

Übungsblatt 4

1. Statistische Modellierung

Herr Mueller leitet die Joghurt-Abteilung eines Supermarktes. Er versucht, ein statistisches Modell fuer die Wahrscheinlichkeit zu entwickeln, dass ein Joghurt schlecht ist. Er verwendet folgende Features: Ablaufdatum, Geschmacksrichtung, und Groesse. Er probiert 10 Joghurts als „Trainingsmenge“:

20.1.	Erdbeer	150ml	Schlecht
20.1.	Himbeer	150ml	Gut
20.1.	Erdbeer	150ml	Schlecht
20.1.	Kirsch	500ml	Schlecht
22.1.	Himbeer	150ml	Gut
22.1.	Himbeer	150ml	Gut
22.1.	Himbeer	500ml	Gut
24.1.	Kirsch	150ml	Schlecht
24.1.	Erdbeer	500ml	Gut
24.1.	Erdbeer	500ml	Gut

- Wie viele Klassen gibt es?
- Wie viele Features gibt es?
- Machen Sie fuer jedes Feature eine Liste aller „gesehenen“ Werte. Wie gross ist der gesamte Feature-Raum?
- Fuer welchen Anteil des gesamten Feature-Raumes haben wir Instanzen „gesehen“? Was bedeutet das fuer die Abdeckung eines Klassifikators, der auf diesem Feature-Raum arbeitet, auf neuen Daten?
- „Lernen“ Sie die Wahrscheinlichkeitsverteilung $P(\text{Klasse} \mid \text{Datum})$ aus den Daten. Wie gut kann dieses Modell die Klasse vorhersagen?
 - Hinweis: Benutzen Sie MLE (siehe Folien). Im aktuellen Fall reduziert sich die Formel zu: $P(\text{Klasse} \mid \text{Datum}) = \frac{f(\text{Klasse}, \text{Datum})}{f(\text{Datum})}$
 - Hinweis: Erstellen Sie eine Tabelle fuer alle Kombinationen aus Feature-Wert und Klasse und tragen Sie die Wahrscheinlichkeit ein.
 - Hinweis: Klassifizieren Sie jede Instanz nach dem Entscheidungskriterium „wahrscheinlichste Klasse“. Wenn es keine wahrscheinlichste Klasse gibt, waehlen Sie eine zufaellige.
- Wiederholen Sie (e) fuer $P(\text{Klasse} \mid \text{Geschmack})$ und $P(\text{Klasse} \mid \text{Groesse})$.
- Wie gut funktionieren die Modelle? Beachten Sie, dass Sie bei einer binaeren Klassifikation (gut/schlecht) durch Raten 50% Akkuratheit erreichen koennen.

- h. Konstruieren Sie jetzt das Modell $P(\text{Klasse} \mid \text{Datum}, \text{Geschmack})$. Wie gut ist dieses Modell?
- i. Wie „robust“ ist dieses Modell – wie stark koennte die Hinzufuegung eines neuen Datenpunktes das Modell veraendern?
 - 1. Hinweis: Denken Sie an Feature-Kombinationen, die nur ein Mal gesehen wurden
- j. Modelle werden i.A. an neuen Daten (Testdaten) ueberprueft. Was erwarten Sie fuer die Modelle, die ein Feature verwenden, das Modell aus (h), das zwei Features verwendet, und das komplette Modell $P(\text{Klasse} \mid \text{Geschmack}, \text{Groesse}, \text{Datum})$?
 - 1. Hinweis: Denken Sie an die Abdeckung des jeweiligen Featureraums

2. Semantische Annotation

Unter <http://www.ps.uni-sb.de/~pado/moin.cgi> finden Sie das Wiki des SALSA-Projekts und darin 20 hand-annotierte Sätze. Unter <http://www.icsi.berkeley.edu/~framenet/> finden Sie Beschreibungen der verwendeten Frames (Menüpunkt “FN Data”). Umschreiben Sie mithilfe der Frame-Definitionen, welche “Bedeutung” Satz 53 hat. (“UNKNOWN”-Frames können Sie ignorieren.)

Beispiel für Satz 52: “Rao nimmt eine Führungsrolle ein (LEADERSHIP, und zwar als Premierminister). Er unternimmt gleichzeitig eine gefährliche Handlung (RISK_ACTION), und zwar eine revolutionäre Wirtschaftreform.”)

3. Syntaktische Analyse

Gegeben sei eine kontextfreie Grammatik mit den Produktionsregeln

S	-> S PP	V	-> verfolgte, schlief
S	-> NP V NP	N	-> Detektiv, Gangster, Motorrad, Sonnenbrille
S	-> NP V	Prp	-> auf, mit
NP	-> NP PP	Det	-> der, dem, den
PP	-> Prp NP		
NP	-> Det N		

- a. Leiten Sie den Satz *Der Detektiv verfolgte den Gangster mit der Sonnenbrille.* auf zwei unterschiedliche Weisen ab, indem Sie die beiden Strukturbäume angeben, die aus der Ableitung resultieren.
- b. Leiten Sie mit dieser Grammatik zwei Wortfolgen ab, die keine korrekten Sätze des Deutschen sind. Kommentieren Sie, gegen welche grammatischen Regeln dabei verstoßen wird.