Blatt 6

16

a) Satz von Bayes: $P(F|W) = \frac{P(W|F) \cdot P(F)}{P(W)}$

Bedingte Wahrscheinlichkeit: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$P(A \cap B)$: Wahrscheinlichkeit, dass A und B gemeinsam auftreten.

$$P(F|W) = \frac{P(W \cap F)}{P(W)} = \frac{\frac{P(W \cap F)}{P(F)} \, P(F)}{P(W)} = \frac{P(W|F) \cdot P(F)}{P(W)}$$

Der Satz von Bayes folgt direkt aus der Definition der bedingten Wahrscheinlichkeit.

b) Zunächst die Wahrscheinlichkeiten für die gegebenen Ereignisse einzeln:

$P_{\text{starker Wind}} = 1/3$

$P_{\text{hohe Feuchtigkeit}} = 1/3$

$P_{\text{kalte Temperatur}} = 1/3$

$P_{\text{sonnig}} = 2/9$

$\Rightarrow P(W|F) = \frac{2}{243} \approx 8{,}23 \cdot 10^{-3}$

Die gesamte Wahrscheinlichkeit, dass Fußball gespielt wird ist $\frac{36}{56} = \frac{9}{14}$.

Nun muss noch berechnet werden wie wahrscheinlich es ist, dass die Wetterbedingungen des heutigen Tages eintreffen. Das ist dann $P(W)$

$P_{\text{starker Wind}} = 6/14$

$P_{\text{hohe Feuchtigkeit}} = 7/14$

$P_{\text{kalte Temperatur}} = 6/14$

$P_{\text{sonnig}} = 3/14$

$\Rightarrow P(W) = \frac{27}{1372} \approx 1{,}967 \cdot 10^{-2}$

Damit muss die berechnete Wahrscheinlichkeit fürs Fußballspielen normiert werden.

$$P(F|W) = \frac{P(W|F) \cdot P(F)}{P(W)} = \frac{\frac{2}{243}}{\frac{27}{1372}} \cdot \frac{9}{14} = \frac{196}{729} \approx 0{,}2689$$

c)

$P_{\text{Wind}} = 2/3$

$P_{\text{Feucht}} = 1/3$

$P_{\text{heiß}} = 0$

$P_{\text{Anblick}} = 2/9$

$\Rightarrow P(W|F) = 0$

Damit wird auch $p(F|W) = 0$ sein. Das Problem liegt darin, dass in dem vorliegenden Datensatz noch nie bei heißem Wetter gespielt wurde und somit die Wahrscheinlichkeit dafür, bei heißem Wetter zu spielen 0 ist.

Man kann das Problem durch einen größeren Datensatz behoben werden, in dem dann auch etwas unwahrscheinlichere Ereignisse auftreten.

17

a)

$$H(Y) = \sum_{z \in Z} P(Y=z) \log_2 P(Y=z)$$

Die möglichen Ereignisse sind "Fußball spielen" und "nicht Fußball spielen". Es ergibt sich also

$$H(\text{Fußball}) = -\left(\frac{n_{\text{Fußball}}}{n} \log_2\left(\frac{n_{\text{Fußball}}}{n}\right) + \frac{n_{\text{kein Fußball}}}{n} \log_2\left(\frac{n_{\text{kein Fußball}}}{n}\right)\right)$$

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

$$\approx 0{,}94029$$

b)

Für die Berechnung des Informationsgewinns braucht man die Entropie nach einem Schnitt bzgl. des Attributs Wind. Allgemein ist die Entropie nach einem Schnitt

$$H(Y|X) = \sum_{m \in M} P(X=m)\, H(Y|X=m)$$

$$= -\sum_{m \in M} P(X=m) \sum_{z \in Z} P(Y=z|X=m) \log_2 P(Y=z|X=m)$$

In diesem Fall wird das zu

$$H(\text{Fußball}|X) = \frac{n_{\text{wind}}}{n} H(\text{Fußball}|\text{wind}) + \frac{n_{\text{kein wind}}}{n} H(\text{Fußball}|\text{kein Wind})$$

$$H(\text{Fußball}|\text{Wind}) = -\left(\frac{n_{\text{Fußball, wind}}}{n_{\text{wind}}} \log_2\left(\frac{n_{\text{Fußball, wind}}}{n_{\text{wind}}}\right) + \frac{n_{\text{kein Fußball, wind}}}{n} \log_2\left(\frac{n_{\text{kein Fußball, wind}}}{n}\right)\right)$$

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

$$= 1$$

$$H(\text{Fußball} \mid \text{keinWind}) = -\left(\frac{n_{\text{Fußball, keinWind}}}{n_{\text{keinWind}}} \log_2\left(\frac{n_{\text{Fußball, keinWind}}}{n_{\text{keinWind}}}\right) + \frac{n_{\text{keinFußball, keinWind}}}{n_{\text{keinWind}}} \log_2\left(\frac{n_{\text{kein Fußball, keinWind}}}{n_{\text{keinWind}}}\right)\right)$$

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

$$\approx 0{,}811278$$

$$\Rightarrow H(\text{Fußball}, X) = \frac{6}{14} H(\text{Fußball} \mid \text{Wind}) + \frac{8}{14} H(\text{Fußball} \mid \text{kein Wind})$$

$$\approx 0{,}89216$$

Der Informationsgewinn nach einem Schnitt auf dem Attribut Wind ist also:

$$IG(\text{Fußball} \mid \text{Wind}) = H(\text{Fußball}) - H(\text{Fußball} \mid X) \approx 0{,}04813$$

Rest: siehe Jupyter Notebook