Homework 3.1 - Machine Learning

Lars Gerne, Niklas Hauschel

09.03.2020

Inhaltsverzeichnis

1 Entropie

1.1 Begriffserklärung

Die Entropie gibt die Unreinheit von Daten an. Wenn der Wert geringer ist, dann können die Daten einfacher klassifiziert werden. Bei einem höherem Wert können die Daten schlecht klassifiziert werden. Bei einer hohen Entropie werden mehr Bits benötigt, um die Information zu beschreiben.

1.2 Formel

$$H(S) = -\sum_{c \in C} p(c) log_2(p(c))$$

H - griechisches E (Eta), steht für Entropie
S - Datensatz
C - Menge aller Kategorien
c - Kategorie

2 Aufgabe

Entscheidungsbaum für einen Datensatz mithilfe des ID3-Algorithmus berechnen.

outlook	temp	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
rainy	mild	high	true	no

Tabelle 1: Playing Tennis Game - data set

1. Schritt: Gesamtentropie berechnen
   Hiefür muss die Gesamtzahl an ja/nein Ereignissen gezählt werden.

$$H(S) = -\left(\frac{9}{14} log_2\left(\frac{9}{14}\right) + \frac{5}{14} log_2\left(\frac{5}{14}\right)\right)$$
$$\approx 0.940$$

2. Schritt: Information Gain für jedes Feature berechnen
   Entropie für jede Klassifizierung berechnen:

outlook	overcast	sunny	rainy	sum
YES	4	2	3	9
NO	0	3	2	5
sum	4	5	5	14

$$H(outlook = overcast) = -\left(\frac{4}{4} log_2\left(\frac{4}{4}\right) + 0 log_2\left(0\right)\right)$$
$$= 0$$
$$H(outlook = sunny) = -\left(\frac{2}{5} log_2\left(\frac{2}{5}\right) + \frac{3}{5} log_2\left(\frac{3}{5}\right)\right)$$
$$\approx 0.971$$
$$H(outlook = rainy) = -\left(\frac{3}{5} log_2\left(\frac{3}{5}\right) + \frac{2}{5} log_2\left(\frac{2}{5}\right)\right)$$
$$\approx 0.971$$

   Information Gain des Features:

$$IG(S, A_{outlook}) = 0.94 - \left(\frac{4}{14}0 + \frac{5}{14}0.971 + \frac{5}{14}0.971\right)$$
$$= 0.246$$

temperature	hot	mild	cool	sum
YES	2	4	3	9
NO	2	2	1	5
sum	4	6	4	14

$$H(temp = hot) = -\left(\frac{2}{4}log_2\left(\frac{2}{4}\right) + \frac{2}{4}log_2\left(\frac{2}{4}\right)\right)$$
$$= 1$$
$$H(temp = mild) = -\left(\frac{4}{6}log_2\left(\frac{4}{6}\right) + \frac{2}{6}log_2\left(\frac{2}{6}\right)\right)$$
$$\approx 0.918$$
$$H(temp = cool) = -\left(\frac{3}{4}log_2\left(\frac{3}{4}\right) + \frac{1}{4}log_2\left(\frac{1}{4}\right)\right)$$
$$\approx 0.811$$

Information Gain des Features:

$$IG(S, A_{outlook}) = 0.94 - \left(\frac{4}{14}1 + \frac{6}{14}0.918 + \frac{4}{14}0.811\right)$$
$$= 0.029$$

humidity	high	normal	sum
YES	3	6	9
NO	4	1	5
sum	7	7	14

$$H(humidity = high) = -\left(\frac{3}{7}log_2\left(\frac{3}{7}\right) + \frac{4}{7}log_2\left(\frac{4}{7}\right)\right)$$
$$\approx 0.985$$
$$H(humidity = normal) = -\left(\frac{6}{7}log_2\left(\frac{6}{7}\right) + \frac{1}{7}log_2\left(\frac{1}{7}\right)\right)$$
$$\approx 0.592$$

Information Gain des Features:

$$IG(S, A_{outlook}) = 0.94 - \left(\frac{7}{14}0.985 + \frac{7}{14}0.592\right)$$

$$= 0.152$$

windy	FALSE	TRUE	sum
YES	6	3	9
NO	2	3	5
sum	8	6	14

$$H(windy = TRUE) = -\left(\frac{3}{6}log_2\left(\frac{3}{6}\right) + \frac{3}{6}log_2\left(\frac{3}{6}\right)\right)$$

$$= 1$$

$$H(windy = FALSE) = -\left(\frac{6}{8}log_2\left(\frac{6}{8}\right) + \frac{2}{8}log_2\left(\frac{2}{8}\right)\right)$$

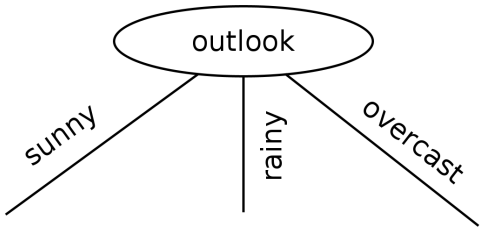
$$\approx 0.811$$

Information Gain des Features:

$$IG(S, A_{outlook}) = 0.94 - \left(\frac{8}{14}0.811 + \frac{6}{14}1\right)$$

$$= 0.049$$

3. Schritt: Das Feature mit dem größten IG wird als Wurzelknoten gewählt. Daraus ergibt sich folgender Baum:



Für jeden Zweig muss rekursiv wieder ein neuer Wurzelknoten bestimmt werden.

1. Gesamtentropie berechnen:
   Für das Subset S_{sunny} ergibt sich folgender Datensatz:

outlook	temp	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
sunny	mild	high	false	no
sunny	cool	normal	false	yes
sunny	mild	normal	true	yes

$$H(S_{sunny}) = -\left(\frac{2}{5}log_2\left(\frac{2}{5}\right) + \frac{3}{5}log_2\left(\frac{3}{5}\right)\right)$$
$$\approx 0.971$$

2. Information Gain für jedes Feature berechnen:

temperature	hot	mild	cool	sum
YES	0	1	1	2
NO	2	1	0	3
sum	2	2	1	5

$$H(temp = hot) = 0$$
$$H(temp = mild) = 1$$
$$H(temp = cool) = 0$$
$$IG(S_{sunny}, A_{temp}) = 0.971 - \left(\frac{2}{5}0 + \frac{2}{5}1 + \frac{1}{5}0\right)$$
$$\approx 0.571$$

humidity	high	normal	sum
YES	0	2	2
NO	3	0	3
sum	3	2	5

$$H(humidity = high) = 0$$

$$H(humidity = normal) = 0$$

$$IG(S_{sunny}, A_{humidity}) = 0.971 - \left(\frac{3}{5}0 + \frac{2}{5}0\right)$$

$$\approx 0.971$$

windy	FALSE	TRUE	sum
YES	1	1	3
NO	1	2	3
sum	2	3	5

$$H(windy = FALSE) = 1$$

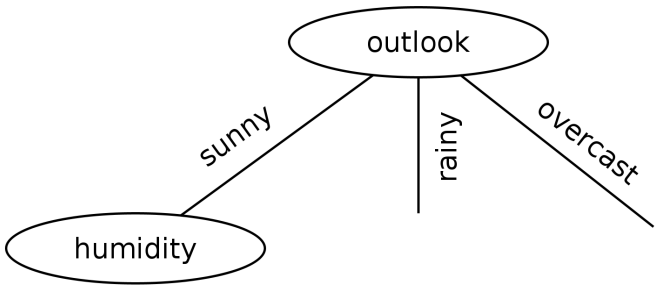
$$H(windy = TRUE) = -\left(\frac{1}{3}log_2\left(\frac{1}{3}\right) + \frac{2}{3}log_2\left(\frac{2}{3}\right)\right)$$

$$\approx 0.918$$

$$IG(S_{sunny}, A_{windy}) = 0.971 - \left(\frac{2}{5}1 + \frac{3}{5}0.918\right)$$

$$\approx 0.020$$

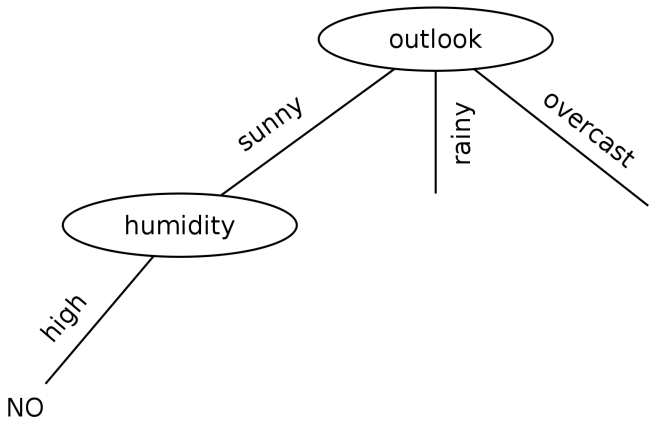
3. Schritt: Das Feature mit dem größten IG wird als Wurzelknoten gewählt. Daraus ergibt sich folgender Baum:



1. Gesamtentropie berechnen:
   Für das Subset $S_{sunny,high}$ ergibt sich folgender Datensatz:

outlook	temp	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
sunny	mild	high	false	no

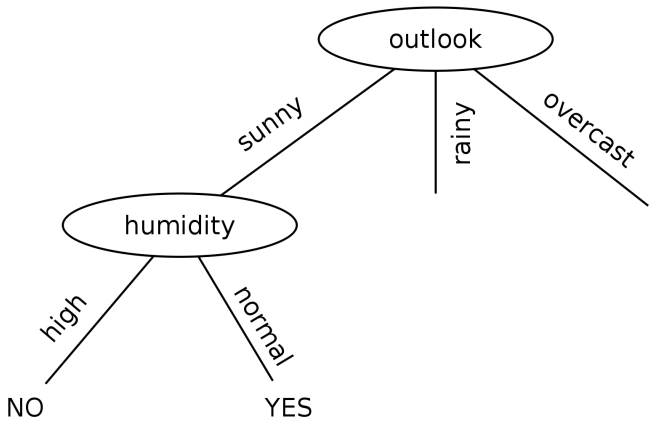
   Es muss keine Entropie berechnet werden, da alle Einträge das Ergebnis „no“ aufweisen.
   Daraus ergibt sich folgender Baum:



1. Gesamtentropie berechnen:
   Für das Subset $S_{sunny,normal}$ ergibt sich folgender Datensatz:

outlook	temp	humidity	windy	play
sunny	cool	normal	false	yes
sunny	mild	normal	true	yes

   Es muss keine Entropie berechnet werden, da alle Einträge das Ergebnis „yes“ aufweisen.
   Daraus ergibt sich folgender Baum:



1. Gesamtentropie berechnen:

outlook	temp	humidity	windy	play
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
rainy	mild	normal	false	yes
rainy	mild	high	true	no

$$H(S_{overcast=rainy}) = -\left(\frac{2}{5}log_2\left(\frac{2}{5}\right) + \frac{3}{5}log_2\left(\frac{3}{5}\right)\right)$$
$$\approx 0.971$$

2. Information Gain für jedes Feature berechnen:

temperature	mild	cool	sum
YES	2	1	3
NO	1	1	2
sum	3	2	5

$$H(temp = mild) = -\left(\frac{2}{3}log_2\left(\frac{2}{3}\right) + \frac{1}{3}log_2\left(\frac{1}{3}\right)\right)$$
$$\approx 0.918$$
$$H(temp = cool) = 1$$
$$IG(S_{rainy}, A_{temp}) = 0.971 - \left(\frac{3}{5}0.92 + \frac{2}{5}1\right)$$
$$\approx 0.019$$

humidity	high	normal	sum
YES	1	2	3
NO	1	1	2
sum	2	3	5

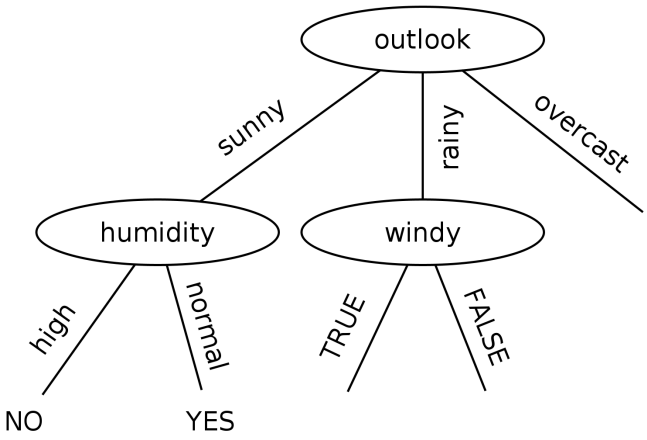
$$H(humidity = high) = 1$$
$$H(humidity = normal) = -\left(\frac{2}{3}log_2\left(\frac{2}{3}\right) + \frac{1}{3}log_2\left(\frac{1}{3}\right)\right)$$
$$\approx 0.918$$

$$IG(S_{rainy}, A_{humidity}) = 0.971 - \left(\frac{3}{5}0.92 + \frac{2}{5}1\right)$$
$$\approx 0.019$$

windy	TRUE	FALSE	sum
YES	0	3	3
NO	2	0	2
sum	2	3	5

$$H(windy = TRUE) = 0$$
$$H(windy = FALSE) = 0$$
$$IG(S_{rainy}, A_{windy}) = 0.971 - \left(\frac{3}{5}0 + \frac{2}{5}0\right)$$
$$\approx 0.971$$

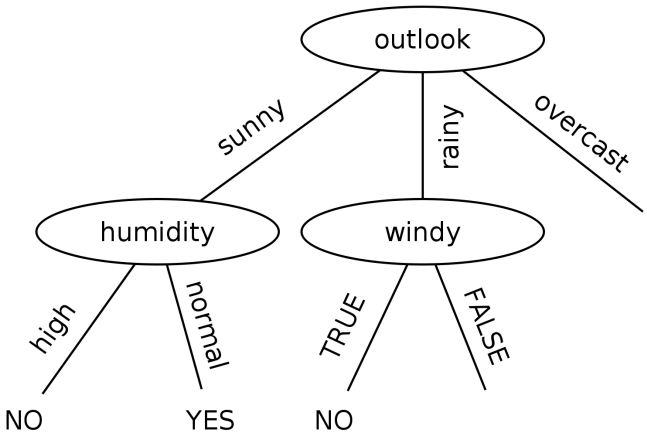
3. Schritt: Das Feature mit dem größten IG wird als Wurzelknoten gewählt. Daraus ergibt sich folgender Baum:



1. Gesamtentropie berechnen:
   Für das Subset $S_{rainy,TRUE}$ ergibt sich folgender Datensatz:

outlook	temp	humidity	windy	play
rainy	cool	normal	true	no
rainy	mild	high	true	no

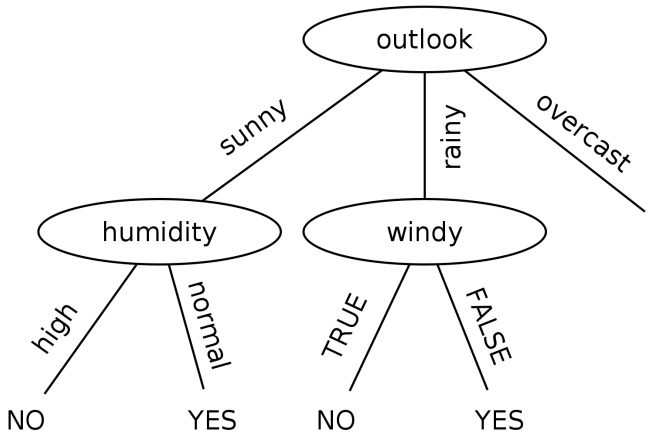
   Es muss keine Entropie berechnet werden, da alle Einträge das Ergebnis „no“ aufweisen.
   Daraus ergibt sich folgender Baum:



1. Gesamtentropie berechnen:
   Für das Subset $S_{rainy,FALSE}$ ergibt sich folgender Datensatz:

outlook	temp	humidity	windy	play
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	mild	normal	false	yes

   Es muss keine Entropie berechnet werden, da alle Einträge das Ergebnis „yes“ aufweisen.
   Daraus ergibt sich folgender Baum:



1. Gesamtentropie berechnen:
   Für das Subset $S_{outlook=overcast}$ ergibt sich folgender Datensatz:

outlook	temp	humidity	windy	play
overcast	hot	high	false	yes
overcast	cool	normal	true	yes
overcast	mild	high	true	yes

   Es muss keine Entropie berechnet werden, da alle Einträge das Ergebnis „yes“ aufweisen.
   Daraus ergibt sich folgender Baum: