Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **13** — 20 Punkte

Aufgabe 1. *Gini (-Simpson) Index* 8 P.

Die Entropie einer diskreten Zufallsvariable V mit Wertemenge $\mathcal{V}$ ist definiert als $H(V) = -\sum_{v\in\mathcal{V}} p(V=v)\log p(V=v)$.

(a) Erklären Sie (kurz) wieso $H(V)$ die Ungewissheit über V quantifiziert.

Es sei nun W eine weitere diskreten Zufallsvariable mit Wertemenge $\mathcal{W}$. Gegeben einem Event $W=w$ können wir die bedingte Entropie $H(V|W=w)$, d.h. die Entropie der bedingten Verteilung $V|W=w$ betrachten. Die bedingte Entropie von V gegeben W ist dann der Erwartungswert

$$H(V|W) = \sum_{w\in\mathcal{W}} p(W=w)H(V|W=w) \ .$$

(b) Erklären Sie, wieso die Größe $H(V)-H(V|W)$ auch (erwarteter) Informationszuwachs (engl. information gain) $IG(V,W)$ genannt wird.

(c) Zeigen Sie, dass $IG(V,W) = IG(W,V)$.

(d) Beim Fitten von Entscheidungsbäumen wird oftmals der Informationszuwachs als Verzweigungskriterium verwendet. Erklären Sie, wieso das eine gute Idee ist und wie dabei V und W durch die Trainingsdaten $(x_1,y_1),\dots,(x_n,y_n)$ bestimmt sind, d.h., welche konkreten Schritte notwendig sind um die nächste Verzweigung zu bestimmen. Gehen Sie dabei sowohl auf kategorische als auch numerische Merkmale ein.

(e) Alternativ wird oftmals der Gini-Zuwachs als Verzweigungskriterium verwendet. Analog zum Informationszuwachs hat er die Form $G(V)-G(V|W)$ mit $G(V|W)=\sum_{w\in\mathcal{W}} p(W=w)G(V|W=w)$.
Wie ist $G(V)$ dabei definiert und ergibt das ein sinnvolles Verzweigungskriterium?

Aufgabe 2. *Entscheidungsbäume* 6 P.

Gegeben seien die folgenden Daten:
Klasse 1: $\{(-2,-3),(-2,-2),(-1,-3),(2,3),(1,3),(2,2)\}$
Klasse 2: $\{(-2,3),(-2,1),(-1,3),(2,-3),(1,-3),(2,-2)\}$

(a) Zeichnen Sie die Daten handschriftlich in ein Koordinatensystem ein.

(b) Bestimmen Sie einen Entscheidungsbaum minimaler Tiefe, der die Trainingsdaten korrekt klassifiziert. Erklären Sie Ihr Vorgehen und warum der resultierende Entscheidungsbaum wirklich minimale Tiefe hat. Ergänzen Sie ihre Zeichnung aus Teilaufgabe (a) um die Entscheidungsregionen des Entscheidungsbaums.

(c) Bestimmen Sie einen Entscheidungsbaum der die Trainingsdaten korrekt klassifiziert mithilfe des CART Algorithmus. Nutzen Sie dabei den Gini-Simpson Index als Kriterium. Fertigen Sie nach jeder Iteration eine Zeichnung der Daten und der resultierenden Entscheidungsregionen an.

(d) Vergleichen Sie die Entscheidungsbäume aus Teilaufgabe (b) und (c).

Aufgabe 3. *Regressionsbäume* 6 P.

Gegeben sind die folgenden Datenpaare.

x	0	1	3	4	6
y	1	2	4	2	4

Wir möchten einen Regressionsbaum der mittels des CART-Algorithmus an die Daten anpassen. Wir nutzen dazu den MSE als Verzweigungskriterium und die arithmetischen Mittelwerte $t \in \{\frac{1}{2}(x_i + x_{i+1})\}$ als mögliche Schwellwerte.

(a) Zeichnen Sie die Daten handschriftlich in ein Koordinatensystem. Ergänzen Sie Ihre Zeichnung um alle Funktionsgraphen der Funktionen, die durch einen Regressionsbaum der Tiefe 1 realisiert werden können.

(b) Bestimmen Sie den Regressionsbaum der Tiefe 1 der vom CART-Algorithmus ausgewählt wird. Welchen MSE hat er?

(c) Wiederholen Sie Teilaufgabe (b) für den Regressionsbaum der Tiefe 2.

(d) Bestimmen Sie ausgehend von ihrem Ergebnis aus (c) die resultierenden Entscheidungsbäume der Tiefe 2 auf den folgenden Daten:

(i)

x	0	1	3	4	6
y	-1	0	4	2	4

(ii)

x	0	1	3	5	6
y	1	2	4	2	4

(iii)

x	0	1	3	4	6
y	2	4	8	4	8