Data Mining WS 2012/2013

Institut für Neuroinformatik
Dr. F. Schwenker

6. Aufgabenblatt (Abgabe am 29.01.2013 in der Vorlesung)

10. Aufgabe (6): Entscheidungsbaum

Gegeben sei die folgende Datenmatrix:

M1	M2	M3	M4	Klasse
0	1	1	0	1
1	0	1	0	1
0	0	1	1	1
1	1	1	1	1
1	0	1	1	2
0	0	0	0	2
0	1	0	0	2
1	1	1	0	2

Es gibt also vier Merkmale (M1,M2,M3,M4) und zwei Klassen in diesem Datensatz.
Bestimmen Sie einen binären Entscheidungsbaum (jeder innere Baumknoten hat also genau zwei Kinder), so dass in den Blattknoten jeweils nur Datenpunkte aus einer der beiden Klassen vorkommen.
Verwenden Sie das Maß Q_m (*misclassification impurity*) zur Bewertung der Knoten. Der Gewinn (gain) durch eine Zerlegung einer Region/Knoten R in zwei Regionen/Knoten R_l und R_r wurde in der Vorlesung definiert durch:

$$\Delta Q(R, R_l, R_r) := Q(R) - Q(R_l)p_{R_l} - Q(R_r)p_{R_r}.$$

Hierbei ist $p_{R_l} = |R_l|/|R|$ und $p_{R_r} = |R_r|/|R|$ der Anteil der Datenpunkte von R in R_l bzw. in R_r.

11. Aufgabe (6 Punkte): impurity measures

In der Vorlesung wurden die folgenden drei Maße zur Bewertung der *Impurity* von Knoten in Entscheidungsbäumen bzw. von Regionen/Mengen vorgestellt:

$$Q_m(p_1, \ldots, p_J) := 1 - \max_{j=1}^{J} p_j, \quad \text{misclassification index} \tag{1}$$

$$Q_g(p_1, \ldots, p_J) := 1 - \sum_{j=1}^{J} p_j^2, \quad \text{Gini index} \tag{2}$$

$$Q_e(p_1, \ldots, p_J) := -\sum_{j=1}^{J} p_j \log_2 p_j, \quad \text{entropy index} \tag{3}$$

Hierbei ist J die Anzahl der Klassen und $p_j := n_j/n$ die relative Häufigkeit der Datenpunkte der Klasse j in der betrachteten Region/Menge/Knoten.
Zeigen Sie, dass für die Maße Q_m, Q_g, Q_e die folgenden Eigenschaften gelten:

1. $Q(p_1, \ldots, p_J)$ ist maximal genau dann wenn $(p_1, \ldots, p_J) = (1/J, \ldots, 1/J)$ ist.
2. $Q(p_1, \ldots, p_J)$ ist minimal genau dann wenn $(p_1, \ldots, p_J)$ ein Einheitsvektor e_i ist (d.h. mit genau einer 1 an der Position i und an jeder anderen Position 0).

Für den Fall $J = 2$ (d.h. die Daten sind aus zwei Klassen) sind die oben definierten Maße wegen $Q(p_1, p_2) = Q(p_1, 1 - p_1)$ Funktionen einer Variablen p_1.
Plotten Sie für diesen Fall ($J = 2$) die drei Maße Q_m, Q_g, Q_e.