Klausur 2011: Quantitative Methoden der CL

1 Schätzen von Wahrscheinlichkeiten

Google übersetzt:

```
Der Mann hat das Ziel getroffen.
Der Mann hat seine Verbündeten getroffen.
Der Mann trifft seinen Freund.
Der Mann hat seine Vorbereitungen getroffen.
Der Mann hat seine Schwester getroffen.
Der Ball hat das Tor getroffen.
Der Abgeordnete trifft seinen Freund mit seiner Behauptung.
Der Abgeordnete trifft seinen Freund.
Der Mann trifft seinen Freund mit dem Ball.
```

als:

```
The man has hit the target.
The man has made his allies.
The man meets his friend.
The man has made his preparations.
The man has taken his sister.
The ball hit the goal.
The deputy hit his friend with his claim.
The deputy meets with his friend.
The man meets his friend with the ball.
```

Schätzen Sie folgenden Wahrscheinlichkeiten, notieren Sie die Fragestellung auch als Wahrscheinlichkeit, z.B. P(dirObj=bel|Subj=bel)

1. P, dass das Subjekt belebt (Subj=bel) ist
2. P, dass das direkte Objekt (dirObj=bel) belebt ist
3. P, dass das Subjekt und das direkte Objekt belebt sind
4. P, dass "treffen" mit "hit" übersetzt wird

5. P, dass "treffen" mit "meet" übersetzt wird
6. P, dass "treffen" mit "hit" oder "meet" übersetzt wird
7. P, dass "treffen" mit "meet" übersetzt wird, gegeben dass das Subjekt belebt ist
8. P, dass "treffen" mit "meet" übersetzt wird, gegeben dass das Subjekt und das direkte Objekt belebt sind
9. P, dass das direkte Objekt belebt ist, gegeben dass das Subjekt belebt ist
10. Sind P(Subj=bel) und P(dirObj=bel) stochastisch voneinander unabhängig? Belegen Sie ihre Aussage.

2 Präzision, Ausbeute

Ihr Tagger hat 260 Wortformen als Adjektive getaggt. Laut Gold Standard gibt es aber 400 Adjektive. 200 Tagger-Entscheidungen sind richtig. Berechnen Sie Präzision und Ausbeute.
Präzision: 200/260 Ausbeute: 200/400

3 Witten-Bell Smoothing

Die Wahrscheinlichkeitsmasse für nicht-gesehene Ereignisse ist:

$$\frac{T}{N+T}$$

mit T=Anzahl der beobachteten Types und N=Anzahl der Token Verteilt auf alle nicht-gesehenen Ereignisse ergibt das:

$$\frac{T}{Z(N+T)}$$

wobei Z die Menge der Nullereignisse ist.

Alle *beobachteten* Ereignisse haben die Schätzung (c_i als Zählfunktion):

$$\frac{c_i}{N+T}$$

Aufgabe: Wahrscheinlichkeit von Lesarten schätzen. Das Wort 'love' hat in WordNet 6 Nomenlesarten. Sie möchten die Wahrscheinlichkeit für diese Lesarten schätzen. Sie machen eine Stichprobe - Sie disambiguieren von Hand 100 Vorkommen des Wortes 'love' in einem Corpus. Nehmen wir der Einfachheit halber an, dass 4 der 6 Lesarten jeweils 25 mal auftreten, 2 Lesarten also nicht angetroffen werden. Berechnen Sie a) nach Witten-Bell die Wahrscheinlichkeit eines einzelnen nicht-gesehenen Ereignisses. Was ist b) die Wahrscheinlichkeit (jeweils) der vier gesehenen Ereignisse?
Masse für nicht-gesehene: 4/(100+4)
Masse pro nicht-gesehenes: 4/(2*(100+4))= 1/52
gesehene: 25/104

4 Parsing

Gegeben folgende Grammatik. Was ist die Wahrscheinlichkeit, das folgender Satz von der Grammatik generiert wird: 'Theo liebt Anna'? Sie brauchen den konkreten Zahlenwert nicht zu ermitteln, nur die Berechnungsvorschrift.

```
0.8  S -> NP VP
0.2  S -> V NP

0.3  NP -> PN
0.8  NP -> Det Noun

0.4  VP -> V
0.6  VP -> V NP

0.1 PN -> Anna
0.1 PN -> Theo
..

0.25 Verb -> liebt
...
```

Lösung: 2* (0.8 * 0.3 * 0.1 * 0.6 * 0.25 * 0.3 * 0.1)
mal 2, da zwei Lesarten

5 Hypothesentesten + Entscheidungsbäume

Das Splitten eines Astes beim Entscheidungsbaumlernen ist nur dann sinnvoll, wenn der Informationsgewinn durch das Splitting steigt. Ansonsten sollte man besser den Ast als Terminal in den Baum einbauen. Steigender Informationsgewinn bedeutet statistisch gesehen, dass die Verteilung positiver und negativer Beispiele nach dem Splitten ungleich der Verteilung vor dem Splitten ist.

Operationalisieren Sie diese Entscheidung als statistischen Test mittels der Chiquadrat-Verteilung.

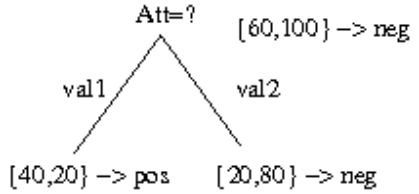
- Formulieren Sie eine entsprechende Nullhypotese (in Worten, keine Formel nötig).
- Berechnen Sie den Prüfwert (der zu ermittelnde Wert, der dann als Wert aus der Verteilung eine Wahrscheinlichkeit bekommt) für das unten gegebene Beispiel:

$$\chi^2_{(k-1)} = \sum_{j=1}^{k} \frac{(f_j - e_j)^2}{e_j}$$

wobei:
e_j = Erwartungswert
f_j = Häufigkeit
k = Anzahl Ereignisse

(Formelinstantiierung reicht!)

Hier die Daten: Gegeben folgender Effekt beim Splitting - soll es durchgeführt werden oder ist es unnötig (nicht wirklich hilfreich)?



[6,10] heisst: 6 positive Beispiele, 10 negative; daher gemäss Mehrheitsklasse wird hier eine negative Klassifikation erfolgen

5.1 Lösung

- Wahrscheinlichkeit vor dem Splitten (insg. 160 Instanzen):

  pos = 60/160, neg = 100/160

- Nullhypothese: p(pos)= 60/160 und p(neg)=100/160 (auch) nach Splitten
- Ergebnis des Splittens: {40,20} und {20,80}
- 4 Erwartungswerte: für jeden Ast, positiv und negativ
- EW pos e_j = (60/160)*60 (bei 60 Instanzen)= 3600/160= 22.5
- EW neg e_j = (100/160)*60 = 6000/160 usw.
- Prüfwert $\chi^2_{k-1} = \Sigma_{j=1}^{k} \frac{f_j - e_j}{e_j} = \frac{(40-22.5)^2}{22.5} + \frac{(20-37.5)^2}{37.5} + ...$