Klausur 2014: Quantitative Methoden der CL (60 Min, 60 Pkt)

1 Maximum Likelihood Schätzung, Smoothing (10 Min, 12 Pkt)

Ein positives Wort kombiniert mit einem negativen ergibt eine positive (z.B. reuiger Sünder) oder negative Phrase (z.B. 'perfektes Desaster'). Gegeben folgende Zählungen (Adj+Nomen=NP)

```
adj, nomen -> NP:  Frequenz

pos, neg   -> neg: 200
pos, neg   -> pos: 100
neg, pos   -> neg: 150
neg, pos   -> pos: 0
neg, neg   -> pos: 0
neg, neg   -> neg: 20
pos, pos   -> neg: 0
pos, pos   -> pos: 30
```

Schätzen Sie ein Modell (9 Pkt) für

- P(NP=neg|adj=pos,nomen=neg)=$\frac{200+1}{300+2}$
- Erläuterung: c ist pos oder neg, also v=2
- P(NP=pos|adj=pos,nomen=neg)=$\frac{100+1}{300+2}$
- P(NP=neg|adj=pos,nomen=pos)=1/32
- Erläuterung: es gibt 30 Fälle für $adj = pos, nomen = pos$ bei v=2
- Verwenden Sie Add-One-Smoothing (3 Pkt)

$$P(NP = c|e_i) = \frac{s_i + 1}{s + v}$$

  - s_i ist die Anzahl der Beispiele für ein c gegeben e_i
  - s ist die Gesamtzahl der Beispiele von e_i

- c ist pos oder neg
- v ist die Anzahl der Klassen pro e_i

- Schätzen Sie nun P(adj=pos,nomen=neg) mit Smoothing. Adaptieren Sie die Smoothing-Formel entsprechend (was ist nun v)? (3 Pkt)
- $P(adj = pos, nomen = neg) = \frac{300+1}{500+4}$, da 4 Kombination von pos,neg

Alternativ (nur 6 Pkt): Berechnen Sie das Modell ohne Smoothing

2 Sprachmodell (5 Min, 6 Pkt)

Braucht man für die Übersetzung von Deutsch nach Englisch mit Bayes ein deutsches, ein englisches Sprachmodell oder braucht man beides?
Notieren Sie die Bayes'sche Formel für das Problem der maschinellen Übersetzung und identifizieren Sie darin das Sprachmodell.

Lösung:

$$P(e|d) = \frac{P(e) * P(d|e)}{P(d)}$$

P(d) entfällt. P(e) ist das englische Sprachmodell, dieses braucht man.

3 Inter Annotator Aggrement (IAA) und Entropie (20 Min, 20 Pkt)

Zwei Classifier liefern für ein Datenset folgende gemeinsame Analyse (Die Spalte ist das Ergebnis von A, Zeile ist B).

```
          X       Y <-- A
     --------------
X |     30      10
Y |     40      20
```

- Berechnen Sie das IAA der beiden Classifier (10 Pkt).

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

  - P(a): empirische Wahrscheinlichkeit für Übereinstimmung
  - P(a) = 50/100
  - P(e): Erwartungswert der Übereinstimmung
  - P(e) = 70/100*40/100 + 60/100*30/100= 46/100

$$\kappa = \frac{1/2 - 46/100}{1 - 46/100} = 2/27$$

- Nehmen wir an, die Präzision für die Klasse X von A sei 80%. Wieviele der X von A sind richtig? (2 Pkt)
- d.h. 80% von 30+40 = 56
- Kann man entscheiden, wie der Recall von A ist? Falls ja, geben Sie den Wert, falls nein, begründen Sie, warum nicht. (2 Pkt)
- Nein, keine Information über den Goldstandard vorhanden
- Wie gross ist die Entropie der Übereinstimmung: Instantiieren Sie die Formel (4 Pkt)
- -(1/2*log(1/2))-(1/2*log(1/2))= 1
- Präzisieren Sie: je grösser die Entropie, desto (kleiner) die Übereinstimmung (2 Pkt)

4 Hypothesentesten (10 Min, 6 Pkt)

Eine Münze wird 100 mal geworfen und gibt 40 mal Zahl und 60 Kopf.

Etwas später findet ein weiteres Experiment statt. Eine (dieselbe?) Münze wird 100 mal geworfen. Ergebnis: 30 mal Zahl, 70 mal Kopf.

Wie kann man die Binomialverteilung verwenden, um die Hypothese zu überprüfen, dass die beiden Versuche mit ein- und derselben Münze durchgeführt wurden, dass also z.B. der zweite Versuch mit derselben Münze wie der erste durchgeführt wurde? Instantiieren Sie auch die Formel: b(s;n,p) mit s = Anzahl der Erfolge, n = Anzahl der Versuche, p = Erfolgswahrscheinlichkeit.

Lösung:

Münze 1: P(Zahl)=4/10
Münze 2: P(Zahl)=3/10

b(30,100,4/10): die Wahrscheinlichkeit 30 mal Zahl zu bekommen bei 100 Würfen und einer Erfolgswahrscheinlichkeit von 4/10 (gemäss MLE Münze 1)

Nun verwendet man die kummulative Formel: P für 30 und weniger Treffer.

Wenn $\sum_{i=0}^{30} b(i, 100, 4/10)$ sehr klein ist, z.B. kleiner als 0.05, dann verwerfen

5 tfidf (5 Min, 6 Pkt)

Was ist (bedeutet) tf, was idf und was leistet tfidf (wozu ist es nützlich)?

tf: Zählung, wie oft Term in einem Dokument vorkommt, df: wie oft in allen Dokumenten, idf = log(N/df), wobei N die Anzahl der Dokumente ist. tfidf wird zur Gewichtung von Termen verwendet: hohes Gewicht für Terme in einem Dokument, die dort sehr häufig sind und selten in anderen auftreten. Das wird beim IR für Aehnlichkeit bzw. Suchrelevanz von Dokumenten verwendet.

6 Was gehört alles zum empirischen Problemlösen mit Machine Learning in der CL (10 Min,10 Pkt)

Beschreiben und erläutern Sie kurz die prototypischen Ingredenzien und zentralen Begriffe.

- annotierte (repräsentative) Daten, IAA dafür
- ML-Modellierung: Feature-Engineering, evtl. Smoothing, Classifier auswählen
- Evaluation: Test- und Trainingset oder Kreuzvalidierung, F-Mass etc.
- Baseline und Signifikanztest