Blatt 10

Von Marian Bruns und Kai Brügge

Aufgabe 1 KNN

a)

Bei stark verschiedenen Größenordnungen der Attribute werden die Attribute unterschiedlich stark berücksichtigt. Um dem gegenzusteuern kann man Gewichtungsfunktionen auf die Attribute anwenden.

b)

Weil die Modellbildung erst zum Zeitpunkt der Anfrage stattfindet.

Siehe auch Wikipedia

c)

Ein tatsächlicher Vergleich der Laufzeiten der Algorithmen macht keinen Sinn, da sie unterschiedliche Fragestellungen beantworten und Ein- und Ausgabe nicht gleich sind.

Die Anfrage an einen Random Forest benötigt $\mathcal{O}(L \cdot T)$ vergleiche. Dabei ist L die Anzahl der Bäume im Forest und T die maximale Tiefe der Bäume. Beide Werte bleiben konstant und sind unabhängig von der Dimension und Größe des Datensatzes.

Bei der naiven Implementierung eines k-NN Algorithmuses wie er auf dem Übungsblatt vorgegeben wird beträgt die Laufzeit der Anfrage

$$\mathcal{O}(n \log(n) \cdot f(d)).$$

Wobei $f(d)$ die Distanzfunktion in Abhängigkeit der Dimension d ist und n die Anzahl der Datenpunkte.

$\mathcal{O}(n \log(n))$ ist die Zeit die zum Sortieren (in-situ) gebraucht wird.

Bei einer Implentierung mit einer Suchstruktur für Raumgeordnete Daten lässt sich die Laufzeit jedich erheblich reduzieren. Je nach Dimension wird der k-NN sogar schneller als der Random Forest.

d)

Implemtierung siehe Abgabe. Die Daten werden hier in einem kd-Tree gespeichert um schnelle Suchanfragen zu erlauben.

e)

Confusion Matrix:

$$\begin{bmatrix} 17606 & 2394 \\ 1146 & 8854 \end{bmatrix}$$

Recall: 0.93889 Precision: 0.8803 Significance: 124.5

f)

Confusion Matrix:

$$\begin{bmatrix} 17474 & 2526 \\ 1064 & 8936 \end{bmatrix}$$

Recall: 0.9426 Precision: 0.8737 Significance: 123.6

g)

Confusion Matrix:

$$\begin{bmatrix} 17388 & 2612 \\ 998 & 9002 \end{bmatrix}$$

Recall: 0.94577 Precision: 0.8694 Significance: 123.0

Aufgabe 2 Entscheidungsbäume

Im folgenden sind alle Einheiten Bits.

a)

Wir vermuten das an dieser Stelle die Entropie des Datensatzes bezüglich der Zielvariable berechnet werden soll. Für eine Zufallsvariable Y mit dem Alphabet Z ist die Entropie definiert als der Erwartungswert der Information

$$H(Y) = -\sum_{z \in Z} p_z \log_2 p_z.$$

Im gegebenen Datensatz ist $Z = True, False$. Die Wahrscheinlichkeiten p_z ergeben sich durch Abzählen zu

$$p_{True} = \frac{9}{14}$$

und

$$p_{False} = \frac{5}{14}.$$

Demnach ist die Entropie der Zielvariable in der Tabelle

$$H(Y) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0,9402.$$

b)

Der Information Gain ist definiert als die Änderung der Entropie der Zielvariable Y bedingt einer weiteren Zufallsvariable X

$$IG(Y, X) = H(Y) - H(Y|X).$$

Dabei ist $H(Y|X)$ die bedingte Entropie. Sei X aus dem Alphabet M dann kann die bedingte Entropie geschrieben werden als

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

Sei X im Datensatz beschrieben durch die Spalte mit dem Namen Wind und dem Alphabet $M = \{True, False\}$

$$H(Y|Wind = True) = -\frac{3}{6} \log \frac{3}{6} - \frac{3}{6} \log \frac{3}{6} = 1$$

$$H(Y|Wind = False) = -\frac{6}{8} \log \frac{6}{8} - \frac{2}{8} \log \frac{2}{8} = 0,8112$$

Daraus ergibt sich die bedingte Entropie zu

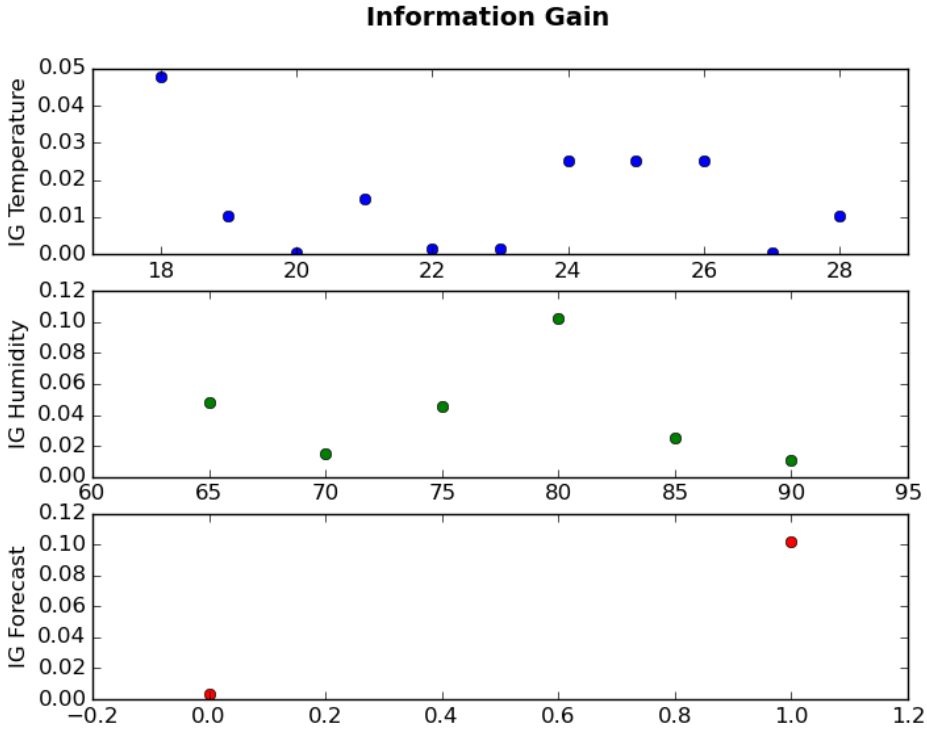
$$H(Y|X) = P(Wind = True) \cdot 1 + P(Wind = False) \cdot 0,8112 = \frac{6}{14} \cdot 1 + \frac{8}{14} \cdot 0,8112 = 0,8921$$

Es folgt der Information Gain zu

$$IG(Y, X) = 0,9402 - 0,8921 = 0,0481$$

c)

Für die Berechnung siehe die Implementation. Die Cuts sind hier als echt größer "<" gemeint.



d)

Der größte Information Gain kann mit der Variable "Wettervorhersage" oder "Luftfeuchtigkeit" erzielt

werden.