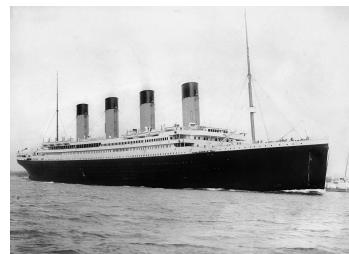


Nichtlineare Datenstrukturen: Bäume

Der ID3-Algorithmus

Die Titanic wurde 1912 gebaut und stach am 14. April zu ihrer Jungfernreise in See. An Bord waren über 2200 Passagiere, die sich in drei Klassen zur ersten Überfahrt des Ozeanriesen eingebucht hatten (legal bezahlt oder illegal eingeschmuggelt). Gegen 23:40 Uhr kollidierte die Titanic mit einem Eisberg und sank zwei Stunden und 40 Minuten später vollständig.

Trotz der einigermaßen langen Zeit bis zum Sinken des Schiffes überlebten nur etwa 686 Menschen das Unglück.



Der vorliegende Datensatz ist ein bereinigter Auszug aus der Passierliste der Titanic. Wir wollen nun mittels des *ID3-Algorithmus* den Computer einen *Entscheidungsbaum* generieren lassen, der für einen Passgier entscheidet, ob er überlebt hat, oder nicht.

Aufgabe 1

Berechnen sie die *Entropie* E_{gesamt} des vollständigen Datensatzes bezüglich des Zielattributes *survived*.

$$E_{gesamt} = \underline{\hspace{10cm}}$$

Die *Entropie* berechnet sich durch $E_{gesamt} = -P_{ja} \cdot \log_2 P_{ja} - P_{nein} \cdot \log_2 P_{nein}$, wobei $P_{ja/nein}$ den Anteil der Datensätze mit der Ausprägung „überlebt = ja“ bzw. „überlebt = nein“ darstellt.

Aufgabe 2

Berechnen sie die *Entropie* für alle Ausprägungen (Werte) des Attributes _____ bezüglich des Zielattributes *survived*.

$$E = \underline{\hspace{10cm}}$$

$$E = \underline{\hspace{10cm}}$$

$$E = \underline{\hspace{10cm}}$$

Die *Entropie* $E_{Attributwert}$ berechnet sich wie oben, allerdings werden nur Zeilen berücksichtigt, bei denen das betreffende Attribut den betreffenden Wert hat (z.B. alle Zeilen bei denen *sex* gleich „male“ ist).

Aufgabe 3

Berechnen sie den *Informationsgewinn* für ihr Attribut.

$$IG = \underline{\hspace{10cm}}$$

Der *Informationsgewinn* $IG_{Attribut}$ eines Attributs berechnet sich durch $E_{gesamt} - \sum P_{Attributwert} \cdot E_{Attributwert}$.