

# Klausur Bioinformatik I, WS 03/04

## Studiengang molekulare Biotechnologie

13.2.04

### Gruppe **B**

#### Aufgabe 1

a) Agglomerative hierarchische Clusteranalyse

Die Dateninstanzen A, B, C, D und E sollen mittels *Weitesten-Nachbar-Algorithmus* (complete linkage clustering) geclustert werden.

Gegeben sei folgende Distanzmatrix:

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>
<b>A</b>	0	2	3	6	1
<b>B</b>	2	0	5	8	4
<b>C</b>	3	5	0	9	7
<b>D</b>	6	8	9	0	6
<b>E</b>	1	4	7	6	0

Skizzieren Sie das Dendrogramm, das den Algorithmusverlauf verdeutlichen soll.

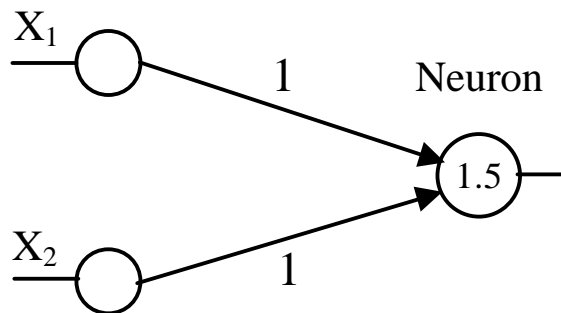
b) Gegeben seien zwei Cluster {A, E, B} und {A, B, C, D, E}.

Der zweite Cluster soll als „gold standard“ angenommen werden. Berechnen Sie das PPR-score (Product of Precision and Recall).

#### Aufgabe 2

Gegeben sei ein Neuron mit Eingaben  $X_1$  und  $X_2$ .

Die Zahlen an den Pfeilen sind die Gewichte, die Zahl im Neuron ist der Schwellwert.



- a) Berechnen Sie die Ausgaben des Neurons für die binären Eingaben  $X_1$  und  $X_2$  (tragen Sie die Werte in die Tabelle ein):

$X_1$	$X_2$	Neuron
0	0	
0	1	
1	0	
1	1	

- b) Welche Funktion realisiert das Neuron?  
 c) Warum braucht man keine verborgene Schicht um die Funktion zu realisieren? (Bitte veranschaulichen Sie Ihre Antwort mit einer Skizze).
- d) Bitte Zutreffendes auswählen (mehreres ist möglich):
- 1 Probleme beim Lernen von Neuronalen Netzen sind:
    - a) schlechte Generalisierung
    - b) gute Generalisierung
    - c) zu wenig Trainingsdaten
    - d) overfitting
    - e) cross-validation
    - f) zu wenig Iterationsschritte
    - g) zu viel Iterationsschritte
    - h) das Lernverfahren endet in einem lokalen Minima
    - i) das Lernverfahren endet in einem globalen Minima
  2. Im Laufe des Backpropagation-Algorithmus wird die Fehlerfunktion:
    - a) minimiert
    - b) maximiert
  3. Das Lernverfahren von Kohonen Netzen ist:
    - a) überwachtes Lernen
    - b) unüberwachtes Lernen
    - c) Wettbewerbslernen
- e) Bitte geben Sie einige Beispiele für die Anwendung von Neuronalen Netzen in der Bioinformatik an.

### Aufgabe 3

Geben seien die beiden Sequenzen

- 1) E I S Z E I T,
- 2) Z E I T I G.

Die beiden Sequenzen sollen mit der Methode des dynamischen Programmierens global aligniert werden.

a) Welchen Algorithmus verwendet man (Name)?

Ein globales Alignment mit folgenden Scoreparametern

- Match: +3  
Mismatch: 0  
Gap: -2

soll erstellt werden. Erstelle dazu eine Backtracking-Matrix und markiere den Pfad für das beste Alignment. Gib das optimale Alignment und seinen Scorewert an.

### Aufgabe 4

a) Gegeben sind die folgenden Datensätze:

Klasse	Länge [m]	Breite [m]
A	2	0.8
A	3	1
A	4	1
A	3	2
B	2	2.5
B	1.5	2.5
B	1	3
B	1	4
B	0.5	2

Skizzieren Sie die Situation im Merkmalsraum.

b) Zu welcher Klasse gehört der Datensatz mit Länge 1m und Breite 1.5m? Verwenden Sie den Klassifikator k-Nächster-Nachbar ( $k=3$ ), die Datensätze aus 1a) als Trainingsdaten und die Euklidische Distanz. Eine Berechnung ist unbedingt erforderlich.

c) Was versteht man unter „Overfitting“ (Überanpassung)?

## Aufgabe 5

- a) Erstellen Sie einen optimalen Entscheidungsbaum, mit dem die Pflanzenklassen A und B getrennt werden können. Eine kurze Begründung ist unbedingt erforderlich. (Die Berechnung des information gain ist **nicht** nötig!)

Klasse	Farbe	Blütenlänge [cm]	Essbar
A	Grün	4.5	Ja
A	Gelb	6.2	Ja
A	Grün	3.9	Ja
B	Gelb	12.1	Nein
B	Grün	9.5	Nein
B	Gelb	2.1	Ja
B	Gelb	3.3	Ja

- b) Eine zu diagnostizierende Pflanze hat die Farbe grün, 12.0 cm Blütenlänge und ist nicht essbar. Gehört sie gemäß dem Entscheidungsbaum aus 2a) zur Klasse A oder B? Eine Begründung ist unbedingt erforderlich.
- c) Zwei Subtypen von Brustkrebs sind mit einem Klassifikator zu 100% trennbar (0% Fehler auf den Trainingsdaten). Der Klassifikationsfehler auf den Validierungsdaten beträgt 12%. Es wird Ihnen vorgeschlagen, eine Ensemble-Methode (Boosting oder Bagging) zu verwenden. Welche der beiden Methoden würden Sie im konkreten Fall wählen? Eine Begründung ist unbedingt erforderlich.

## Aufgabe 6

- a) Die Prävalenz einer Hepatitis-C-Infektion (HCV) sei 0.5%. Die Sensitivität des Diagnostetests beträgt 85% und die Spezifität 90%. Wie hoch ist die Wahrscheinlichkeit, dass ein positiv getesteter Patient auch tatsächlich mit HCV infiziert ist?
- b) Sollte bei einer Krebsvorsorgeuntersuchung die Spezifität oder Sensitivität des diagnostischen Tests besonders hoch sein? Eine Begründung ist unbedingt erforderlich.
- c) Wie groß ist die Entropie eines diagnostischen Tests, wenn die Prävalenz der Erkrankung, auf die getestet wird, 100% beträgt? Eine Begründung ist unbedingt erforderlich.

## Aufgabe 7

(Statistik) Erklären Sie die Unterschiede zwischen Median, (arithmetischem) Mittelwert

und Erwartungswert (einer kontinuierlich verteilten Zufallsvariablen). Unter welchen Umständen zieht man den Median dem Mittelwert vor?

## Aufgabe 8

(Sequenzanalyse)

- (a) Erklären Sie das Prinzip der log-odds-Scores.
- (b) Als Häufigkeitsverteilung der Nukleotide in beliebigen Sequenzen (keine Splice Sites) nehmen Sie an:

$$\pi_A = 0.2721, \quad \pi_C = 0.2274, \quad \pi_G = 0.2278, \quad \pi_T = 0.2727$$

Tabelle 1 zeigt die beobachteten Häufigkeiten der 4 Nukleotide um die 5'-Donorstelle von Splice Sites. Berechnen Sie den log-odds-score (mit natürlichen Logarithmen) für folgende Sequenz:

caggtaaaa

Berechnen Sie den Wert auf zwei Nachkommastellen genau!

**Tab. 1:** Häufigkeitsverteilung der vier Nukleotide um die 5' Donorstelle von Splice Sites (in Prozent)

	-3	-2	-1	0	1	2	3	4	5
A	34.08	60.36	9.14	0.00	0.00	52.57	71.26	7.08	15.98
C	36.24	12.90	3.27	0.00	0.00	2.82	7.56	5.50	16.46
G	18.31	12.48	80.34	100.00	0.00	41.94	11.76	81.35	20.90
T	11.38	14.25	7.24	0.00	100.00	2.55	9.29	5.88	46.16

## Aufgabe 9

- a) i) Nenne zwei heuristische Verfahren, mit denen mehr als zwei Sequenzen aligniert werden können (Begriffe, Schlagworte).
- ii) Gib kurz jeweils einen Vorteil und einen Nachteil der beiden Verfahren an (Stichworte).
- b) Zeige, dass folgende Formeln äquivalent sind:

$$(1) \quad \mathbf{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$(2) \quad \mathbf{b}_1 = \frac{(\sum_{i=1}^n x_i y_i) - n \bar{x} \bar{y}}{(\sum_{i=1}^n x_i^2) - n \bar{x}^2}$$

wobei  $\bar{x}$  (bzw.  $\bar{y}$ ) der Mittelwert aller  $x_i$  (bzw.  $y_i$ ) ist, i.e.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i .$$