6. Übungsblatt

Ksenia Klassen
ksenia.klassen@udo.edu

Dag-Björn Hering
dag.hering@udo.edu

Henning Ptaszyk
henning.ptaszyk@udo.edu

20. Dezember 2016

1 Aufgabe 1

1.1 a)

Haben die Attribute stark verschiedene Größenordnungen, ist es sehr wichtig die Attribute zu normieren. So wird vermieden ,dass Attribute, die eine vergleichsweise größere Größenordnung haben, stärker berücksichtigt werden als andere. Dies würde ohne Normierung passieren, da Abstände gebildet werden.

1.2 b)

Der **k-NN-Algorithmus** speichert beim Lernen einfach die Abständsvektoren der Trainingsdaten ab. Da somit also eigentlich nichts mit den Trainingsdaten passiert kann der Algorithmus als **"lazy-learner"** bezeichnet werden. Somit sind beschränkt sich die Laufzeit beim Lernen auf die Zeit, die benötigt wird um die Trainingsdaten abzuspeichern. In der Anwendungsphase müssen jeweils die Abstände der zu klassifizierenden Daten zu den Trainingsdaten bestimmt werden und anschließend noch sortiert werden. Diese Eigenschaften unterscheiden sich stark von anderen Algorithmen wie zB. einem **Random-Forest**, der mehr Aufwand in das Lernen steckt und dafür beim klassifizieren schneller ist.

1.3 c)

Siehe **aufgabe1.py**.

1.4 d)

Für die zu ermittelnden Größen ergeben sich:

$$
\begin{aligned}
\text{Reinheit} &= 0{,}645 \\
\text{Effizienz} &= 0{,}942 \\
\text{Signifikanz} &= 0{,}314
\end{aligned}
$$

1.5 e)

Nach Logarithmieren der Hits ergibt sich:

$$
\begin{aligned}
\text{Reinheit} &= 0{,}645 \\
\text{Effizienz} &= 0{,}942 \\
\text{Signifikanz} &= 0{,}314
\end{aligned}
$$

Das Ergebnis verändert sich nicht. Obwohl man hätte erwarten können dass es sich verbessert, weil die **Anzahl Hits** jetzt in die selbe Größenordnung fällt wie **x** und **y**.

1.6 f)

Bei Verwendung von 20 anstatt von 10 nächsten Nachbarn, ergeben sich:

$$
\begin{aligned}
\text{Reinheit} &= 0{,}633 \\
\text{Effizienz} &= 0{,}899 \\
\text{Signifikanz} &= 0{,}300
\end{aligned}
$$

Es ist zu erkennen, dass sich das Ergebnis leicht verschlechtert wenn zu viele Nachbarn hinzugezogen werden.

Aufgabe 3:

a) Für die Entropie gilt:

$$I(p,n) = \frac{-p}{p+n} \log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2\left(\frac{n}{p+n}\right)$$

$p = 9$ (Anzahl: Fußball = True)

$n = 5$ (Anzahl: Fußball = False)

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

$$\approx 0{,}94$$

b) $E(a) = \sum_{i=1}^{m} \frac{p_i + n_i}{p+n} I(p_i, n_i)$ (Informationsgewinn)

$gain(a) = I(p,n) - E(a)$ (Informationsgehalt)

$m = 2$

$p_1 = 3$ (~~Körnd~~ Anzahl: Wind = True & Fußball = True)

$n_1 = 3$ (Anzahl: Wind = True & Fußball = False)

$p_2 = 6$ (Anzahl: Wind = False & Fußball = True)

$\cancel{p}n_2 = 2$ (Anzahl: Wind = False & Fußball = False)

$$\Rightarrow gain(Wind) = 0{,}94 - \left(\frac{6}{14}\left(-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right) + \frac{8}{14}\left(-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)\right)\right)$$

$$\approx 0{,}048$$

Abbildung 1

.

c) In den Abbildungen 2- 4 ist der Informationsgewinn in Abbhängigkeit der jeweiligen Schnitte auf den unterschiedlichen Attributen aufgetragen.

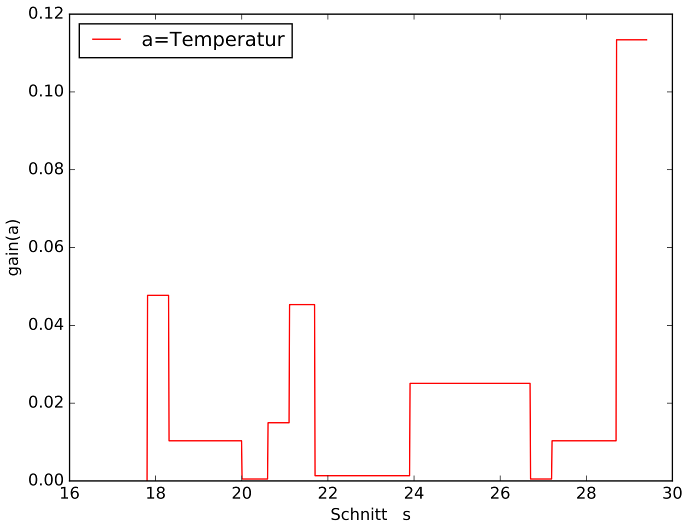

Abbildung 2: Der Informationsgewinn gain(a) in Abhängigkeit von dem Schnitt s auf dem Attribut a=Temperatur .

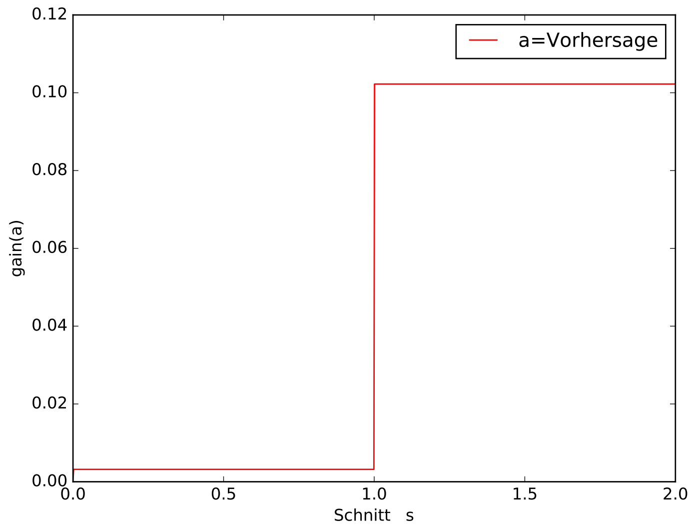

Abbildung 3: Der Informationsgewinn gain(a) in Abhängigkeit von dem Schnitt s auf dem Attribut a=Wettervorhersage .

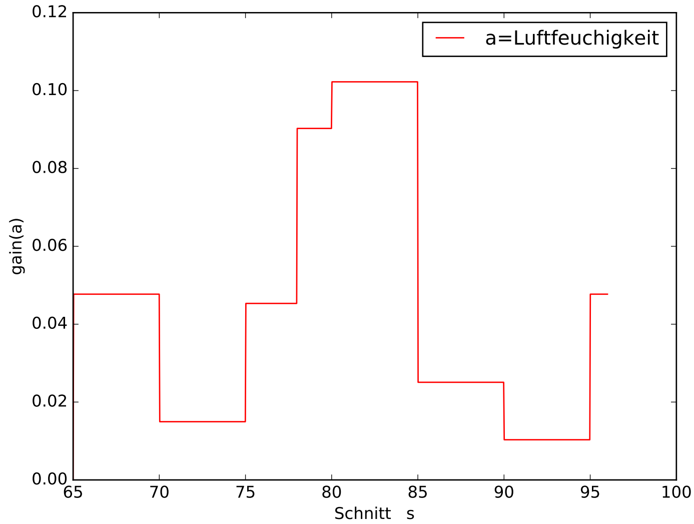

Abbildung 4: Der Informationsgewinn gain(a) in Abhängigkeit von dem Schnitt s auf dem Attribut a=Luftfeuchigkeit .

d) Ein Schnitt s auf dem Attribut Temperatur liefert für

$$s = 28,7 \tag{1}$$

den größten Informationsgewinn gain(Temperatur) von:

$$\text{gain(a)} \approx 0,11 \tag{2}$$

und eignet sich somit am besten zum Trennen der Daten.