1 Enzscheidungsbäume

Das Maschinelle Lernen durch Entscheidungsbäume läuft so ab, dass Daten nach Ja-Nein-Fragen bezüglich der Features in bestimmte Kindknoten einsortiert werden, bis eine eindeutuge Trennung der Datensets vorliegt. Diese Fragen werden so optimiert, dass der größtmögliche Informationsgewinn dabei entsteht. Praktisch gibt es also für jeden Knoten eine Eingabe und eine Funktion, die es zu maximieren gilt. Damit solche Bäume aber keine allzu-große Tiefe erreichen, wird diese Tiefe begrenzt.
Um nun den Informationsgewinn genauer zu definieren wird der Begriff der Unreinheit genutzt. Unreinheit ist hierbei die ein Maß der Trennung richtiger und falscher Klassifizierung, die später durch Funktionen genauer definiert werden. Der Informationsgewinn berechnet sich also aus dem Verlust der Falsch-Klassifizierungen vom Eltern- zum Kindknoten.

$$IG(f) = I(D_p) - \sum_{j=1}^{n} \frac{N_j}{N_p} I(D_j) \tag{1}$$

Dabei ist der Bruch in der Summe ein Maß für die Aufteilung der Daten in jedem Kindknoten. Eine kleine Anzahl an Klassifizierungen führt also auch zu schlechterer Unreinheit.

Nun gibt es ebenfalls verscheidene Wege die Unreinhet zu messen.

1.1 Die Entropie

Die Entropie ist ein Maß der Verteilung der Daten im Baum und die Etropie ist maximal, wenn die Daten auf alle Knoten gleichmäßig verteilt sind und ist mit $k(i|t)$ als der Anzahl der Daten im t Knoten mit der Klassifizierung i definiert als:

$$I_E = -\sum_{i=1}^{c} k(i|t) \log_2(k(i|t)) \tag{2}$$

1.2 Der Gini-Koeffizient

Der Gini-Koeffizient ist ebenfalls ein Maß der Verteilung der Daten, ganz ähnlich der Entropie, minimiert aber die Wahrscheinlichkeit einer Fehlklassifizierung.

$$I_G = 1 - \sum_{i=1}^{c} k(i|t)^2 \tag{3}$$

1.3 Der Klassifizierungsfehler

Der Klassifizierungsfehler dient dabei eher der Fehlerbehandlung, da es weniger sensitiv als die oben-genannten Alternativen ist.

$$I_K = 1 - \max(k(i|t)) \tag{4}$$

1.4 Random-Forests

Das Random-Forests-Verfahren generiert dabei aus Teilmengen der Trainingsdaten einzelne Entscheidungsbäume. Es werden zufällige Features der Daten gewält und der Entscheidungsbaum teilt sich dann anhand des Merkmales, welches den größten Informationsgewinn zulässt. Am Ende werden dann durch eine Mehrheitsentscheidungen die Merkmale ausgewählt, die den Informationsgewinn maximieren.