

Lernverfahren autonomer Roboter - Übung 8

Tobias Hahn
3073375

January 11, 2017

1 Decision Tree

1.1 Continous data

1.2 Same path twice

1.3 Information gain

Der Gain einer Partitionierung ist gegeben durch die Formel

$$Gain(X, T) = Info(T) - Info(X, T)$$

Um also zu zeigen dass der Gain von der Partitionierung 0 ist, muss gezeigt werden dass $Info(T) = Info(X, T)$. Dafür ist es hilfreich, sich deren Definitionen zu vergegenwärtigen. Die verwendeten Zeichen sind:

- n = Anzahl der negativen Beispiele
- p = Anzahl der positiven Beispiele
- $|T_i|$ = Anzahl der Elemente in Partition i
- $|T|$ = Gesamtanzahl der Elemente
- n_i = Anzahl der negativen Beispiele in Partition i
- p_i = Anzahl der positiven Beispiele in Partition i

$$Info(T) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\left(\frac{p}{p+n} * \log_2\left(\frac{p}{p+n}\right) + \frac{n}{p+n} * \log_2\left(\frac{n}{p+n}\right)\right)$$
$$Info(X, T) = \sum_i \frac{|T_i|}{|T|} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

Wir wissen, dass $\frac{p_i}{p_i+n_i}$ für alle Partitionen gleich ist. Nun müssen wir noch zeigen dass daraus auch folgt dass $\frac{n_i}{p_i+n_i}$ für alle Partitionen gleich ist, dass hätten wir gezeigt dass wir uns das gewichtete Mitteln sparen können, da das gewichtete Mitteln von immergleichen Zahlen immer die gleiche Zahl ergibt. Dies ist jedoch einfach zu zeigen, da $\frac{n_i}{p_i+n_i} = 1 - \frac{p_i}{p_i+n_i}$, d.h. wenn das eine für alle gleich ist ist das andere Verhältnis auch für beide gleich. Damit kann man die zweite Gleichung vereinfachen zu:

$$Info(X, T) = I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

Diese vereinfachte Form sieht nun schon fast so aus wie der Informationsgehalt von $Info(T)$, nur dass statt der Gesamtanzahl von n und p jeweils die Anzahlen in einer der Partitionen stehen. Da wir aber gezeigt haben dass beide Verhältnisse in allen Partitionen gleich sind, so folgt daraus dass das Verhältnis auch in der Summe der Partitionen das gleiche ist. D.h. $Info(T) = Info(X, T)$, und damit $Gain = 0$.

1.4 Decision tree learning