Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **12** — 26 Punkte

Aufgabe 1. *Gini (-Simpson) Index* 6 P.

Zeigen Sie die folgenden Eigenschaften des Gini (-Simpson) Index:

(a) Der Gini Index G misst die Wahrscheinlichkeit, dass zwei zufällig ausgewählte Datenpunkte einer unterschiedlichen Klasse angehören. Erläutern Sie, welche Verteilungsannahme mit zufällig gemeint ist.

(b) Folgern Sie daraus, wie die Berechnung des Gini Index auf die Gewichtungen im Adaptive Boosting angepasst werden kann.

(c) Im Falle von n Datenpaaren $(x_1, y_1), \ldots, (x_n, y_n)$ aus zwei Klassen $y_i \in \{0, 1\}$, lässt sich der Gini Index $G = G(r)$ als Funktion des Anteils $r = \frac{\text{Anzahl Daten der Klasse 1}}{\text{Anzahl Daten}}$ ausdrücken. Zeigen Sie: $G(r)$ ist symmetrisch um $r = \frac{1}{2}$, $G(r)$ ist minimal für $r = \frac{1}{2}$ und $G(|r - \frac{1}{2}|)$ ist strikt monoton steigend.

Aufgabe 2. *Entscheidungsbäume* 10 P.

Gegeben sind die folgenden Marktforschungsdaten.

Ort	Tageszeit	Kundentyp	Position	Kauft Frucht?
Supermarkt	Morgen	Stammkunde	Vorne	Ja
Supermarkt	Nachmittag	Stammkunde	Hinten	Nein
Supermarkt	Nachmittag	Neu	Vorne	Ja
Marktstand	Morgen	Stammkunde	Vorne	Ja
Marktstand	Nachmittag	Neu	Hinten	Nein
Marktstand	Abend	Stammkunde	Vorne	Ja
Marktstand	Morgen	Neu	Vorne	Ja
Supermarkt	Abend	Stammkunde	Hinten	Ja
Supermarkt	Morgen	Neu	Vorne	Ja
Supermarkt	Morgen	Stammkunde	Hinten	Nein
Supermarkt	Morgen	Neu	Hinten	Nein
Marktstand	Abend	Neu	Hinten	Ja
Supermarkt	Nachmittag	Neu	Vorne	Nein
Marktstand	Morgen	Stammkunde	Hinten	Ja
Supermarkt	Abend	Stammkunde	Vorne	Ja

Tabelle 1: Datensatz zur Vorhersage von Fruchtkäufen

Bestimmen Sie mittels des CART Algorithmus einen Entscheidungsbaum, der ausgehend von den Merkmalen Ort, Tageszeit, Kundentyp und Position klassifiziert, ob ein Kunde eine bestimmte Fruchtsorte kauft. Der Entscheidungsbaum soll die Tiefe 2 haben und den Gini Koeffizienten als Kriterium nutzen. Interpretieren Sie die Tageszeit als kategorisches Merkmal.

Aufgabe 3. 10 P.

Bestimmen Sie mittels Adaptivem Boosting ein Ensemblemodell auf den Daten aus Tabelle 1, das erneut ausgehend von den Merkmalen Ort, Tageszeit, Kundentyp und Position klassifiziert, ob ein Kunde eine bestimmte Fruchtsorte kauft. Das Ensemblemodell soll aus $M = 3$ Entscheidungsbäumen der Tiefe 1 bestehen.

Geben Sie (für $m \in \{1, 2, 3\}$) die gewichteten Fehler ϵ_m der Basismodelle F_m, die Modellgewichte β_m und die (ungewichteten) Fehler der Ensembles f_m, $m \leq M$ an. Kennzeichnen Sie außerdem in der Tabelle die Datenpunkte, welche von den jeweiligen Basismodellen falsch klassifiziert werden.

Hinweis: Um irrationale Fehler- bzw. Datengewichte zu vermeiden, empfiehlt es sich die äquivalente Gewichtung $w_{m+1,i} = w_{m,i} \exp\left((1 - y_i F_m(\mathbf{x}_i))\beta_m\right)$ zu verwenden.