

# Sprachtechnologie

## 10. Übung

Prof. Tobias Scheffer  
Uwe Dick

Sommer 2016

Ausgabe am: 27.06.16  
Besprechung am: 04.07.16

### 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-Merkmalvektoren und repräsentieren Sie diese im Vektorraum-Modell.

### Aufgabe 2

*ROC-Kurve*

Nachdem Sie zwei unterschiedliche Klassifikatormodelle gelernt haben, möchten Sie diese evaluieren und vergleichen. Nehmen Sie an, dass Modell 1 einen Gewichtsvektor  $w_1 = \begin{pmatrix} 0.2 \\ -0.15 \end{pmatrix}$  hat und Modell 2  $w_2 = \begin{pmatrix} -1 \\ -3 \end{pmatrix}$ . Sie erhalten den Entscheidungsfunktionswert des  $i$ -ten Modells für ein Textbeispiel  $x$  durch  $f_i(x) = \langle w_i, \text{TF-IDF}(x) \rangle$ . Nachdem Sie die Modelle trainiert 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 die gelernten Klassifikatoren jeweils die ROC-Kurve an und bestimmen Sie den AUC-Wert. Welcher Klassifikator ist besser?

### Aufgabe 3

Geben Sie außerdem jeweils die Precision-Recall-Kurven an. Bestimmen Sie zudem das

---

**Algorithm 0.1** Precision-Recall-Kurve

---

**Ensure:**  $L$  ist Liste der Länge  $N$  von Instanzen  $x$ , absteigend sortiert nach  $\text{Score}(x)$ .

Setze  $TP = 0, FP = 0$ .  $P$  = Anzahl positiver Labels in  $L$ .

**for**  $i = 1..N$  **do**

**if**  $\text{Label}(L[i]) = +1$  **then**

$TP+ = 1$

**else**

$FP+ = 1$

**end if**

    Setze  $\text{Precision} = TP/(TP + FP)$ ,  $\text{Recall} = TP/P$ .

    Verbinde neuen Punkt (Precision,Recall) mit altem Punkt.

**end for**

---

jeweilige maximale F-Measure.