

Musterlösungen zum 4. Übungsblatt

Sebastian Padó

14. Januar 2005

Aufgabe 1a

Zwei Klassen (gut, schlecht)

Aufgabe 1b

Drei Features (Ablaufdatum, Geschmacksrichtung, Grösse)

Aufgabe 1c

- Ablaufdatum: 20.1., 22.1., 24.1.
- Geschmack: Erdbeer, Himbeer, Kirsche
- Grösse: 500ml, 150 ml.

Groesse des Featureraumes: $3 * 3 * 2 = 18$

Aufgabe 1d

Wir haben 7 von 18 Instanzen gesehen, also etwa 38%. Ein Klassifikator, der alle Features verwendet, wird also vermutlich keine gute Abdeckung haben: wenn neue Daten klassifiziert werden sollen, die diese Features zufällig kombinieren, haben wir die Kombination mit über 60% Wahrscheinlichkeit nicht gesehen.

Aufgabe 1e

Die Zellen enthalten die Wahrscheinlichkeiten:

Klasse	P(Klasse 20.1.)	P(Klasse 22.1.)	P(Klasse 24.1.)
Gut	1/4	3/3 = 1	2/3
Schlecht	3/4	0/3 = 0	1/3

Für den 20.1. ist die wahrscheinlichste Klasse also *schlecht*, für die beiden anderen Daten *gut*. Wenn man jede Instanz mit der wahrscheinlichsten Klasse für ihr Datum klassifiziert, macht man 2 Fehler auf 10 Datenpunkten, also ist die Akkuratheit des Klassifikators 80% auf den Trainingsdaten.

Aufgabe 1f

Klasse	P(Klasse Himbeer)	P(Klasse Erdbeer)	P(Klasse Kirsch)
Gut	4/4 = 1	2/4 = 1/2	0/2 = 0
Schlecht	0/4 = 0	2/4 = 1/2	2/2 = 1

Für Himbeer ist die wahrscheinlichste Klasse also *gut*, für Kirsch *schlecht*, und für Erdbeer sind die beiden Klassen gleich wahrscheinlich. Wenn man jede Instanz mit der wahrscheinlichsten Klasse für ihren Geschmack klassifiziert, macht man auch 2 Fehler auf 10 Datenpunkten, also ist die Akkuratheit des Klassifikators ebenfalls 80% auf den Trainingsdaten. (Egal, welche Klasse man für Erdbeer wählt: die Klassifikation ist gleich gut).

Klasse	P(Klasse 150ml)	P(Klasse 500ml)
Gut	3/6 = 1/2	3/4
Schlecht	3/6 = 1/2	1/4

Für 150 ml sind beide Klassen gleich wahrscheinlich; für 500ml ist *gut* wahrscheinlicher. Wenn man jede Instanz mit der wahrscheinlichsten Klasse für ihre Grösse klassifiziert, macht man auch 4 Fehler auf 10 Datenpunkten, also ist die Akkuratheit des Klassifikators nur 60% auf den Trainingsdaten.

Aufgabe 1g

Offenbar sind der Geschmack und das Ablaufdatum gute, aber nicht perfekte Features, um vorherzusagen, ob ein Joghurt noch geniessbar ist. Die Grösse der Packung scheint nur eine geringe Vorhersagekraft zu besitzen.

Aufgabe 1h

Klasse	P(Kl 20,Erd)	P(Kl 20,Him)	P(Kl 20,Him)	P(Kl 22,Erd)	
Gut	0/2 = 0	1/1 = 1	0/1 = 0	n.d.	
Schlecht	2/2 = 1	0/1 = 0	1/1 = 1	n.d.	
	P(Kl 22,Him)	P(Kl 20,Kir)	P(Kl 24,Erd)	P(Kl 24,Him)	P(Kl 24,Kir)
Gut	3/3 = 1	n.d.	2/2 = 1	n.d.	0/1 = 0
Schlecht	0/3 = 0	n.d.	0/2 = 0	n.d.	1/1 = 1

Wenn man wieder jede Instanz mit der Klasse klassifiziert, die am wahrscheinlichsten für ihre Datum-Geschmack-Kombination ist, werden alle Instanzen richtig klassifiziert.

Aufgabe 1i

Das Modell ist nicht besonders robust: Viele Feature-Kombinationen wurden gar nicht oder nur ein Mal gesehen. Für die einmal gesehenen Feature-Kombinationen ist sich der Klassifikator sehr sicher (1 von 1 Instanzen = 100% Sicherheit), aber die Hinzufügung auch nur einer einzigen Instanz kann die Wahrscheinlichkeit von 100% auf 50% (Gleichstand!) drücken und damit die Klassifikation ändern. Dies beleuchtet ein grundlegendes Problem der Statistik: je mehr Features man verwendet, desto besser ist die theoretische Vorhersagekraft des Modells; andererseits sinken bei einer konstanten Trainingsmenge die Frequenzen der einzelnen Featurekombinationen, weil es mehr solche gibt, womit das Modell anfälliger wird.

Aufgabe 1j

Die einfachen Modelle, die nur je ein Feature verwenden, werden nicht besonders gut sein, aber dafür eine ziemlich gute Abdeckung haben: sie werden zumindest irgendwelche Voraussagen machen können. Das komplette Modell hat nur eine Abdeckung von 38% (siehe 1d), wird also für die meisten Datenpunkte überhaupt nichts aussagen können.

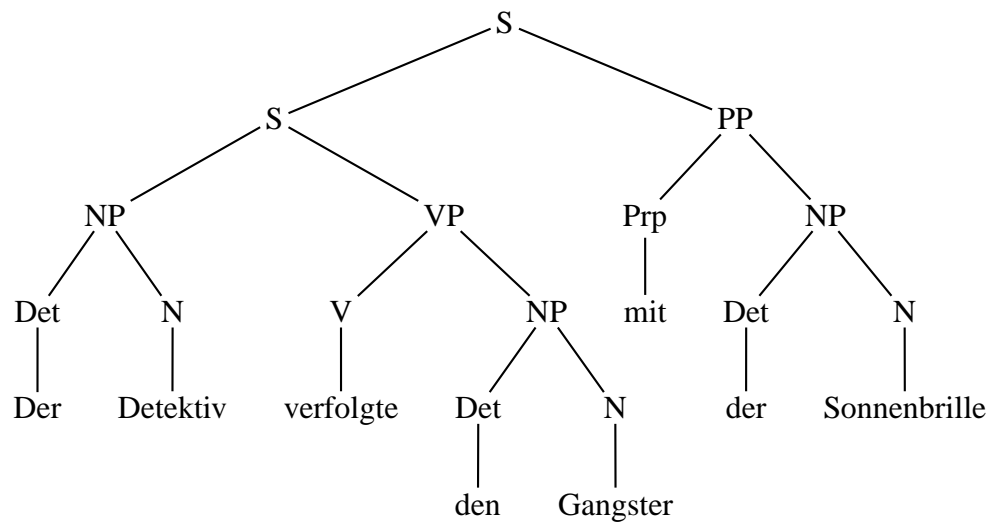
Da das Modell $P(Klasse|Datum, Geschmack)$ ziemlich gut ist, bietet es sich an, dieses Modell zu verwenden, wenn es die zu klassifizierende Feature-Kombination schon einmal gesehen hat, und ansonsten auf die Modelle $P(Klasse|Datum)$ und $P(Klasse|Geschmack)$ zurückzugreifen (back-off)!

Aufgabe 2

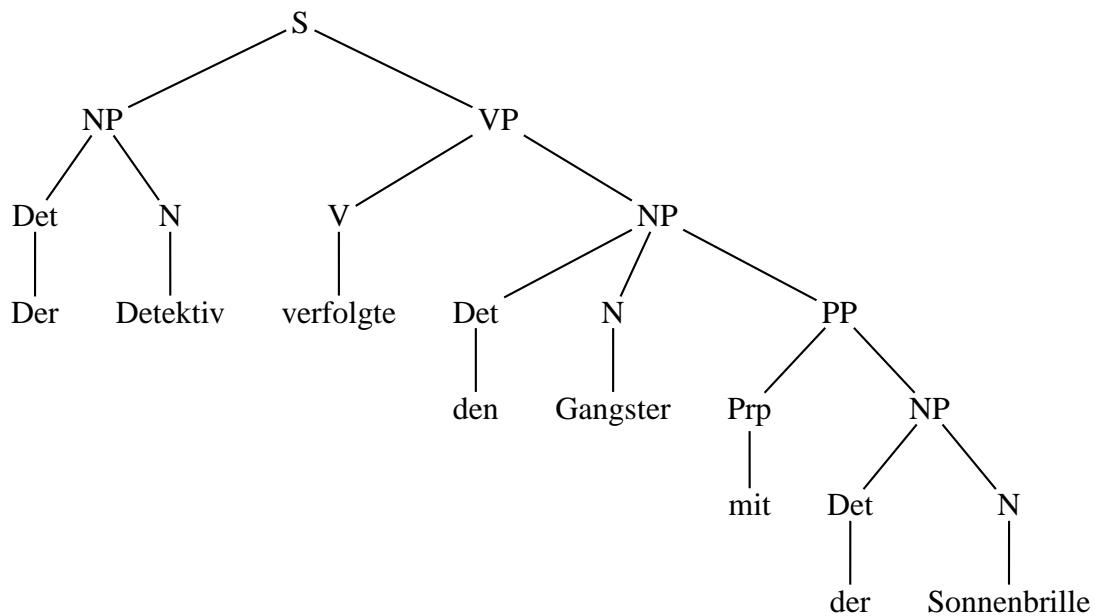
Der 71 Jahre alte Narasimha Rao hat eine Führungsposition inne (Rolle LEADER eines LEADERSHIP-Frames), und zwar regiert er über Indien (JURISDICTION). Diese Rolle (ROLE) hat er aber erst seit kurzem (NEW LEADER eines CHANGE_OF_LEADERSHIP-Frames). Ausserdem war sich "kaum jemand" (COGNIZER) sicher (Frame CERTAINTY, was folgendes angeht (CONTENT): nämlich, dass er (ITEM) einen STANDARD übertreffen würde (Frame SURPASSING). Der Standard war gesetzt durch die Einschätzung (CATEGORISATION) von ihm (ITEM) durch jemanden, nämlich diejenigen (ROLE, die die Congress-Partei (JURISDICTION anführten (Frame LEADERSHIP).

Aufgabe 3a

Lesart 1: Die Verfolgung findet mit der Sonnenbrille statt.



Lesart 2: Der Gangster besitzt die Sonnenbrille.



Aufgabe 3b

Drei wichtige Mechanismen des Deutschen, die beachtet werden müssen, um wohlgeformte Sätze zu erzeugen, sind Subkategorisierung, Kongruenz und Kasuszuweisung (Rektion). Alle können von der Grammatik verletzt werden:

- “Den Motorrad schläft”. Der Satz verletzt die *Kongruenz in der Nominalphrase*: Artikel und Kopfnomen müssen dieselben Kasus/Genus/Numerus haben. Weil die vorliegende Grammatik nur Nomina in der 3. Person Singular kennt, kann sie die *Subjekt-Verb-Kongruenz* nicht verletzen: Subjekt und Verb müssen in Kasus und Numerus übereinstimmen (nicht *du schläft*).
- “Der Gangster verfolgte”. Dieser Satz verletzt die Subkategorisierung des Verbs: verfolgte ist transitiv, muss also ein direktes Objekt haben.
- “mit die Sonnenbrille”: *mit* “regiert” den Dativ (d.h. die NP in einer *mit*-PP muss im Dativ stehen) – diese Fallzuweisung wird von der Grammatik nicht erzwungen. Entsprechend erzwingt sie nicht, dass das Subjekt eines Satzes im Nominativ stehen muss, sondern erlaubt Sätze wie “den Gangster schläft”.