Data Mining Tutorial

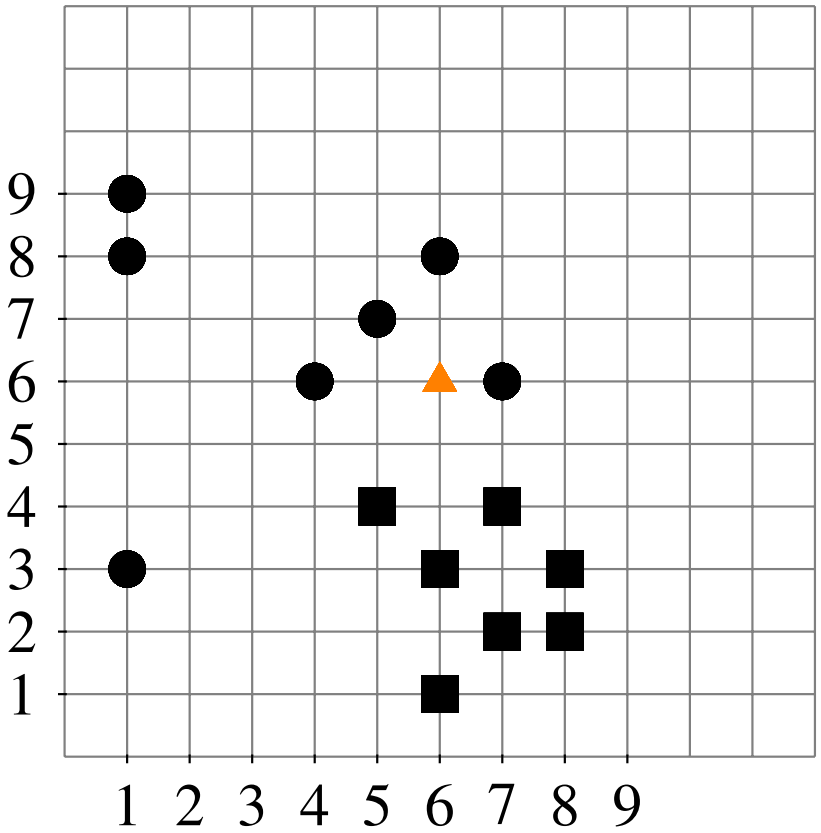
Klassifikation II

Erich Schubert, Arthur Zimek

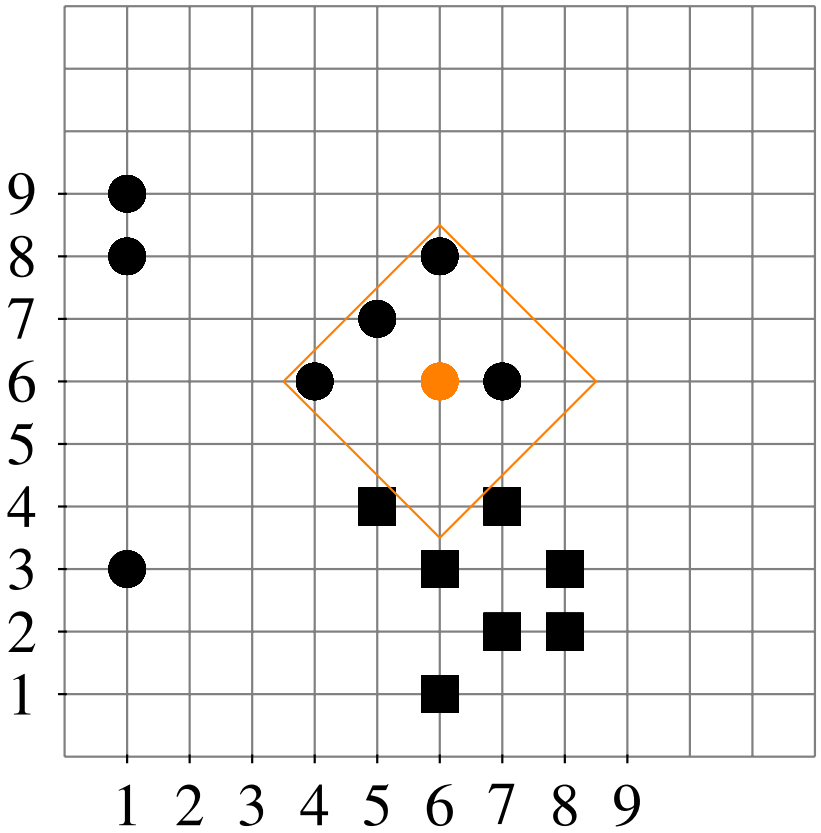
Ludwig-Maximilians-Universität München

2014-07-01 — KDD Übung

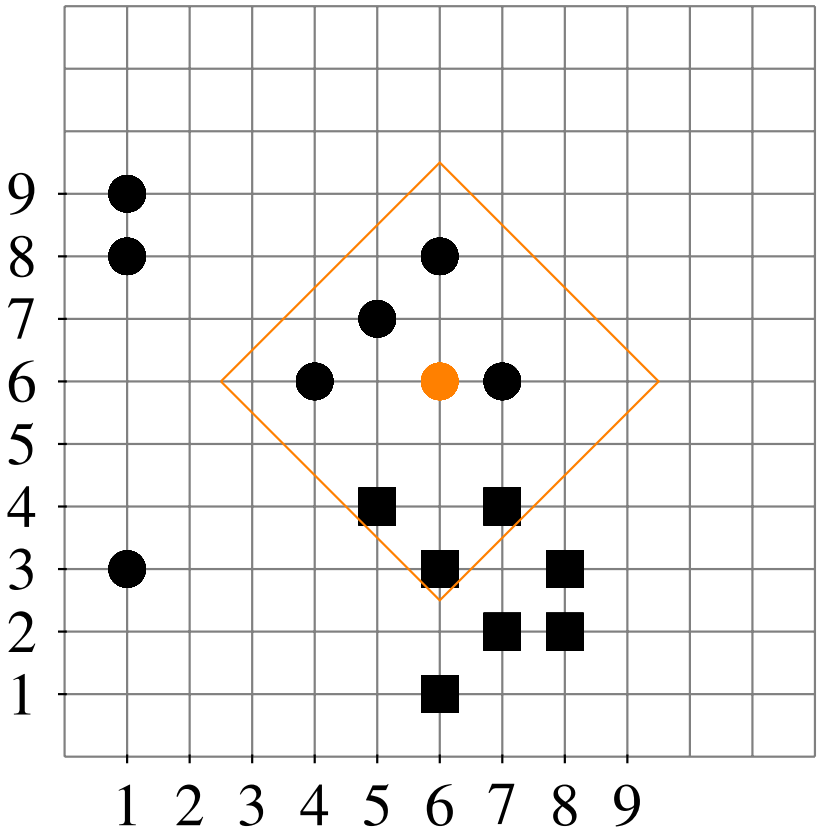
Nächste-Nachbarn Klassifikation



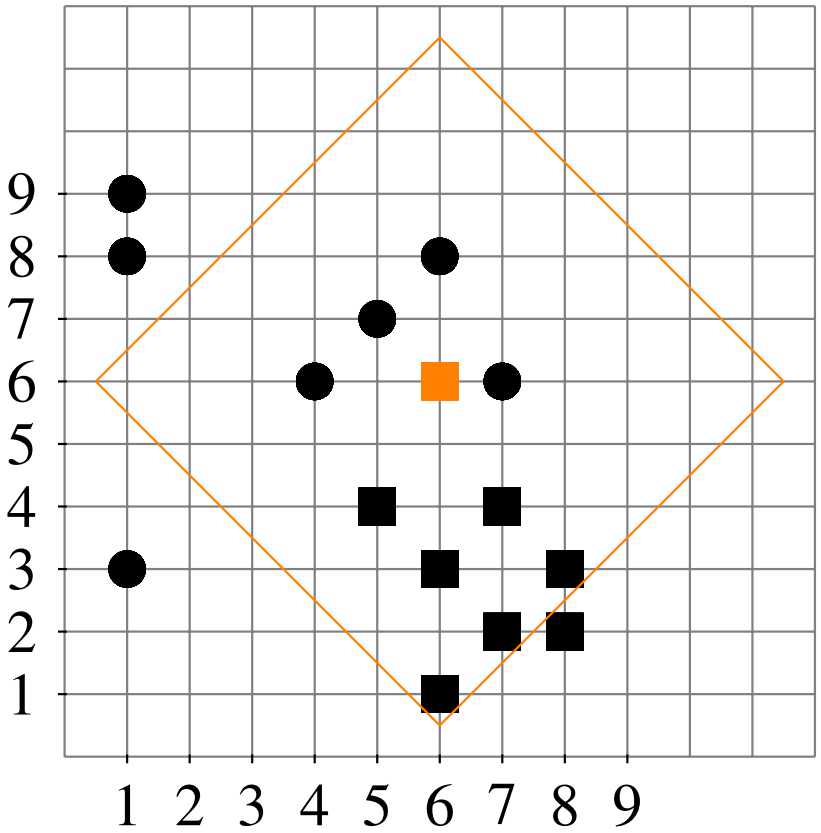
Nächste-Nachbarn Klassifikation



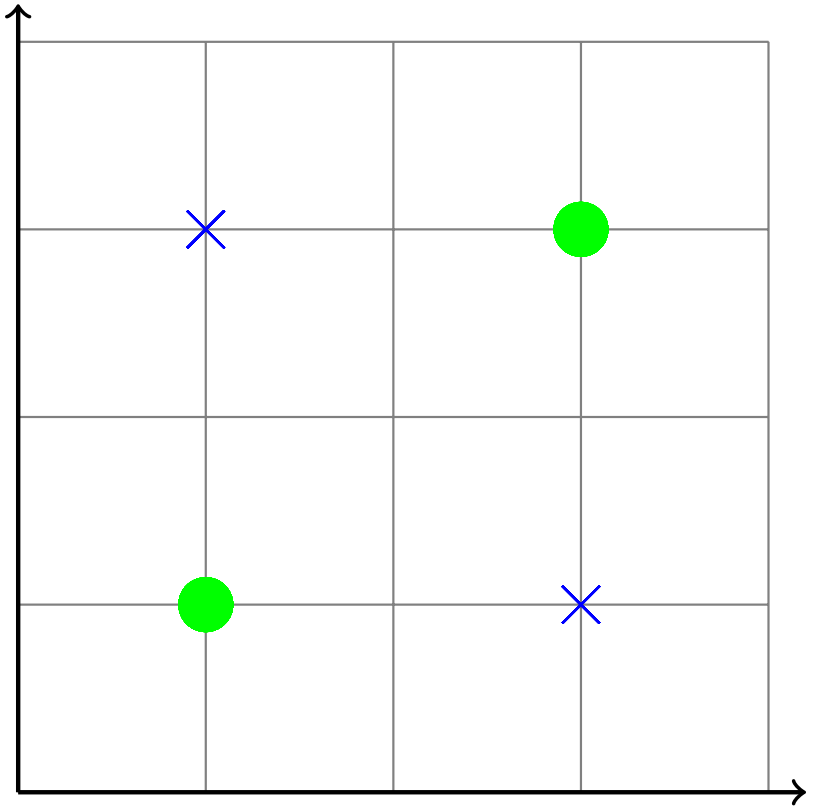
Nächste-Nachbarn Klassifikation



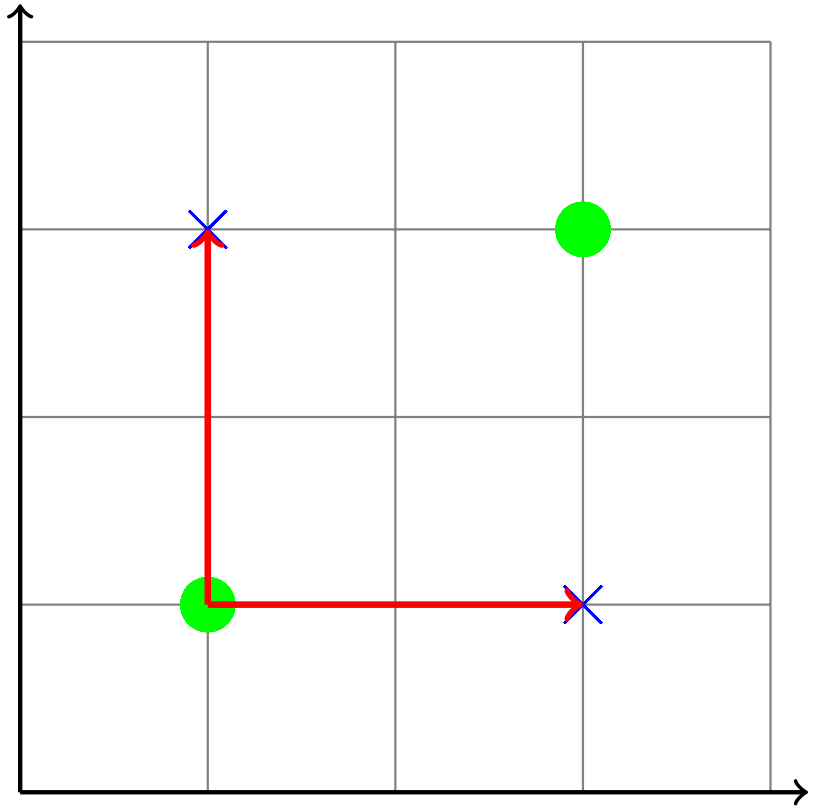
Nächste-Nachbarn Klassifikation



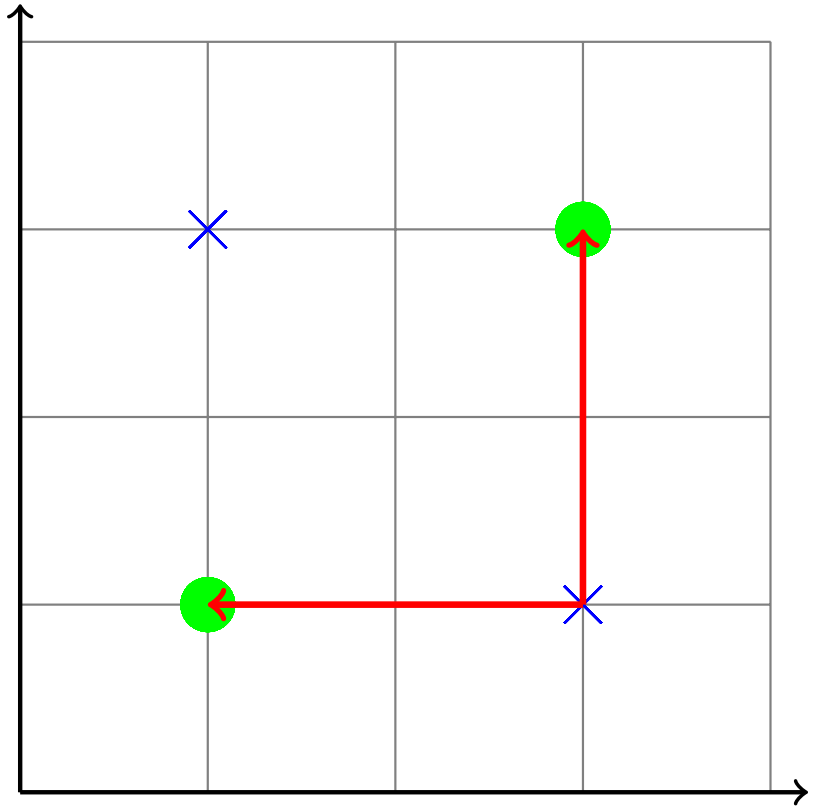
Nächste-Nachbarn Klassifikation



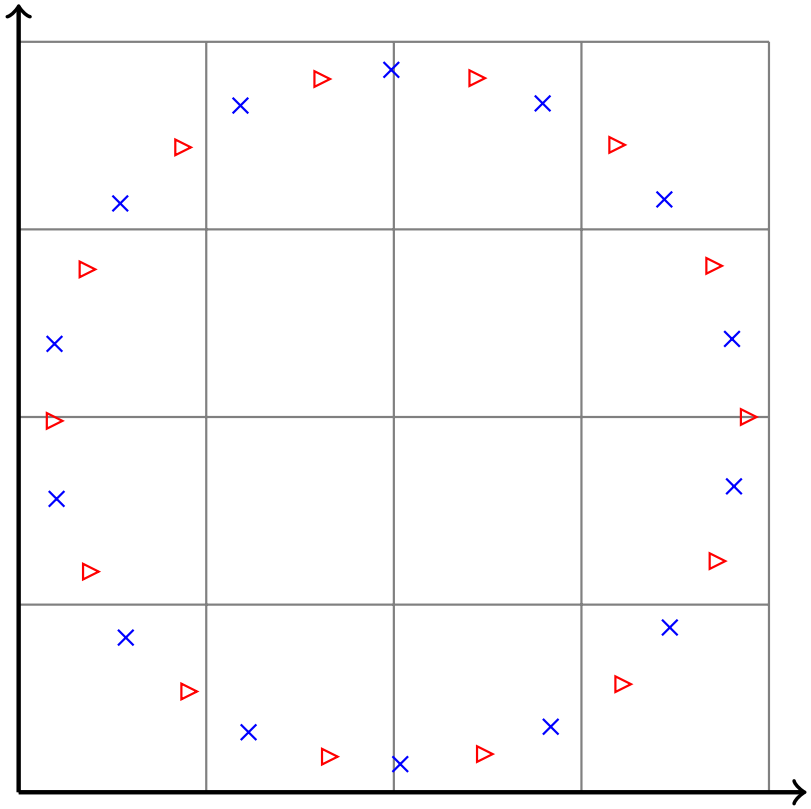
Nächste-Nachbarn Klassifikation



Nächste-Nachbarn Klassifikation



Nächste-Nachbarn Klassifikation



Entscheidungsbäume

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

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

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

Entscheidungsbäume

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

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

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

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

Entscheidungsbäume

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

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

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

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

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

Entscheidungsbäume

Informationsgewinn im Attribut Zeit: Entropie für T_1

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

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

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

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

Entscheidungsbäume

Informationsgewinn im Attribut Zeit: Entropie für T_2

2-7 Jahre: $T_2 =$ Personen 2,7,8

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

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

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

Entscheidungsbäume

Informationsgewinn im Attribut Zeit: Entropie für T_3

> 7 Jahre: $T_3 =$ Personen 3,5

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

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

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

Informationsgewinn für das Attribut Zeit:

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

Informationsgewinn im Attribut Geschlecht: Entropie für T_1

m: $T_1 =$ Personen 1,2,5,6,8

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

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

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

Informationsgewinn im Attribut Geschlecht: Entropie für T_2

w: $T_2 =$ Personen 3,4,7

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

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

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

Informationsgewinn für das Attribut Geschlecht:

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

Entscheidungsbäume

Informationsgewinn im Attribut Wohnort: Entropie für T_1

Stadt: $T_1 =$ Personen 1,7,8

$$p(R = \texttt{low}) = 1$$
$$p(R = \texttt{high}) = 0$$
$$\mathit{Entropie}(T_1) = 0$$

Informationsgewinn im Attribut Wohnort: Entropie für T_2

Land: $T_2 =$ Personen 2,3,4,5,6

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

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

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

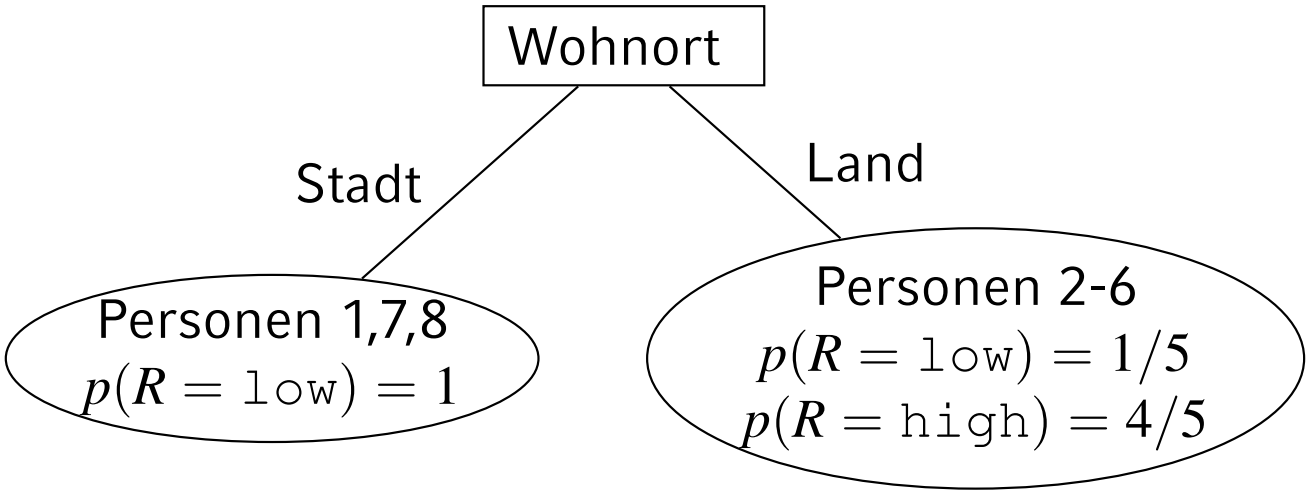
Entscheidungsbäume

Informationsgewinn für das Attribut Wohnort:

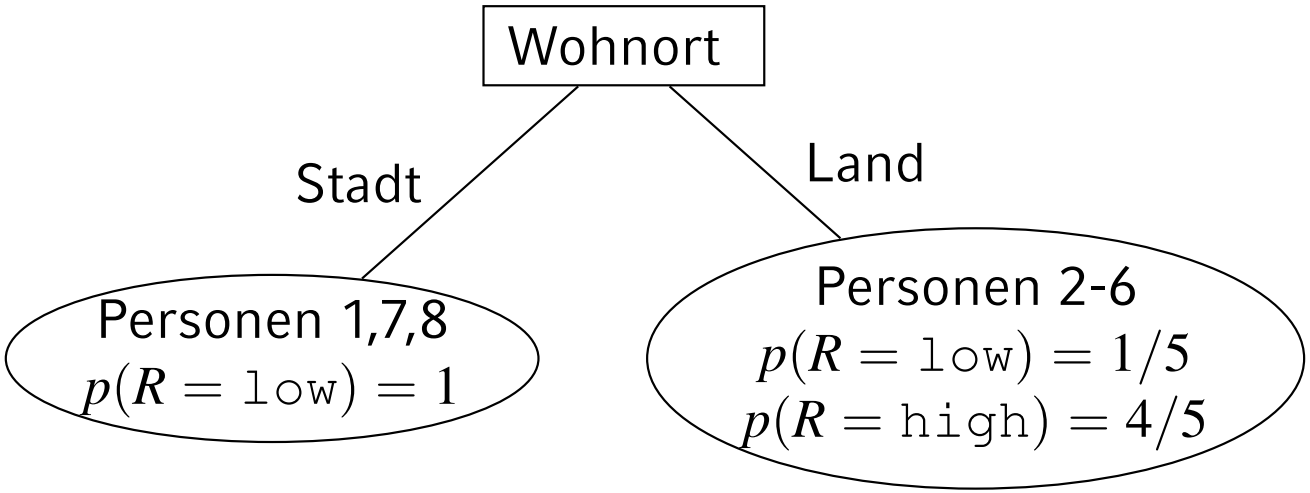
$$\begin{aligned} & \mathit{Informationsgewinn}(T, \mathit{Geschlecht}) \\ &= 1 - \left(0 + \frac{5}{8} \cdot 0.722\right) \\ &\approx 0.55 \end{aligned}$$

Gewinn maximal für Attribut Wohnort.

Entscheidungsbäume



Entscheidungsbäume



Rechter Zweig:

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

Informationsgewinn im Attribut Zeit: Entropie für T_1

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

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

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

Informationsgewinn im Attribut Zeit: Entropie für T_2

2-7 Jahre: T_2 = Person 2

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

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

Entscheidungsbäume

Informationsgewinn im Attribut Zeit: Entropie für T_3

> 7 Jahre: $T_3 =$ Personen 3,5

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

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

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

Informationsgewinn für das Attribut Zeit:

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

Entscheidungsbäume

Informationsgewinn im Attribut Geschlecht: Entropie für T_1

m: $T_1 =$ Personen 2,5,6

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

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

Informationsgewinn im Attribut Geschlecht: Entropie für T_2

w: $T_2 =$ Personen 3,4

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

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

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

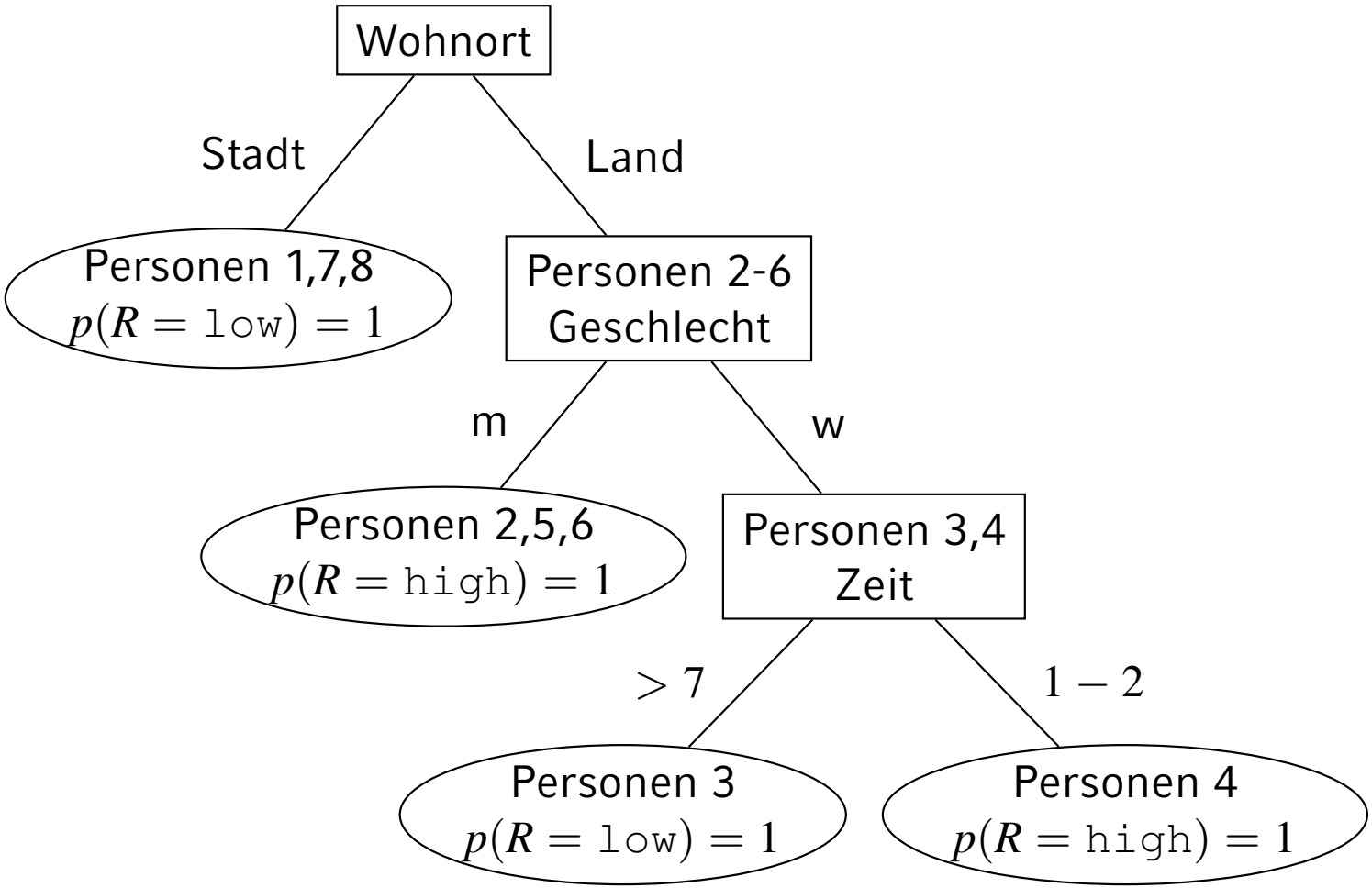
Entscheidungsbäume

Informationsgewinn für das Attribut Geschlecht:

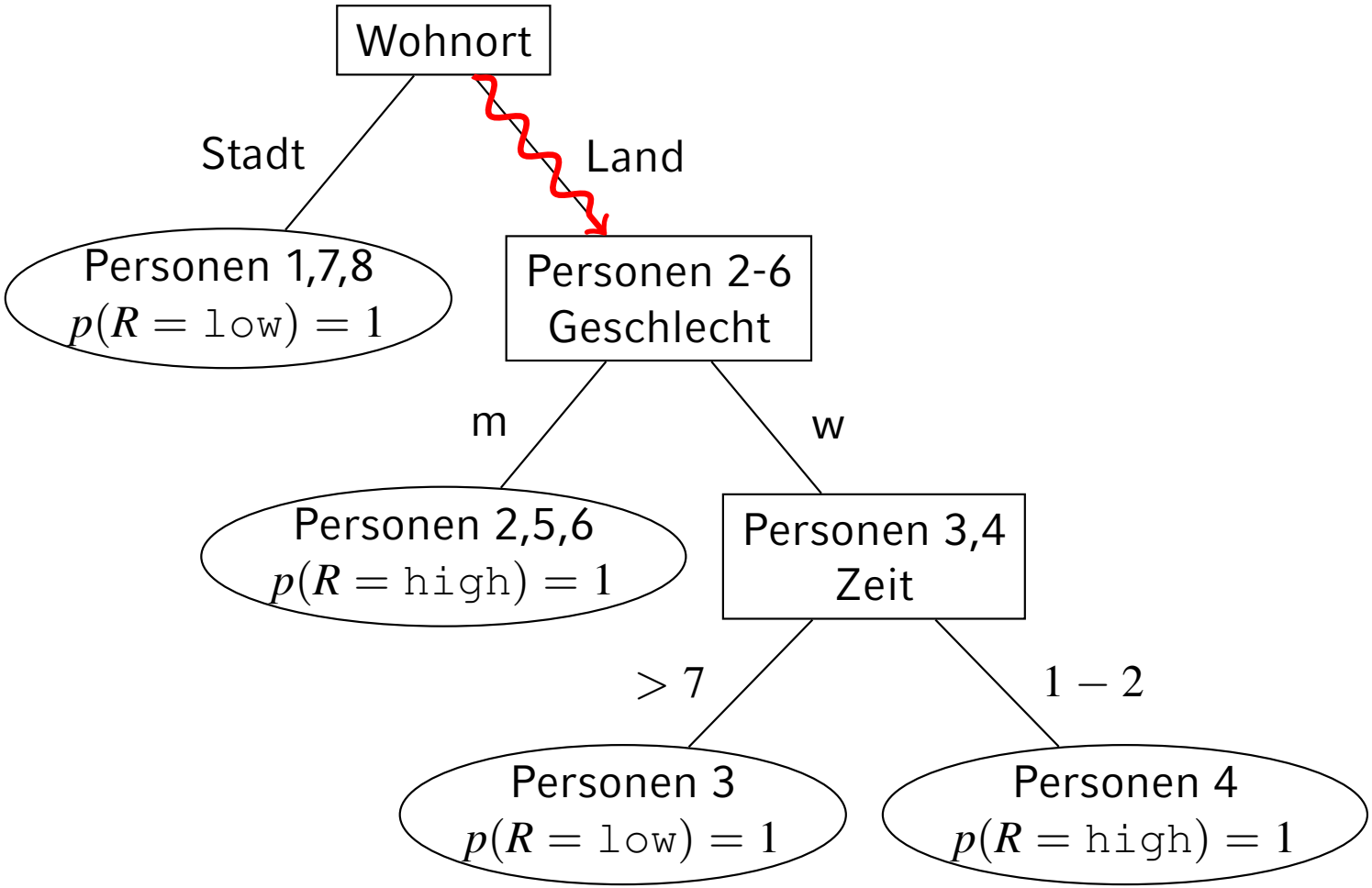
$$
\begin{aligned}
&\mathit{Informationsgewinn}(T, \mathit{Geschlecht}) \\
&= \mathit{Entropie}(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} \mathit{Entropie}(T_i) \\
&= 0.722 - \left(\frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 1 \right) \\
&\approx 0.322
\end{aligned}
$$

Gleicher Gewinn in beiden. Egal, welches verwendet wird.

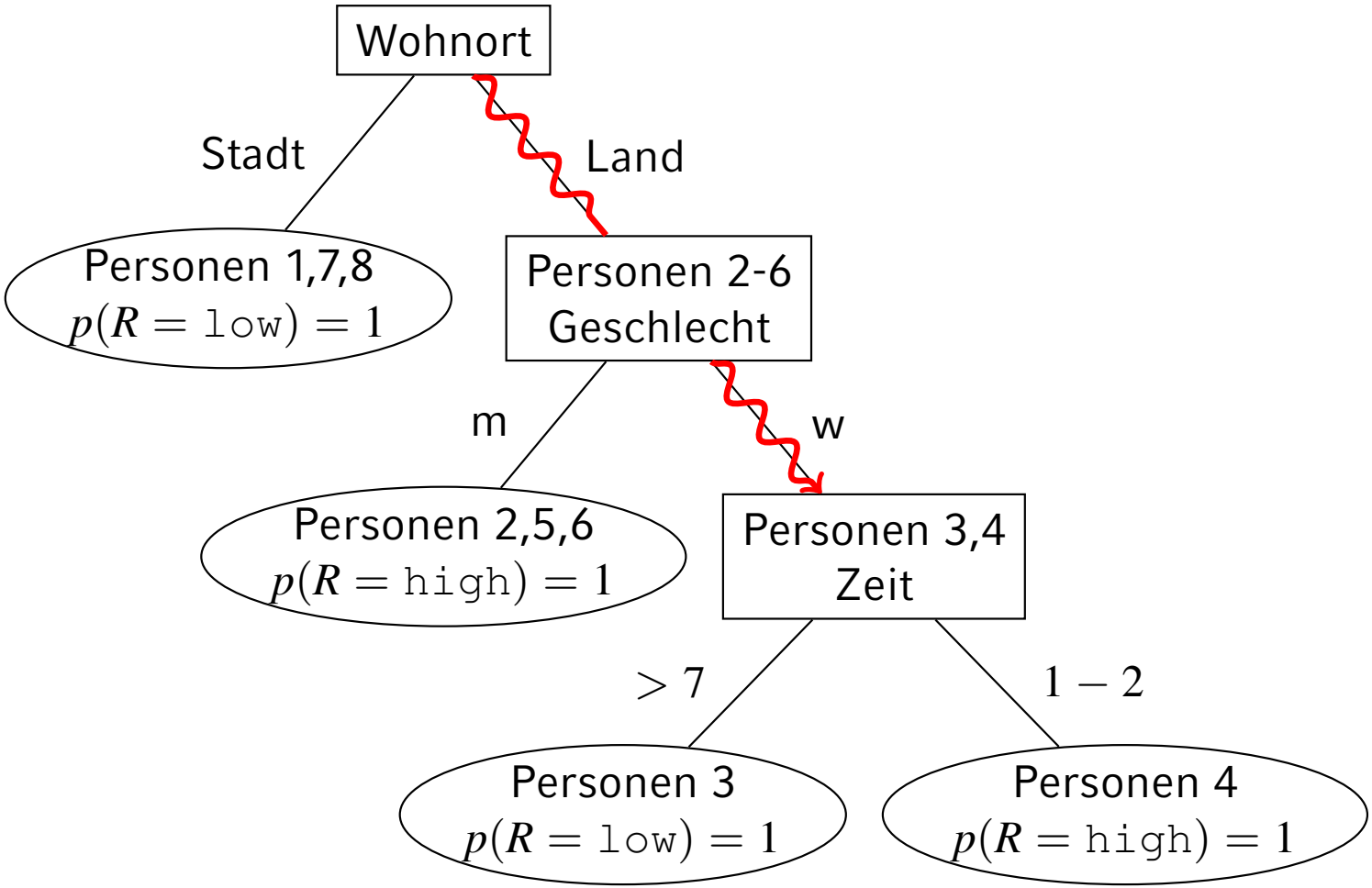
Entscheidungsbäume



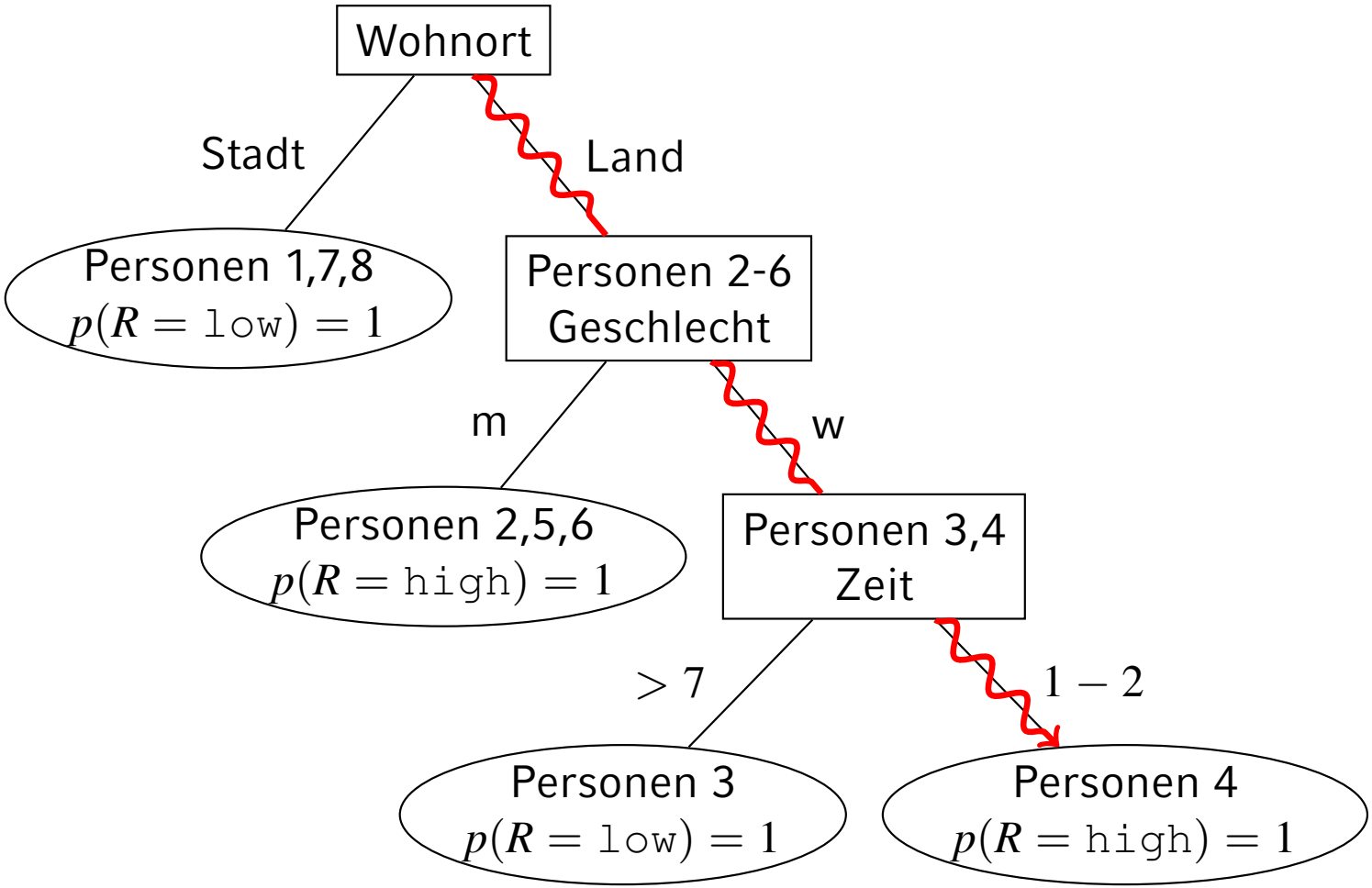
Entscheidungsbäume



Entscheidungsbäume



Entscheidungsbäume



Entscheidungsbäume