Sprachtechnologie

3. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2016

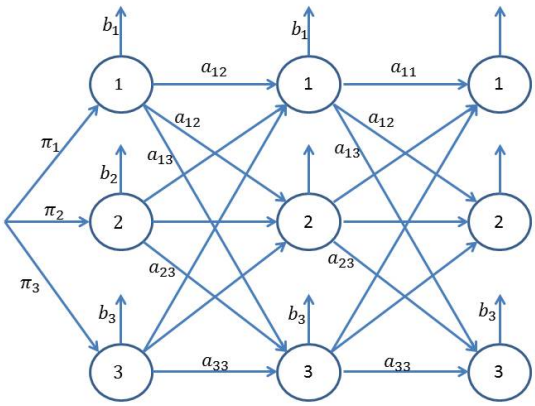
Ausgabe am: 03.05.16
Besprechung am: 10.05.16

Aufgabe 1 *HMM*

Auf einer Ufologen-Konferenz erklärt Erich von Däniken die Regeln des Hochmarsianisch (die offizielle Sprache der MarsianerInnen):

- Zuerst einmal gibt es nur drei Laute: *argh*, *bob* und *zonk*
- Es kommunizieren immer drei Marsianer miteinander.
- Nachdem ein Laut gesagt wurde sagt einer der beiden anderen Marsianer den nächsten Laut mit einer Wahrscheinlichkeit von je 40%, in 20% der Fälle fügt der aktuelle Redner noch einen weiteren Laut hinzu.
- Wer das Gespräch beginnt ist gleichwahrscheinlich.

Sie erinnern sich an Herrn von Däniken als Sie an einem Vorhang vorbeigehen, hinter dem die drei einzigen Marsianer auf der Konferenz gerade ein Gespräch beginnen. Zufällig wissen Sie, dass die drei aus unterschiedlichen Regionen kommen und verschiedene Dialekte sprechen. Der erste benutzt nur die Laute *zonk* und *bob* im Verhältnis $\frac{2}{5}$ zu $\frac{3}{5}$, der zweite nur *argh* und *zonk* im Verhältnis $\frac{1}{11}$ zu $\frac{10}{11}$ und der dritte eine Kombination aller drei Laute *bob*, *argh* und *zonk* in den Verhältnissen $\frac{4}{9} : \frac{2}{9} : \frac{3}{9}$. Sie hören die ersten drei Laute des Gespräches: *zonk*, *bob* und *argh*. Können Sie Herrn von Däniken sagen, welcher Marsianer welchen Laut gesagt hat?
Solch eine Aufgabe lässt sich gut mit einem HMM lösen. Dazu müssen folgende Parameter identifiziert werden: Welche Zustände gibt es? Was sind die möglichen Beobachtungen? Welche Start-, Übergangs- und Beobachtungswahrscheinlichkeiten nehmen Sie an?
Abbildung sollte Ihnen helfen, die Aufgabenstellung, zusammen mit Folie 32, besser zu verstehen. Zuerst einmal sollten Zustände mit Marsianern identifiziert werden, d.h. der Zustand y_1 zum Zeitpunkt $t = 1$ beschreibt, welcher Marsianer zu diesem Zeitpunkt spricht. Es gibt also drei Zustände $Y = \{1, 2, 3\}$. Es gibt auch drei Beobachtungen $X = \{argh, bob, zonk\}$. Die Startwahrscheinlichkeiten π, also etwa $\pi_1 = P(y_1 = 1) = \frac{1}{3}$

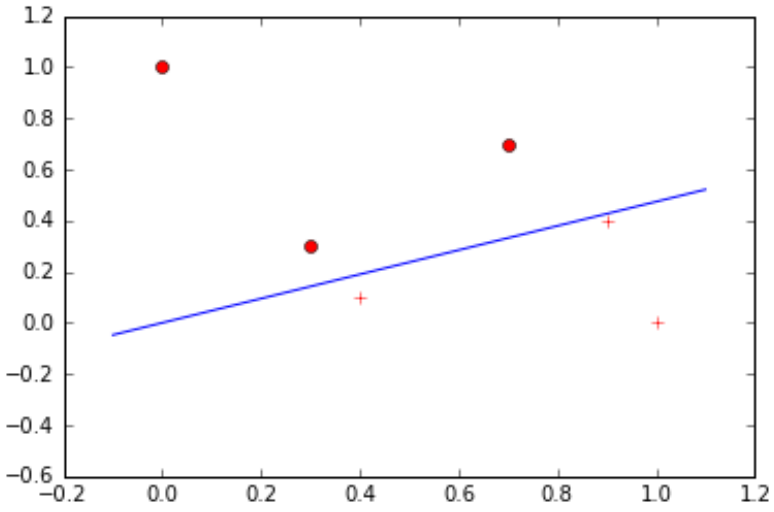


können direkt aus der obigen Beschreibung gelesen werden. Das gleiche gilt für die Übergangswahrscheinlichkeiten, etwa $a_11 = P(y_{t+1} = 1|y_t = 1) = 0.2$, und die Beobachtungswahrscheinlichkeiten, z.B. $b_1(zonk) = \frac{2}{5}$.

1. Bestimmen Sie nun die Likelihood des Modells (Hint: Folien 38ff, Forward-Algorithmus).
2. Welcher Zustand ist der wahrscheinlichste (gegeben die Sequenz) für den zweiten Laut (also zum Zeitpunkt $t = 2$) (Hint: Forward-Backward-Algorithmus)?
3. Bestimmen Sie die wahrscheinlichste Zustandsfolge. (Hint: Viterbi-Algorithmus)

Aufgabe 2 *Lineare Modelle*

Ein Dienstleister hat von einem Kunden den Auftrag bekommen, zu schätzen, wieviele Techblogs sich vorwiegend mit Anwendungen unter Windows beschäftigen. Der Dienstleister geht so vor, dass er jeweils alle Artikel von 500 Techblogs herunterlädt und dann von Mitarbeitern einschätzen lässt, ob sich der entsprechende Blog vorwiegend mit Windows beschäftigt. Anschließend möchte er ein lineares Modell lernen, das dieses Label aufgrund von relativen Häufigkeiten von Wörtern in den Artikeln vorhersagt. Das trainierte Modell soll dann verwendet werden, um für 10000 gecrawlte Blogs das gewünschte Label zu schätzen.
Unglücklicherweise hat der Praktikant einen Fehler begangen und für alle gecrawlten Seiten nur die relativen Häufigkeiten der beiden Wörter Iphone und Android gespeichert. Um seinen Fehler zu kaschieren, trainiert er einfach ein Modell, dass nur diese beiden Wörter



verwendet. Um zu testen, ob seine Notlösung funktioniert, nimmt er die Daten für 6 Blogs und testet den gelernten Klassifikator. Die relativen Häufigkeiten der beiden Wörter auf den 6 Blogs werden in der Matrix $X = (x_1, ..., x_6)^\top$ dargestellt, die entsprechenden Labels im Vektor $Y = (y_1, ..., y_6)^\top$. Das gelernte Modell hat den Parametervektor θ.

$$X = \begin{pmatrix} 1.0 & 0.0 & 0.9 & 0.7 & 0.3 & 0.4 \\ 0.0 & 1.0 & 0.4 & 0.7 & 0.3 & 0.1 \end{pmatrix}^\top \tag{1}$$

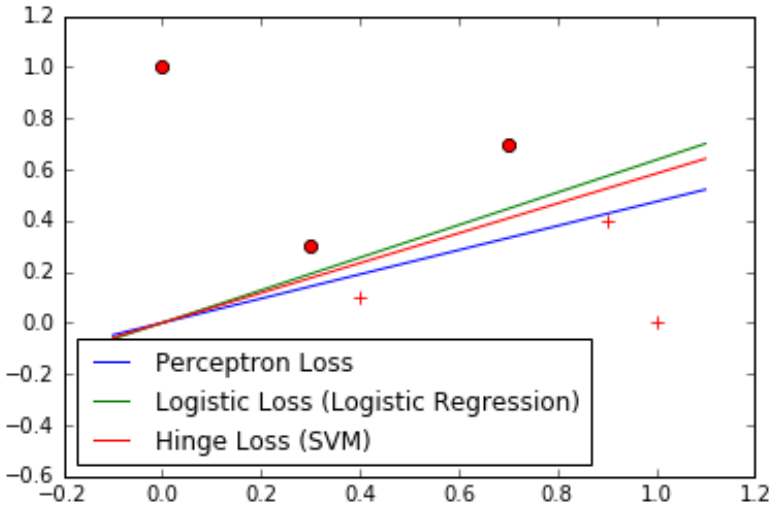
$$Y = (1, -1, 1, -1, -1, 1)^\top \tag{2}$$

$$\theta = (0.9 - 1.9)^\top \tag{3}$$

1. Berechnen Sie die Entscheidungsfunktionswerte (Folien S. 56,57) für alle Blogs.
2. Berechnen Sie dann den 0/1-Fehler des Klassifikators auf den Daten (S.70).
3. Anschließend Berechnen Sie auch den Loss entsprechend den anderen 3 vorgestellten Verlustfunktionen (S. 71-74).
4. Denken sie, dass θ einen guten Klassifikator beschreibt? (Siehe Abbildung)

Der Praktikant ist nun der Meinung, dass der Klasssifikator nicht sonderlich gut ist und trainiert zwei neue Klassifikatoren (S.61ff). Den ersten trainiert er, indem er als Verlustfunktion ℓ den Hinge Loss verwendet, den zweiten mit Logistic Loss. Die folgenden beiden Parametervektoren werden gelernt.

$$\theta_{logistic} = (12.3, -19.3)^\top, \theta_{hinge} = (4.7, -8)^\top \tag{4}$$



Überlegen Sie, warum diese beiden Klassifikatoren (Abbildung) möglicherweise besser sein könnten als die erste Lösung, die durch ein Perzeptron gelernt wurde (Perceptron Loss).

Aufgabe 3 *T9*

Die Eingabehilfe T9 ermöglicht die Texteingabe auf Zifferntastaturen. Diskutieren Sie, wie Sie T9 mit Hilfe eines N-Gramm-Modells oder eines HMM auf Buchstabenebene implementieren können. Das System soll stets das wahrscheinlichste Wort für die getippten Ziffern ermitteln.

Eine naive Implementierung, bei der alle möglichen Buchstabenfolgen $w_1, \dots w_T$ durchsucht werden, hat eine exponentielle Laufzeit in T (Siehe Folie 33). In der Vorlesung wurde erklärt, dass man durch eine geeignete Dekodierung eine Laufzeit von $\mathcal{O}(TV^N)$ erreichen kann, wobei V die Größe des Alphabets ist und N der Parameter des N-Gram-Modells. Geben Sie einen konkreten Algorithmus mit dieser Laufzeit an.