

Data Mining Tutorial

Klassifikation III

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2014-06-13 — KDD Übung

Wdh: beim Split von T im Attribut A in Partitionen $T_1 \dots T_m$:

$$\text{Entropie}(T) = - \sum_{i=1}^k p_i \cdot \log p_i$$

$$\text{Informationsgewinn}(T, A) = \text{Entropie}(T) - \sum_{i=1}^m \frac{|T_i|}{|T|} \text{Entropie}(T_i)$$

Wdh: beim Split von T im Attribut A in Partitionen $T_1 \dots T_m$:

$$\text{Entropie}(T) = - \sum_{i=1}^k p_i \cdot \log p_i$$

$$\text{Informationsgewinn}(T, A) = \underbrace{\text{Entropie}(T)}_{\text{vorher}} - \underbrace{\sum_{i=1}^m \frac{|T_i|}{|T|} \text{Entropie}(T_i)}_{\text{mittlere Entropie nachher}}$$

Mittlere Entropie, Gewichtet nach *Anteil* an der Datenbank!

Wdh: beim Split von T im Attribut A in Partitionen $T_1 \dots T_m$:

$$\text{Entropie}(T) = - \sum_{i=1}^k p_i \cdot \log p_i$$

Mittlere Entropie, Gewichtet nach *Anteil* an der Datenbank!

Komplette Datenbank:

$$\text{Entropie}(T) = 1, \text{ da } p(R = \text{low}) = \frac{1}{2} = p(R = \text{high})$$

(Hier: \log_2 – eine andere Basis erzeugt aber den gleichen Baum!)

Informationsgewinn im Attribut Zeit: Entropie für T_1

1-2 Jahre: $T_1 =$ Personen 1,4,6

$$p(R = \text{low}) = \frac{1}{3}$$

$$p(R = \text{high}) = \frac{2}{3}$$

$$\begin{aligned} \text{Entropie}(T_1) &= - \sum_{i=1,2} p_i \log p_i \\ &= - \left(\frac{1}{3} \log \frac{1}{3} + \frac{2}{3} \log \frac{2}{3} \right) \\ &\approx 0.918 \end{aligned}$$

Informationsgewinn im Attribut Zeit: Entropie für T_2
2-7 Jahre: $T_2 = \text{Personen } 2,7,8$

$$p(R = \text{low}) = \frac{2}{3}$$

$$p(R = \text{high}) = \frac{1}{3}$$

$$\begin{aligned} \text{Entropie}(T_2) &= \text{Entropie}(T_1) \\ &\approx 0.918 \end{aligned}$$

Informationsgewinn im Attribut Zeit: Entropie für T_3
> 7 Jahre: $T_3 = \text{Personen } 3,5$

$$p(R = \text{low}) = \frac{1}{2}$$

$$p(R = \text{high}) = \frac{1}{2}$$

$$\begin{aligned} \text{Entropie}(T_3) &= - \left(\frac{1}{2} \log \frac{1}{2} \right) \cdot 2 \\ &= 1 \end{aligned}$$

Informationsgewinn für das Attribut Zeit:

$$\begin{aligned} & \text{Informationsgewinn}(T, \text{Zeit}) \\ &= \text{Entropie}(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} \text{Entropie}(T_i) \\ &= 1 - \left(\frac{3}{8} \cdot 0.918 + \frac{3}{8} \cdot 0.918 + \frac{2}{8} \cdot 1 \right) \\ &\approx 0.06 \end{aligned}$$

Informationsgewinn im Attribut Geschlecht: Entropie für T_1
m: $T_1 = \text{Personen } 1,2,5,6,8$

$$p(R = \text{low}) = \frac{2}{5}$$

$$p(R = \text{high}) = \frac{3}{5}$$

$$\text{Entropie}(T_1) \approx 0.971$$

Informationsgewinn im Attribut Geschlecht: Entropie für T_2
w: $T_2 = \text{Personen } 3,4,7$

$$p(R = \text{low}) = \frac{2}{3}$$

$$p(R = \text{high}) = \frac{1}{3}$$

$$\text{Entropie}(T_2) \approx 0.918$$

Informationsgewinn für das Attribut Geschlecht:

$$\begin{aligned} & \text{Informationsgewinn}(T, \text{Geschlecht}) \\ &= \text{Entropie}(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} \text{Entropie}(T_i) \\ &= 1 - \left(\frac{5}{8} \cdot 0.971 + \frac{3}{8} \cdot 0.918 \right) \\ &\approx 0.05 \end{aligned}$$

Informationsgewinn im Attribut Wohnort: Entropie für T_1
Stadt: $T_1 = \text{Personen } 1,7,8$

$$p(R = \text{low}) = 1$$

$$p(R = \text{high}) = 0$$

$$\text{Entropie}(T_1) = 0$$

Informationsgewinn im Attribut Wohnort: Entropie für T_2
Land: $T_2 = \text{Personen } 2,3,4,5,6$

$$p(R = \text{low}) = \frac{1}{5}$$

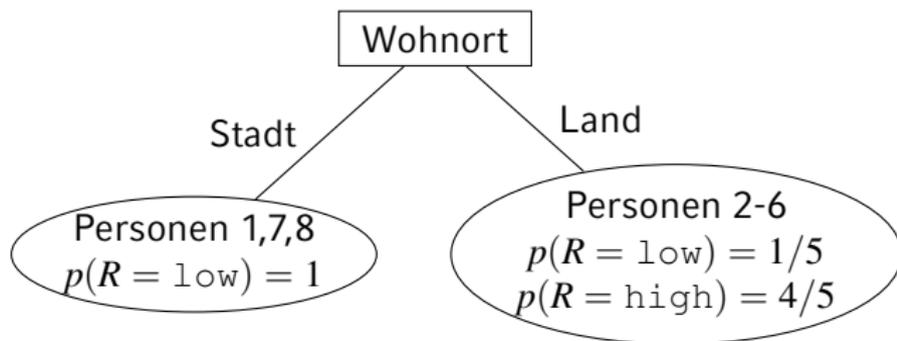
$$p(R = \text{high}) = \frac{4}{5}$$

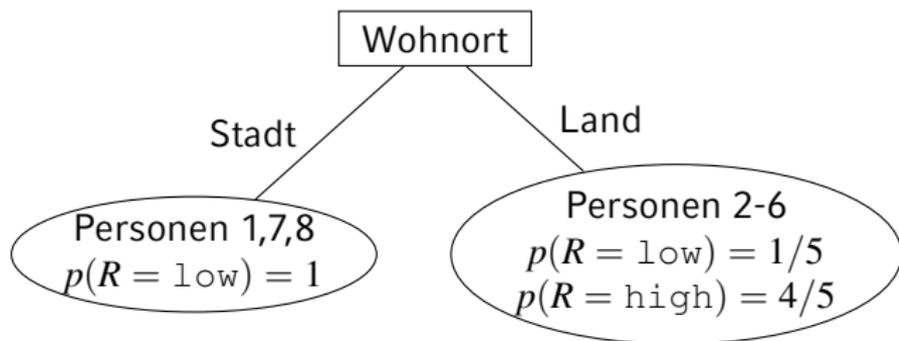
$$\text{Entropie}(T_2) \approx 0.722$$

Informationsgewinn für das Attribut Wohnort:

$$\begin{aligned} & \text{Informationsgewinn}(T, \text{Geschlecht}) \\ &= 1 - \left(0 + \frac{5}{8} \cdot 0.722 \right) \\ &\approx 0.55 \end{aligned}$$

Gewinn maximal für Attribut Wohnort.





Rechter Zweig:

$$\text{Entropie}(T) = - \left(\frac{1}{5} \log \frac{1}{5} + \frac{4}{5} \log \frac{4}{5} \right) \approx 0.722$$

Informationsgewinn im Attribut Zeit: Entropie für T_1

1-2 Jahre: $T_1 =$ Personen 4,6

$$p(R = \text{high}) = 1$$

$$\text{Entropie}(T_1) = 0$$

Informationsgewinn im Attribut Zeit: Entropie für T_2
2-7 Jahre: $T_2 = \text{Person 2}$

Aufgabe 9-1

Aufgabe 9-2

$$p(R = \text{high}) = 1$$

$$\text{Entropie}(T_2) = 0$$

Informationsgewinn im Attribut Zeit: Entropie für T_3
> 7 Jahre: $T_3 =$ Personen 3,5

$$p(R = \text{low}) = \frac{1}{2}$$

$$p(R = \text{high}) = \frac{1}{2}$$

$$\begin{aligned} \text{Entropie}(T_3) &= - \left(\frac{1}{2} \log \frac{1}{2} \right) \cdot 2 \\ &= 1 \end{aligned}$$

Informationsgewinn für das Attribut Zeit:

$$\begin{aligned} & \text{Informationsgewinn}(T, \text{Zeit}) \\ &= \text{Entropie}(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} \text{Entropie}(T_i) \\ &= 0.722 - \left(\frac{2}{5} \cdot 0 + \frac{1}{5} \cdot 0 + \frac{2}{5} \cdot 1 \right) \\ &\approx 0.322 \end{aligned}$$

Informationsgewinn im Attribut Geschlecht: Entropie für T_1
m: $T_1 = \text{Personen } 2,5,6$

$$p(R = \text{high}) = 1$$

$$\text{Entropie}(T_1) = 0$$

Informationsgewinn im Attribut Geschlecht: Entropie für T_2
w: $T_2 =$ Personen 3,4

$$p(R = \text{low}) = \frac{1}{2}$$

$$p(R = \text{high}) = \frac{1}{2}$$

$$\text{Entropie}(T_2) = 1$$

Informationsgewinn für das Attribut Geschlecht:

$$\begin{aligned} & \text{Informationsgewinn}(T, \text{Geschlecht}) \\ &= \text{Entropie}(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} \text{Entropie}(T_i) \\ &= 0.722 - \left(\frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 1 \right) \\ &\approx 0.322 \end{aligned}$$

Gleicher Gewinn in beiden. Egal, welches verwendet wird.

