Klausur zur Vorlesung
20.07.2005

Freie Universität Berlin, SS 2005
Stefan Bentink · Jochen Jäger · Utz J. Pape · Claudio Lottaz · Rainer Spang

Vorlesung Genomische Datenanalyse

Name, Vorname	Matrikelnummer

- Schreiben Sie Ihren Namen und Matrikelnummer oben auf das Deckblatt und Ihren Namen auf jedes Zusatzblatt.
- Notieren Sie Ihre Antworten direkt auf die entsprechenden Aufgabenblätter
- Es stehen 120 Minuten zur Verfugung.
- Als Gedächtnisstütze lassen wir ein von Ihnen handschriftlich (auch beidseitig) beschriebenes DIN A4 Blatt zu. Dieses muss mit der Klausur abgegeben werden. Weitere Unterlagen sind nicht erlaubt.
- Einfache Taschenrechner ohne Programmier- und Speicherfunktionen dürfen im Gegensatz zu Notebooks und Ähnlichem benutzt werden.

Aufgabe 1 (Fallstudie Genexpressionsdaten, 14 Punkte).
Ein besonderer Typ von Leukämie wird durch die Fusion der beiden Gene BCR und ABL characterisiert. Mediziner haben herausgefunden, daß in diesen Patienten vor allem die Tyrosinkinase ABL die Bösartigkeit der Krankheit ausmacht. Deshalb ist für diese Patienten ein Tyrosinkinase Hemmer als Medikation angezeigt und nicht die übliche Chemotherapie. Die genaue und sichere Diagnose der BCR-ABL-Fusion ist also ein sehr wichtiger Aspekt der Leukämiediagnose.

Nehmen Sie an, daß zwei molekulare Tests zur Verfügung stehen. Test BCR1 sei 100% zuverlässig aber für den klinischen Alltag viel zu teuer und viel zu langsam. Der Test BCR2 wäre in der Klinik einsetzbar, erreicht aber nur eine Zuverlässigkeit von 60%.

Sie sind nun für die Auswertung einer medizinischen Studie zuständig. Für 100 Patienten wurden mit dem Test BCR1 das Vorhandensein der BCR-ABL-Fusion bestimmt und in 30 Patienten gefunden. Für alle 100 Patienten wurden auch mit einem modernen Microarray Genexpressionsprofile für 22000 Gene erstellt.

Ihre Projektpartner haben vorab bereits eine statistische Analyse der Microarray-Daten gemacht und berichten begeistert, daß Ihr auf 100 Patienten trainierter Klassifikator 90% der Patienten korrekt einteilt und somit sehr viel besser funktioniert als der Test BCR2 mit seinen 60%.

a) [4] Erklären Sie in eigenen Worten die folgenden Begriffe:
   - Trainingsdaten
   - Testdaten

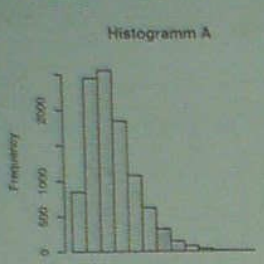
b) [2] Weshalb sollten Sie den Enthusiasmus Ihrer Projektpartner dämpfen?

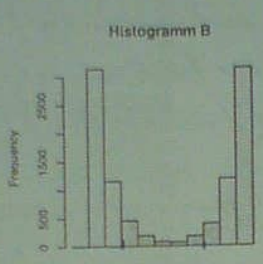
c) [4] Schlagen Sie eine Prozedur vor, wie man den Klassifikator objektiver evaluieren kann.

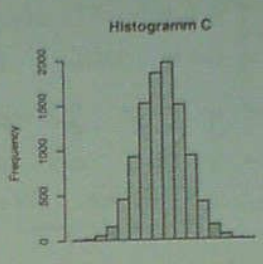
d) [4] Sie sollen einen Klassifikator, der sich auf Genexpressionswerte eines einzelnen Gens bezieht, bauen. Welches Gen würden Sie benutzen, wenn
   - Sie die Microarraydaten für die Entscheidung NICHT benutzen können.
   - Sie die Microarraydaten für die Entscheidung zu Hilfe nehmen können.

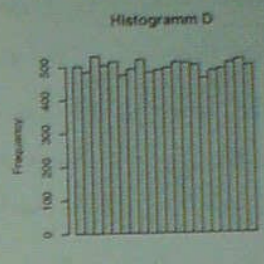
Aufgabe 2 (Kontinuierliche Verteilungen, 16 Punkte).

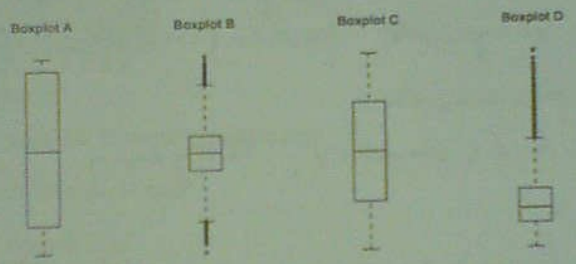
a) [8] Ordnen Sie folgende Histogramme, Boxplots und Verteilungsfunktionen zu:

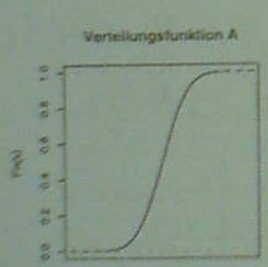


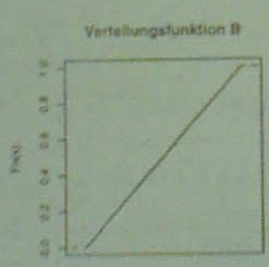


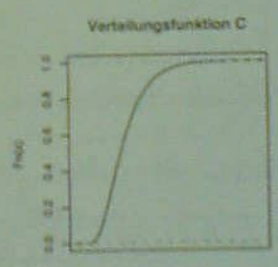


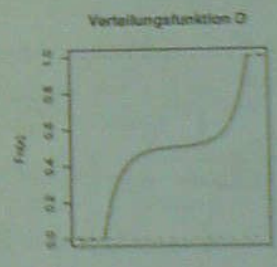












Histogramm	Boxplot	Verteilungsfunktion
A		
B		
C		
D		

b) Betrachten Sie eine kontinuierliche Verteilung deren Dichte die Form eines gleichschenkligen Dreiecks hat, wobei das Maximum dem Schnittpunkt der beiden Schenkel entspricht.

- [2] Sie benötigen mindestens zwei Parameter, um die Verteilung zu definieren. Beschreiben Sie ein (mögliches) Parameterpaar.
- [4] Bestimmen Sie die Dichtefunktion explizit in Abhängigkeit der beiden Parameter.
- [2] Geben Sie den Erwartungswert in Abhängigkeit der Parameter an. Es genügt eine Begründung ihres Ergebnisses - eine Rechnung ist nicht notwendig!

Aufgabe 3 (Blosum, 12 Punkte).
Hier eine Abbildung der Matrix Blosum62.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

a) [4] Beschreiben Sie wofür man die Blosum62 verwendet.

b) [4] Wie sind die Einträge zu interpretieren?

c) [4] Wie wurde die Matrix hergeleitet?

Aufgabe 4 (Transitionsmatrix, 12 Punkte).
In den im folgenden verwendeten Transitionsmatrizen entspricht jeweils Spalte 1 dem Nukleotid "A", Spalte 2 entspricht "C", Spalte 3 entspricht "G" und Spalte 4 entspricht "T".

a) [4] Erklären Sie in eigenen Worten was eine Transitionsmatrix ist.

b) [8] Gegeben sind die folgenden Transitionsmatrizen:

$$T_1 = \begin{pmatrix} 0.91 & 0.03, & 0.03 & 0.03 \\ 0.03 & 0.91, & 0.03 & 0.03 \\ 0.03 & 0.03, & 0.91 & 0.03 \\ 0.03 & 0.03, & 0.03 & 0.91 \end{pmatrix} \quad T_2 = \begin{pmatrix} 0.10 & 0.30, & 0.30 & 0.30 \\ 0.30 & 0.10, & 0.30 & 0.30 \\ 0.30 & 0.30, & 0.10 & 0.30 \\ 0.30 & 0.30, & 0.30 & 0.10 \end{pmatrix} \quad T_3 = \begin{pmatrix} 0.47 & 0.03, & 0.47 & 0.03 \\ 0.03 & 0.47, & 0.03 & 0.47 \\ 0.47 & 0.03, & 0.47 & 0.03 \\ 0.03 & 0.47, & 0.03 & 0.47 \end{pmatrix}$$

In den folgenden Sequenzpaaren ist jeweils die obere Sequenz unbhängig uniform verteilt. Die untere Sequenz ist aus der oberen erzeugt, indem jeweils das Zeichen an Position n unten aus dem Zeichen an Position n oben durch eine der drei Transitionsmatrizen generiert wird. Geben Sie für jedes Sequenzpaar an, Welche der drei Matrizen T_1, T_2 und T_3 benutzt wurde.

```
1 Original: TGGAAGGCGTGCCCACTTGCGATTCTAGAACCAGGTTTGATCTGTCCCGTCTTTCCGGAC
  Generated: CCCCACCGGCCAATACAAGTCCCTTCCGCGATACCAATCGACAGGAACAAGCAAGGCCC

2 Original: CTGACTGGAGAGCTGGCATCCAAGGGCGGTTTCTGTTATCGAAATCCATCCTAGCGTCGG
  Generated: CTGACTGGAGACGTGGCATCAAAGGGCGGTTTCTGTTATTGAGACCCACCCTAGCGTCGG

3 Original: TCGCACTTAATTACAATCGTCAGCATGCACTCCCATCCTAATGCACTTGGTAATCGTATTA
  Generated: TTGTGCCGACCGCGACCTCCGTTACGTGGTGTCAGTCCGGGTGTGTTTGGTAGCTATAACG

4 Original: CCCGCAACCTAGTTCTAAACCTTAGTAATCAGACTAGCTTGCCGACAGGATTGCATAATC
  Generated: GCGTTGGAGCCCAGAGCTCTCGCTACGTACTAGTGCCGCGATGTGAAGCGCCTCTCCGGG
```

Aufgabe 5 (Effizienz des Schätzers, 12 Punkte).

a) [8] Beschreiben Sie folgenden R-Code Zeile für Zeile.

```
> r <- rnorm(30000, mean = 5, sd = 5)
> X <- matrix(r, nrow = 1000, ncol = 30)
> s1 <- apply(X, 1, sd)
> Y <- matrix(rep(3, times = 5000), ncol = 5, nrow = 1000)
> Z <- cbind(X, Y)
> s2 <- apply(Z, 1, sd)
> summary(s1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.160   4.529   4.922   4.948   5.368   7.180

> summary(s2)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.999   4.254   4.615   4.637   5.026   6.703

> var(s1)

[1] 0.403074

> var(s2)

[1] 0.3388929
```

b) [4] Interpretieren Sie das Ergebnis mit Hinblick auf die Varianz/Bias Zerlegung eines Schätzers.

Aufgabe 6 (Transkriptionsbindestellen suchen, 18 Punkte).
Sie sollen eine Transkriptionsfaktorbindestelle mit einem Sequenzprofil charakterisieren und in einem bakteriellen Genom nach Kandidaten für solche Bindestellen suchen. Experimentell haben Biologen verifiziert, daß der Transkriptionsfaktor an folgende Nukleotidfolgen binden kann:

- TATA (70%)
- TATT (10%)
- TATC (2%)
- TCTA (10%)
- TGTA (3%)
- TTTA (5%)

a) [4] Erklären Sie in eigenen Worten die folgenden Begriffe:

- Likelihood
- Hintergrundmodell

b) [4] Berechnen Sie positions-spezifische, relative Häufigkeiten (entsprechend dem Donor-Profil bei Splice Sites) aufgrund dieser Daten (ohne Pseudocounts).

c) [4] Leiten Sie daraus einen Score zum Detektieren von Transkriptionsfaktorbindestellen her. Benutzen Sie dabei als Hintergrundmodell Sequenzen in denen Nukleotide an jeder Position unabhängig uniform verteilt sind. (Hilfestellung: log likelihood ratio).

d) [3] Berechnen Sie entsprechende Scores für die folgenden Nukleotidfolgen

- GATA:
- TCTA:
- TGTC:

e) [3] Geben Sie in der folgenden Sequenz die drei Positionen an, für die sich die höchsten Scores ergeben. Schreiben Sie auch die entsprechenden Scores dazu:

```
0.........1.........2.........3.........4.........
12345678901234567890123456789012345678901234567890
AAGCGGCGGTAATCTGGCCCTATCTTTAGGCGTATACCCTCGTTAGGTTT
```