Sprachtechnologie

6. Übung

Prof. Tobias Scheffer
Thomas Vanck

Sommer 2010

Ausgabe am: 07.06.10
Besprechung am: 14.06.10

Aufgabe 1 *Vektorraum-Modell*

Entwickeln Sie einen Textklassifikator für das Marsianische. Der Klassifikator soll Pläne für die Invasion der Erde (Klasse +1) von Texten anderen Inhalts (Klasse −1) unterscheiden. Die Liste der relevanten Terme umfasst nur *argh* und *zonk*, die in 79 bzw. 90 von 100 Texten vorkommen.

Als Trainingsmenge liegen vier von SETI abgefangene marsianische Texte vor:

a) „*argh bob argh*“, Klasse +1

b) „*zonk zonk bob*“, Klasse −1

c) „*argh zonk bob*“, Klasse +1

d) „*zonk zonk argh*“, Klasse −1

Bestimmen Sie die TF-IDF-Merkmalsvektoren und repräsentieren Sie diese im Vektorraum-Modell (Hinweis: Verwenden Sie den natürlichen Logarithmus ln statt des Logarithmus zur Basis 10).

Aufgabe 2 *Lineare Klassifikatoren*

Gegeben sind zwei lineare Klassifikatoren $h^{(1)}(x)$ und $h^{(2)}(x)$. Wobei gilt

$$h^{(i)}(x) = \text{sign}(w^{(i)T} \cdot x + w_0^{(i)}), \text{für } i \in \{1, 2\}$$

Berechnen Sie für $w^{(1)} = (1, -1)^T$ und $w_0^{(1)} = 0$, bzw $w^{(2)} = (2, -2)^T$ und $w_0^{(2)} = -\frac{1}{4}$, die Klassifikation der TD-IDF-Vektoren aus Aufgabe 1 per Hand nach und stellen sie das Ergebnis mit Trennebene graphisch dar.

Aufgabe 3 *ROC-Kurve*

Nachdem Sie die Labels in Aufgabe 2 bestimmt haben, erhalten sie fünf weitere entschlüsselte Nachrichten. Die entsprechenden TF-IDF-Vektoren und die dazugehörigen Klassen sind in der folgenden Tabelle dargestellt.

ID	1	2	3	4	5
TF-IDF	$\begin{pmatrix}0.02\\0.99\end{pmatrix}$	$\begin{pmatrix}0.35\\0.94\end{pmatrix}$	$\begin{pmatrix}0.60\\0.80\end{pmatrix}$	$\begin{pmatrix}0.86\\0.50\end{pmatrix}$	$\begin{pmatrix}0.99\\0.09\end{pmatrix}$
Klasse	-1	+1	-1	+1	+1

Geben Sie für Ihre beiden Klassifikatoren $h^{(1)}$ und $h^{(2)}$ jeweils die ROC-Kurve an und bestimmen Sie den AUC-Wert. Welcher Klassifikator ist besser?