Klausur 2011: Quantitative Methoden der CL (80 Min., 60 Pkt)

1 MLE von Wahrscheinlichkeiten (15 Min.)(10 Pkt)

```
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.

wir von Google übersetzt:

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 P, 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? Begründen Sie ihre Aussage.

1.1 Lösung

1. P(Subj=bel) = 8/9
2. P(Obj=bel) = 6/9
3. P(Subj=bel,Obj=bel) = 6/9
4. P(hit|treffen) = 3/9
5. P(meet|treffen) = 3/9
6. P(meet OR hit |treffen) = 3/9 + 3/9
7. P(meet|treffen, Subj=bel) = 1/8
8. P(meet|treffen, Subj=bel, Obj=bel) = 3/6
9. P(Obj=bel|Subj=bel) = 6/8
10. Nein, da P(Subj=bel)* P(Obj=bel) $\neq$ P(Subj=bel,Obj=bel)

2 Konfusionsmatrix, Präzision, Ausbeute (10 Min.)(10 Pkt)

Gegeben folgende Konfusionsmatrix (Zeile ist die wahre Klasse gemäss Gold Standard, Spalte Ergebnis eines ML-Verfahren, z.B. Timbl)

```
        X      Y
    -------------
X |    670     15
Y |      3    102
```

1. Erläutern Sie die Tabelle (4 Pkt)
2. Berechnen Sie die Akkuratheit insgesamt (2 Pkt)
3. Berechnen Sie pro Klasse Präzision und Ausbeute (4 Pkt)

2.1 Lösung

1. Timbl hat 670 richtig als X erkannt und 102 richtig als Y, 15 X hat Timbl fälschlicherweise als X klassifiziert etc.
2. Akkuratheit= 772/790
3. P(X) = 670/673, P(Y) = 102/117, R(X) = 670/685, R(Y)= 102/ 105

3 Hypothesentesten + Entscheidungsbäume (15 Min.)(10 Pkt)

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.

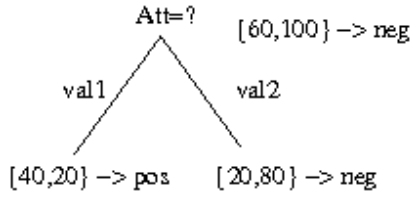
Operationalisieren Sie diese Entscheidung als statistischen Test mittels der Chiquadrat-Verteilung.

- 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: (Formelinstantiierung reicht!)

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

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

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

3.1 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}
- ein Ast reicht, {40,20}, 2 Erwartungswerte: 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= 37.5
- Prüfwert $\chi^2_{2-1} = \Sigma_{j=1}^{k=2} \frac{f_j - e_j}{e_j} = \frac{(40-22.5)^2}{22.5} + \frac{(20-37.5)^2}{37.5}$

4 Machine Learning (15 Min.) (10 Pkt)

Welche Lernverfahren haben wir in der Vorlesung behandelt? Was sind ihre Eigenarten und Unterschiede (stichwortartig)?

4.1 Lösung

- (1 Pkt) Bayes, Entscheidungsbaum, ähnlichkeitsbasiert (z.B. Timbl)
- (3 Pkt) Bayes: Statistisches Modell, apriori P etc. aus Trainingsdaten mit MLE und Smoothing ableitbar, Naive Bayes: Approximation durch Unabhängigkeitsannahme

- (3 Pkt) Entscheidungsbaum: Datenstruktur Baum mit Attributen als Knoten und Werten als Kanten, InfoGain (Entropiereduktion) als Splittingkriterium
- (3 Pkt)ähnlichkeitsbasiert: Speichern aller Beispiele, n-ähnlichsten suchen, Mehrheitsentscheidung, Aehnlichkeitsmasse erforderlich (z.B. Kosinus)

5 Bayes'sche Formel (10 Min.)(6 Pkt)

1. Erläutern sie den Unterschied zwischen P(A,B) und P(B|A), z.B. bei der Bigramm-Modelierung. Warum ist P(B|A) nützlicher als P(A,B)?
2. Wieso ist die Art der Anwendung der Bayesformel vom Typus P(B|A) auf CL-Probleme (z.B. Textklassifikation) oft 'naiv'?

5.1 Lösung

1. P(A,B): apriori P für ein Bigramm, Eigenschaft: unabhängig von Kontext
2. P(B|A): kontextuelle P - sehr viel genauer, da im Kontext verankert
3. wegen der Unabhängigkeitsannahme

6 Modellierung (15 Min.)(14 Pkt)

Sie wollen ein System zur Textklassifikation implementieren. Die drei Klassen sind Hardware, Software, Sonstiges. Sie haben vorklassifizierte Texte (100 pro Klasse) und können Morphologie, Tagging, Parsing und das deutsche Wortnetz GermaNet verwenden (bei Bedarf).

Beschreiben Sie, wie Sie zu Ihrem System kommen, welchen Ansatz sie wählen, welche Schritte nötig sind, welche Optionen Sie haben und wie Sie das Ganze dann evaluieren.

6.1 Lösung

- alles Taggen und Lemmatisieren, Stopwörter entfernen
- Splitten in Test und Training
- Verwendung von Bayes
- Vokabular bestimmen und Vektoren erzeugen
- Modell schätzen
- simple Baseline: Mehrheitsklasse
- Anwendung auf Trainingsmenge
- Eval-Masse
- evtl. Kreuzvalidierung