Mathematische Grundlagen der Computerlinguistik III: Statistische Methoden

Probeklausur

Garance Paris

Sommersemester 2011
16.07.2010

1. Es stehen 90 Minuten zur Bearbeitung zur Verfügung.
2. Erlaubte Hilfsmittel sind eine 4–5-seitige selbsterstellte (handgeschriebene) Zusammenfassung des Stoffes und einen Taschenrechner (kein Handy!).
3. Die einzelnen Lösungsschritte sollen soweit begründet werden, dass der Lösungsansatz nachvollziehbar ist.
4. Zeiteinschätzung: 10 Minuten jeweils für Seiten 2, 3 und 4 (ohne Bonusfrage), 15 Minuten für Seiten und 5, 6 und 7. Es bleiben also 15 Minuten übrig.

Viel Erfolg!!!

1. **Korpora und Statistische Modelle allgemein**

   a) Was ist eine Konkordanz und wofür werden sie in der Sprachwissenschaft eingesetzt? Gibt ein kleines Beispiel (4 Zeilen).

   b) Nenne ein großes, allgemeines Problem, das bei der Annotation von Korpora auftritt ("allgemein": auf alle linguistischen Ebenen)?
   Welche Lösungsansätze gibt es?

   c) Was ist ein *n-gramm-Modell*? Nenne zwei Anwendungen, bei denen ein n-gramm-Modell eingesetzt wird.

2. **Stochastisches Parsing**

   a) Vergleiche mit eigenen Worten die *"Inside"-Prozedur* und den *"Viterbi"-Algorithmus*: Was berechnen sie, was haben sie gemeinsam, was unterscheidet sie?

   b) Welchen Vorteil bieten diese Algorithmen gegenüber dem naïven Ansatz?

   c) Welche Rolle spielen die Variablen β, δ und ψ?

3. **Unix-Tools**

   a) Was machen die unteren zwei Befehlsketten, wenn man sie auf einem deutschen Korpus laufen lässt? Beschreibe das Ergebnis mit geeigneten linguistischen Begriffen, und erkläre, wie es technisch zu Stande kommt.

      i. `tr -sc 'a-zA-Z' '\n' < KORPUS | tr 'A-Z' 'a-z' | grep 'heit$' | sort | uniq -c`

      ii. `tr -sc 'a-zA-Z' '\n' < KORPUS | tr 'A-Z' 'a-z' | grep 'heit$' | sort | wc -l`

   b) Welche(r) Befehl(e) kannst Du verwenden, um aus einem Korpus die fünf häufigsten Wörter, die mit *"ge-"* anfangen, mit ihren Häufigkeiten auszugeben, unabhängig davon, ob sie mit einem großen oder einem kleinen Buchstaben anfangen?

   c) (Bonusfrage) Wie muss Du den oberen Befehl abändern, wenn die Häufigkeiten der einzelnen Wörter nicht angezeigt werden sollen?

4. **Clustering**

   Wir möchten die Wörter *gehen*, *Freund*, *Baum*, und *sprechen* in Kategorien einteilen, verfügen aber über keinen trainierten PoS-Tagger. Weil es aber nur wenige Artikel oder Präpositionen gibt, können wir leicht in einem Korpus zählen, wie oft jeder Artikel oder jede Präposition unmittelbar vor oder nach einem der Wörter vorkommt.

Gesamtzahl	gehen	sprechen	Freund	Baum
Artikel	0	2	1	4
Präpositionen	2	4	1	0

   a) Wende den Algorithmus *"k-means"* an, um die Wörter unbetreut ihrer Ähnlichkeit nach in Kategorien aufzuteilen. Als Zentren verwende bei der Initialisierung $c_1 = (3, 1)$ und $c_2 = (1, 2)$. Vergiss nicht zu zeigen, wann und warum der Algorithmus abbricht!

   b) Als Mensch weißt Du, zu welcher PoS-Kategorie diese Wörter gehören. Evaluiere als Experte die Cluster, die von dem Algorithmus erzeugt wurden, indem Du die Konfusionsmatrix angibst und Präzision und Erinnerung berechnest.

5. **Naiv-Bayes**

Wir betrachten Sätze mit einer NP-V-NP-PP-Struktur wie *"Peter baked the cake with friends"* oder *"Peter baked the cake with almonds"*.

In solchen Sätzen kann die Präpositionalphrase strukturell entweder an das vorausgehende Verb *(high attachment*, 1. Beispiel) oder an den zweiten Substantiv *(low attachment*, 2. Beispiel) angehängt werden.

In einem kleinen Korpus wurde jede Instanz einer solchen Ambiguität als *low-* oder *high-attachment* annotiert, sowie einige weitere Merkmale des Satzes festgestellt: das Verb, den 1. Substantiv des Satzes, die Präposition und den 2. Substantiv.

Instanz	Verb	N1	Präposition	N2	Attachment
1	baked	cake	with	almonds	low
2	saw	cake	with	almonds	low
3	saw	movie	with	friends	high
4	saw	movie	on	Tuesday	high

a) Gib in die folgende Tabelle die A-Posteriori-Wahrscheinlichkeiten an, die ein Naiv-Bayes-Klassifikator verwenden würde, um vorauszusagen, ob die PP an das Verb oder an die NP angehängt werden soll.

b) Was sagt der Klassifikator für den Satz *"Peter baked the cake on Tuesday"* voraus?

c) Auf welches Problem stoßt Du beim beantworten der vorigen Aufgabe?

d) Gib in die Tabelle die neuen A-Posteriori-Wahrscheinlichkeiten an, wenn mit *Add-One* geglättet wird.

e) Klassifiziere den Satz erneut. Welche Voraussage macht der Klassifikator jetzt?

6. **Informationstheorie**
   Wir betrachten eine Sprache, die aus nur fünf Wörtern (Types) besteht. In einem kleinen Korpus (450 Tokens, 150 Sätze) werden die folgenden Häufigkeiten beobachtet:

Bill	Sue	kisses	pizza	likes
120	130	100	50	50

   a) Bestimme die Entropie pro Wort für diese Sprache auf Basis der oberen Werten.

   b) Aufgrund einer Korpusuntersuchung werde aber jetzt festgestellt, dass in der Sprache nur vier Sätze möglich sind:

      "Bill kisses Sue" tritt 80 Mal auf.
      "Sue kisses Bill" tritt 20 Mal auf.
      "Bill likes pizza" tritt 20 Mal auf.
      "Sue likes pizza" tritt 30 Mal auf.

      Berechne erneut mit Hilfe dieser Angaben die durchschnittliche Entropie pro Wort (Hinweis: Betrachte jeden Satz als eine unteilbare Einheit). Erkläre den Unterschied zu dem Ergebnis aus 6a.