NAME:

Klausur 2013: Quantitative Methoden der CL (75 Min., 60 Pkt.)

1 Statistik (10 Min., 5 Pkt.)

Kreuzen Sie die wahren Aussagen an. Bei Bedarf können Sie Ihre Entscheidung auch näher begründen.

[ ] das Grenzwerttheorem besagt, dass Mittelwerte Ziehungen aus der Standardnormalverteilung sind

[ ] die Mittelwerte aller n-elementigen Teilmengen über der Grundmenge (Population) gruppieren sich eng um ihren Erwartungswert

[X] Konfidenzintervalle geben an, wie gross die maximale Abweichung des Mittelwerts vom wahren Mittelwert (der Population) ist

[ ] je grösser das Konfidenzintervall ist, desto vertrauenswürdiger ist die zugrunde liegende empirische Studie

[X] Stichproblengrösse: eine Verkleinerung des Unschärfewerts e bei sonst gleichbleibenden Annahmen (z.B. z-Wert bleibt) führt zu einer steigenden Stichprobengrösse

[ ] die z-Transformation bildet einen Wert auf eine beliebige Normalverteilung ab (z.B. in Abhängigkeit der Varianz)

[ ] falls der Prüfwert eine Wahrscheinlichkeit unterhalb des jeweiligen Signifikanzniveaus hat, ist die Nullhypothese eindeutig widerlegt

[ ] repräsentative Stichproben sind solche, bei denen alle Ereignisse gleichverteilt sind

[ ] wenn eine Stichprobe repräsentativ ist, dann ist ihr Mittelwert identisch mit dem Mittelwert der Grundgesamtheit

[X] bei der Berechnung der optimalen Stichprobengrösse wird in der Regel die grösste denkbare Varianz als Schätzwert verwendet

2 Hypothesentesten (15 Min., 15 Pkt.)

Gegeben folgende idealisierte Paare von accuracy-Werten einer fünffachen Kreuzvalidierung.

Baseline	System
50	55
62	65
53	55
60	62
61	64

Stellen Sie die Nullhypothese (kein Unterschied) und Alternativhypothese auf. Bestimmen Sie den Prüfwert (t-Wert). Instantiieren Sie die R-Funktion pb (pb(Prüfwert,df=?)) mit Ihren Werten (Mittelwert und Varianz ausrechnen, den Rest nicht).

$$t_{(n-1)} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2_{\bar{x}}/n}}$$

mit

$$\sigma^2_{\bar{x}} = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

Nehmen wir als Signifikanzniveau 0.05 und nehmen wir an, der ermittelte Prüfwert ist: 0.025. Was gilt?
Lösung:
(1 Pkt.): Nullhypothese: $\mu_1 - \mu_2 = 0$
Alternativhypothese: $\mu_1 - \mu_2 \neq 0$

(5 Pkt.): Summe Diff = -15, Mittelwert = -3
(5 Pkt.): Varianz = $(-5+3)^2 + 0 + -1^2 + -1^2 + 0 = 6/(4)$ (korr.Varianz)= 1.5
(2 Pkt.): 1-pt(-3/sqrt(1.5/5),df=4)
(2 Pkt.): Wir können die Nullhypothese verwerfen.

3 Interannotator Agreement, Präzision, Ausbeute (15 Min., 10 Pkt.)

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

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

1. Berechnen Sie die Akkuratheit insgesamt (2 Pkt)

   772/790

2. Berechnen Sie pro Klasse Präzision und Ausbeute (4 Pkt)

   Präzision X: 670/673, Y: 102/117

   Recall X: 670/685, Y: 102/105

3. Berechnen Sie Interannotator Agreement (4 Pkt)

   p(a) = 772/790, p(e) = p(X) + p(Y) = 673/790 * 685* 790 + 105/790 * 117/790

- Cohens Kappa

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

  - P(a): empirische Wahrscheinlichkeit für Übereinstimmung
  - P(e): Erwartungswert der Übereinstimmung

4 Entscheidungsbäume, Entropie, InfoGain (15 Min., 10 Pkt.)

Gegeben folgende Vektoren und ihre Klassenzugehörigkeit:
⟨2,j,i,j⟩ Klasse = 0
⟨2,g,2,g⟩ Klasse = 1
⟨b,j,2,g⟩ Klasse = 0

⟨i,g,i,j⟩ Klasse = 1
⟨i,j,2,j⟩ Klasse = 0

a) Wie gross ist die Entropie der Beispielmenge? (4 Pkt.)

$H(\{0,1,0,1,0\}) = p(1)*log_2p(1)+p(0)*log_2p(0) = 2/5*log_2(2/5)+3/5*log_2(3/5)$

b) Wie gross ist der Entropiegewinn, wenn wir die 4. Dimension als Wurzelknoten vorsehen (4 Pkt) ?

$gain(\{0,1,0,1,0\}, 4.Attribut) = H(\{0,1,0,1,0\})-(\frac{2}{5}H(\{1,0\})+\frac{3}{5}H(\{1,0,0\}))$

c) Bei welcher Dimension (Attribut) ist der Entropiegewinn maximal? (2 Pkt)
2

Formeln:

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

$$Gain(S, A) \equiv Entropy(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

wobei S_v die Menge der Beispiele ist, die für Attribut A den Wert v aufweisen.

Es reicht, wenn Sie die Formel instantiieren, Sie brauchen den Wert nicht ausrechnen.

5 MLE, Bayes (20 Min.,15 Pkt.)

Gegeben folgendes Polaritätslexikon:

- positiv: Erfolg, Wohlstand, Beförderung, Sieg
- negativ: Lüge, Niederlage, Arroganz
- neutral: Geld, Banken, Parteien

Es gibt drei Verben, die Nomen treten mit diesen Verben wie angegeben auf und zwar je ein Mal (dadurch ist auch die Verbhäufigkeit gegeben, zudem gibt es nur die 3 Verben).

- 'hoffen': Erfolg, Sieg, Niederlage (des Gegners), Beförderung, Parteien
- 'bedauern': Niederlage, Sieg (des Gegners), Lüge, Banken
- 'verachten': Geld, Erfolg, Wohlstand, Arroganz, Banken, Parteien

Aufgabe: Das Wort 'Dividende' ist noch nicht im Lexikon, tritt aber im Corpus mit allen drei Verben auf.

Berechnen Sie:

(a) P(Dividende=positiv|hoffen,bedauern,verachten) bzw.

(b) P(Dividende=negativ|hoffen,bedauern,verachten)

Es geht also um P(Wort=positiv|...), P(Wort=negativ|...)

Wir nehmen Unabhängigkeit an, i.e. die Wahrscheinlichkeit der Nomenpolarität ist abhängig vom Einzelverb. In einem ersten Schritt ist also die Formeln in a) und b) unter dieser Annahme zu approximieren.

(c) Wenden Sie die Bayes'sche Formel auf Ihre Approximation an.

(d) Schätzen Sie das statistische Modell: Verwenden Sie Add-One-Smoothing, um Nullwahrscheinlichkeiten zu vermeiden.

Tipp: Berechnen Sie die Aprioripolaritäten anhand des Lexikons.

(wer das mit dem Add-one-Smoothing nicht durchschaut, mache es ohne, das gibt dann halt Abzug)

Ist Dividende eher positiv oder negativ? (1 Pkt).

Lösung c): (2 Pkt):
Unabhängigkeitsannahme und Anwendung der Bayes'schen Formel:

$$P(N = Pol|v_1, .., v_n) \approx \prod_i P(N = Pol|v_i) = \prod_i \frac{P(N = Pol) * P(v_i|N = Pol)}{P(v_i)}$$

Folgender Schritt wurde nicht unbedingt erwartet - ich erkenne natürlich immer alternative Lösungswege an:

wir können vereinfachen (Nenner weg, P(N=POL) vor Produkt).

Der Nenner, die Wahrscheinlichkeit der Verben, ist für alle Klassen (positiv, negativ) gleich und daher nicht diskriminierend (daher entfernbar):

$$P(N = Pol|v_1, .., v_n) \approx P(N = Pol) \prod_i P(v_i|N = Pol)$$

Lösung d): statistisches Modell

Kommentar: Ich hatte in der Klausur mündlich die Restriktion gemacht, dass Sieg und Niederlage ganz normal zählen (also wie im Polartitätslexikon angegeben), also Sieg immer pos, Niederlage immer neg. Ausserdem hat niemand add-one smoothing gemacht, daher in der Lösung auch ohne.
Lösungshinweis: P(hoffen|N = pos) heisst: gegeben ein positives Nomen, wie ist dann die P von "hoffen".

- zähle alle Vorkommen der positiven Nomen bei den drei Verben insgesamt (pos. Nomenzählung: Erfolg kommt 2 x vor, 2 x Sieg, 1x Beförderung, 1x Wohlstand = 6 pos. Nomen insgesamt)
- pro Verb: bei hoffen sind es 3, bei bedauern ist es 1 (Sieg) ...
- P(hoffen|N=pos) = 3/6 etc.

Lösung d nun zusammen:

(3 Pkt): $P(N = pos) = 4/10, P(N = neg) = 3/10, P(N = neu) = 3/10$

(3 Pkt): $P(hoffen) = 5/15, P(bedauern) = 4/15, P(verachten) = 6/15$

(3 Pkt):

$P(hoffen|N = pos) = 3/6, P(hoffen|N = neg) = 1/4, P(hoffen|N = neu) = 1/5$

$P(bedauern|N = pos) = 1/6, P(bedauern|N = neg) = 1/2, P(bedauern|N = neg) = 1/5$

$P(verachten|N = pos) = 2/6, P(verachten|N = neg) = 1/4, P(verachten|N = neu) = 3/5$

Lösung a + b):(3 Pkt):

$$P(Dividende = pos|hoffen, bedauern, verachten) =$$

$P(N = pos)*p(hoffen|pos)*p(bedauern|pos)*p(verachten|pos) = 4/10*1/2*1/6*1/3 = 4/360$

$P(Dividende = neg|hoffen, bedauern, verachten) = 3/10*1/4*1/2*1/4 = 3/320$

Lösung der zentralen Frage: (1 Pkt) $3/320 < 4/320$, ergo: eher positiv als negativ (nach neutral war nicht gefragt)