Klausur 2015: Quantitative Methoden der CL (60 Min, 60 Pkt)

1 Theorie (10 Min, 6 Pkt)

Erläutern Sie den Begriff Machine Learning. Was ist es, wozu braucht man es in der CL. Was sind die Grundprinzipien, welche Verfahren gibt es etc.

2 Anwendung (45 Min, 54 Pkt)

Wir haben 1000 Punkte manuell annotieren lassen (von 2 Nicht-Linguisten). Ein Punkt ist entweder eine Abkürzung (AB) wie in 'Dr.' oder ein Satzendpunkt (SE).

2.1 IAA (10 Min, 10 Pkt)

Die zwei Annotatoren haben folgende Uebereinstimmung erzielt. Zeile ist Annotator A, Spalte B.

	AB	SE
AB	200	100
SE	100	600

Figure 1: Konfusionsmatrix 1

a) Wie ist das IAA? b) Wie beurteilen Sie dieses?

Lösung a):
p(a) = 800/1000 = 8/10, p(e) = P(AB)+P(SE)= 3/10*3/10 + 7/10 * 7/10 = 9/100 + 49/100 = 58/100
Kappa= (0.8- 0.58)/0.42 = 0.22/0.42 = 0.5

Lösung b)
sehr gering

2.2 Hypothesentesten

Es wurden nochmals zwei Annotatoren (diesmal zwei Linguisten) eingesetzt. Ergebnis (wobei die Zeile Annotator C ist und die Spalte D):

	AB	SE
AB	260	40
SE	80	620

	AB	SE
AB	200	100
SE	100	600

Figure 2: Konfusionsmatrix 2 und 1 (wiederholt)

Ist die Differenz der beiden Annotatorenteams signifikant? Besteht bei der Annotation der Linguisten (signifikant) mehr Uebereinstimmung als bei den Nicht-Linguisten? Mittels Kappa ist die Anwort eindeutig ja, da ein höherer Kappa-Wert erzeugt wird. Aber statistisch gesehen: ist das vielleicht eine Variation im Rahmen des Zufalls, in dem Sinne, dass beide Konfusionsmatrizen eigentlich gleich sind?

Verwenden Sie, ausgehend von den Konfusionsmatrizen den t-Test.
a) Beschreiben Sie Ihre Lösungsidee, b) instantiieren Sie die Formel und c) diskutieren Sie, was ein t-Wert von -0.009325 bedeuten würde (vgl. die Tabelle in der Formelsektion): welche Zeile ist relevant für das von Ihnen gewählte Signifikanzniveau (welches wählen Sie?)?

Lösung:

- a) Uebereinstimmung ist 1, Nichtuebereinstimmung ist 0, 0-1 ist dann mehr Uebereinstimmung, einseitiger Test mit fallenden Werten, H_0 : $\mu_1 >= \mu_2$,H_1 : $\mu_1 < \mu_2$, d.h. der linke Rand ist relevant, da wir μ_2 abziehen.
- Mittelwertsdifferenz $\bar{x}$ = Tabelle 1 hat 800 mal die 1, 200 mal die 0, Tabelle 2 hat 880 mal die 1, sonst die Null, ergo gilt als Differenz: -80/1000 (gemäss 80 mal 0-1 geteilt durch Gesamtmenge)
- $\bar{x} = -8/100$
- b) $t_{999} = \sigma_{\bar{x}} = \sqrt{((-1-(-8/100))^2 * 80)+(0-(-8/100))^2 * 920))/1000 * 999)}$
- t-Wert: $\bar{x}/\sigma_{\bar{x}} = -0.08/8.579044 = -0.009325048$

- c) die Hypothese, dass die Uebereinstimmung der beiden Teams identisch ist, kann auf dem Signifikanzniveau von 1% verworfen werden.

2.3 Naive Bayes (15 Min, 16 Pkt)

Wir harmonisieren die Annotation von C und D und bekommen dadurch einen Gold Standard mit 700 SE und 300 AB.

Mittels eines Pythonprogramms errechnen wir folgende Statistiken (wir betrachten Bigramme als Indikatoren):

- auf Buchstabe n folgen direkt 600 SE, 100 AB
- auf Buchstabe r folgen direkt 100 SE, 200 AB
- kurrioserweise kommen keine anderen Endbuchstaben im Gold Standard vor
- vorletzter Buchstube grossgeschrieben: 100 mal SE, 200 AB
- vorletzter Buchstube kleingeschrieben: 500 mal SE, 200 AB

Beispiel: die Zeichenkette 'Dr.' hat die Form Gr - ein grossgeschriebener Buchstabe gefolgt von einem r.

1. Berechnen Sie $P(SE|Gn)$ und $P(AB|Gn)$ (Verwendung der Bayes'schen Formel)
2. Können Sie den Nenner bei der Anwendung der Bayes'schen Formel ignorieren? Diskutieren Sie ihre Entscheidung.
3. Was folgt wahrscheinlicher auf Gn: SE oder AB?
4. Welche anderen Features könnten hilfreich sein bei der Punkteklassifikation?

`Lösung`

```
P(SE|Gn)= P(SE)*P(Gn|SE)= P(SE)* P(G|SE) * P(n|SE)
P(AB|Gn)= P(AB)*P(Gn|AB)= P(AB)* P(G|AB) * P(n|AB)
```

```
P(SE)=7/10
P(n|SE) =  6/7 (600/(600+100)
P(G)=3/10
P(n|AB) = 1/3 (100/(100+200))
P(G|SE)=1/6
P(G|AB)=1/2

P(SE|Gn) = P(SE) * P(n|SE) * P(G|SE) = 7/10 * 6/7 * 1/6 = 42/420 = 1/10
P(AB|Gn) = P(AB) * P(n|AB) * P(G|AB) = 3/10 * 1/3 * 1/2 = 3/60 = 1/20
```

1. macht mathematisch keinen Unterschied
2. Gn ist eher Satzendpunkt
3. Andere Buchstaben, evtl. Wortlänge, evtl. initiale Gross- bzw. Kleinschreibung des nachfolgenden Wortes

2.4 Add-One-Smoothing (10 Min, 8 Pkt)

Gegeben seien die 26 Buchstaben (ohne Umlaute etc.). Die Aufgabe bezieht sich auf obiges Setting.

1. Wie gross ist die Wahrscheinlichkeit eines einzelnen, nicht-gesehenen Buchstabens an letzter Stelle, jeweils für die beiden Klassen AB und SE (also etwa: P(x|AB) etc., wobei x hier für einen beliebigen, nicht-gesehenen Kleinbuchstaben steht), wenn Sie Add-one Smoothing verwenden?
2. Wie gross ist die Wahrscheinlichkeit für $P(SE|Gn)$ nach dem Smoothing?

Lösung

1. $P(x|SE) = 1/726$, wobei x ein nicht-gesehener Buchstabe ist; analog für $P(x|AB) = 1/326$
2. 601/726 * 1/6 (kein Smoothing bei $P(G|SE)$)

2.5 Kombinatorik (5 Min, 6 Pkt)

Gegeben seien die 26 Buchstaben (ohne Umlaute etc.). Die Aufgabe bezieht sich auf obiges Setting, i.e. Gross- und Kleinschreibung des vorletzten Buchstabens.

1. Wieviele 2er Sequenzen sind grundsätzlich möglich?
2. Nehmen wir an, dass Sequenzen mit identischen Buchstaben nicht vorkommen (z.B. aa oder Aa). Wieviele sind es dann?
3. Nehmen wir an, dass Gross- und Kleinschreibung und die Reihenfolge keine Rolle spielt, so dass also Ab= ab = ba nur einmal zählen. Wieviele sind es dann?

1. 52*26=1352
2. (52*26) - 52 = 1300
3. (26*26 - 26)/2 =312

3 Formeln

Formel 1:

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

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

Formel 2:

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

wobei $\sigma_{\bar{x}}^2$ die korrigierte Varianz der Mittelwertsdifferenzverteilung ist

Formel 3:

$$P(Ereignis\ x) = \frac{|Ereignis\ x| + 1}{|alle\ Ereignisse| + |verschiedene\ Ereignisse|}$$

mit

- |Ereignis x|: Zählung Ereignis x
- |alle Ereignisse|: Zählung über alle Ereingisse hinweg (inkl. x), i.e. Tokenzählung
- |verschiedene Ereignisse|: Anzahl der Ereignistypen

Die Formel muss bei bedingten Wahrscheinlichkeit entsprechend restringiert werden.

Schranken: Wahrscheinlichkeit, dass ein Wert im Intervall kleiner dem t-Wert liegt

	t-Wert	Wahrscheinlichkeit
a)	-2.33	0.01
b)	2.33	0.99
c)	-1.65	0.0496
d)	1.6	0.95