Universität Salzburg Florian Graf

Machine Learning

Übungsblatt **13** 20 Punkte

Aufgabe 1. *Adaptives Boosting I* 13 P.

Gegeben seien n Daten-Label Paare $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \subset \mathbb{R}^d \times \{-1, 1\}$.
Beim Adaptive-Boosting (AdaBoost) Verfahren werden M Basismodelle $F_m : \mathbb{R}^d \to \{-1, 1\}$ mit Gewichten $\beta_m > 0$ zu einem Ensemblemodell f_M kombiniert,

$$f_M(\mathbf{x}) := \operatorname{sign}\left(\sum_{m=1}^{M} \beta_m F_m(\mathbf{x})\right) \ .$$

Die Basismodelle F_m werden üblicherweise aus einer Klasse $\mathcal{F}$ an Modellen ausgewählt, die nur eine geringe Genauigkeit auf den Trainingsdaten erreichen, dafür aber mit wenig Aufwand an diese angepasst werden können. Dabei werden die F_m iterativ ausgewählt, sodass ein gewichteter Fehler $\operatorname{err}(F, \mathbf{w}_m)$ minimiert wird, d.h.,

$$F_m = \operatorname*{argmin}_{F \in \mathcal{F}} \operatorname{err}(F, \mathbf{w}_m) \ .$$

Dieser gewichtete Fehler $\operatorname{err}(F, \mathbf{w})$ gegeben $\mathbf{w} \in \mathbb{R}^n_+$ (d.h. mit Einträgen $w_1, \dots, w_n > 0$) ist definiert als

$$\operatorname{err}(F, \mathbf{w}) := \frac{1}{Z(\mathbf{w})} \sum_{i=1}^{n} w_i \mathbb{1}_{F(\mathbf{x}_i) \neq y_i}, \qquad Z(\mathbf{w}) := \sum_{i=1}^{n} w_i \ ,$$

und die Fehlergewichte $\mathbf{w}_{m+1} = (w_{m+1,1}, \dots, w_{m+1,n})$ sind

$$w_{m+1,i} := \begin{cases} w_{m,i} \exp(-y_i \beta_m F_m(\mathbf{x}_i)) & m > 0 \\ 1 & m = 0 \end{cases} \ .$$

Außerdem ist $\beta_m := \frac{1}{2} \log\left(\frac{1 - \operatorname{err}(F_m, \mathbf{w}_m)}{\operatorname{err}(F_m, \mathbf{w}_m)}\right)$ und $Z_m = Z(\mathbf{w}_m)$. In dieser Aufgabe zeigen Sie:

Falls die Klasse $\mathcal{F}$ reichhaltig genug ist, sodass für alle Gewichte $\mathbf{w} \in \mathbb{R}^n_+$, ein gewichteter Fehler

$$\min_{F \in \mathcal{F}} \operatorname{err}(F, \mathbf{w}_m) \leq \frac{1}{2} - \gamma \ , \quad 0 < \gamma \leq \frac{1}{2} \ , \tag{1}$$

erreicht werden kann, dann nimmt der (ungewichtete) Fehler

$$\operatorname{err}(f_M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f_M(\mathbf{x}_i) \neq y_i} \leq \exp(-2M\gamma^2) \tag{2}$$

des geboosteten Modells f_M exponentiell mit der Anzahl M der Basismodelle F_m ab.

Gehen Sie dabei folgendermaßen vor.

(a) Zeigen Sie zunächst, dass $\mathbb{1}_{f_M(\mathbf{x}_i) \neq y_i} = \mathbb{1}_{y_i \sum_{m=1}^{M} \beta_m F_m(\mathbf{x}_i) \leq 0}$. Folgern Sie daraus für den ungewichteten Trainingsfehler, dass

$$n \operatorname{err}(f_M) = \sum_{i=1}^{n} \mathbb{1}_{f_M(\mathbf{x}_i) \neq y_i} \leq Z_{M+1} \ .$$

(b) Zeigen Sie außerdem, dass die Normierungskonstanten Z_m für $m > 0$ die folgende Rekursion erfüllen

$$Z_{m+1} = Z_m \sqrt{4 \operatorname{err}(F_m, \mathbf{w}_m)(1 - \operatorname{err}(F_m, \mathbf{w}_m))} = Z_m \sqrt{1 - 4\left(\frac{1}{2} - \operatorname{err}(F_m, \mathbf{w}_m)\right)^2} \ .$$

Hinweis: Betrachten Sie die Fälle $F_m(x_i) = y_i$ und $F_m(x_i) \neq y_i$ getrennt. Nutzen Sie dann die Definition von β_m

(c) Folgern Sie, dass $\mathrm{err}(f_M) \leq \exp(-2M\gamma^2)$. Hinweis: Es gilt $1 - t \leq \exp(-t)\ \forall t \in \mathbb{R}$.

Wir betrachten nun die Gewichte β_m der Basismodelle F_m genauer.

(d) Sie haben bereits gezeigt, dass

$$\mathrm{err}(f_M) \leq \prod_{m=1}^{M} \frac{Z_{m+1}}{Z_m} \quad ,$$

Insbesondere gilt diese Oberschranke an den Fehler unabhängig von der Wahl von β_m. Zeigen Sie, dass $\frac{Z_{m+1}}{Z_m}$ für $\beta_m = \frac{1}{2}\log\left(\frac{1-\mathrm{err}(F,\mathbf{w}_m)}{\mathrm{err}(F,\mathbf{w}_m)}\right)$ minimal ist.

(e) Zeigen Sie, dass diese Wahl von β dazu führt, dass jedes Basismodell F_{m+1} die korrekt und falsch klassifizierten Daten des Modells F_m insgesamt gleich gewichtet, d.h, dass

$$\frac{1}{Z_m}\sum_{i=1}^{n} \mathbb{1}_{F_m(\mathbf{x}_i)=y_i} w_{i,m+1} = \frac{1}{Z_m}\sum_{i=1}^{n} \mathbb{1}_{F_m(\mathbf{x}_i)\neq y_i} w_{i,m+1} \quad .$$

Wieso ist dies kein Widerspruch zu der Grundidee, dass jedes Basismodell F_m die zuvor falsch klassifizierten Daten zusätzlich gewichtet?

Aufgabe 2. *Beispiel* 7 P.

Gegeben sind die folgenden Marktforschungsdaten.

Verkaufsort	Tageszeit	Kundentyp	Produktplatzierung	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	Nein
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 Adaptivem Boosting ein Ensemblemodell auf den Daten aus Tabelle 1, das ausgehend von den Merkmalen Ort, Tageszeit, Kundentyp und Position klassifiziert, ob ein Kunde eine bestimmte Fruchtsorte kauft. Interpretieren Sie die Tageszeit als kategorisches Merkmal. Das Ensemblemodell soll aus $M = 3$ Entscheidungsbäumen der Tiefe 1 bestehen. Wählen Sie in jeder Iteration den Baum, der den Klassifizierungsfehler minimiert.

Geben Sie (für $m \in \{1, 2\}$) 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.

Erklären Sie außerdem, wie im Falle des Gini-Index als Kriterium vorzugehen wäre.

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