neben der Komplexität des Modells ist auch die Vorverarbeitung relevant für den Speicherdearf des Modells. Wenn erst man ein Bild analysieren muss, dann benötigt man mehr Speicher, als wenn man schon Text hat.

(b) Wenn wir ein komplexes Modell aus unseren Trainingsdaten erstellt haben, muss es im Normallfall perfekt auf diese passen. Wir müssen jedoch bedenken, dass die Trainingsdaten nur eine Stichprobe der Realität sind und wir ein generalisierungsfähiges Modell bauen wollen. Daher kann es teilweise sinnvoll sein ein weniger komplexes Modell zu bauen, welches einen etwas größen Fehler bei den Trainingsdaten produziert, dafür aber allgemeingültig ist.

Man könnte sagen, man soll den gesunden Menschenverstand nutzen und einen Kompromiss finden.

(c) Einen neuen Datensatz mit einem komplexen Modell zu klassifizieren ist deutlich aufwändiger. Man hat mehr Merkmale und Dimensionen was dazu führt, dass man einen längeren Algorithmus, kompliziertere Operationen, mehr Speicher sowie Leistungsbedarf hat. Auch die Möglichkeit Fehler im Modell zu haben oder während der Klassifizierung der Daten zu machen ist höher.

Weiterhin kann auch, wie schon unter (b) genannt, das Modell weniger generalisierungsfähig sein, als ein weniger komplexes Modell.

Aufgabe 3 – Möglichen Aufgabe Analysis

Entscheiden Sie, ob die gegebenen Folgen

$$(a_n) = \frac{5n^3 + (-1)^n n + 3}{n^2 - 2}$$
$$(b_n) = \frac{2n^2 - 5}{-5n^2 + 2}$$
$$(c_n) = \frac{(-1)^n n^2}{6n^7 - 4(-1)^n n^3 + n}$$
$$(d_n) = \frac{4n^2 + 2(-1)^n}{(-1)^n n^3 - 3}$$
$$(e_n) = \frac{(-2)^n n^3}{3(-1)^n n^3 - 5n^2 + n}$$

konvergent, divergent, bestimmt divergent gegen ∞, bestimmt divergent gegen $-\infty$ oder unbestimmt divergent sind.

Aufgabe 4 – Mögliche Aufgabe theoretische Informatik

Konstruieren Sie reguläre Ausdrücke, die genau die jeweilige Sprache über dem Alphabet $\Sigma = \{0, 1\}$ beschreiben:

1.) $L_1 = \{w \in \Sigma^* \mid w$ enthält 11 nicht als Teilwort $\}$

2.) $L_2 = \{w \in \Sigma^* \mid w$ enthält genau 4 Einsen $\}$

3.) $L_3 = \{w \in \Sigma^* \mid w$ enthält eine gerade Anzahl von Nullen $\}$

4.) $L_4 = \{w \in \Sigma^* \mid |w|$ ist ungerade $\}$