Übungen Musterklassifikation
Sommersemester 2007

Abgabe am 27.4.2007

Prof. Dr. Stefan Posch

Dr. Birgit Möller

(birgit.moeller@informatik.uni-halle.de)

Institut für Informatik

Universität Halle

Blatt 3

Eines der aktuellen Forschungsgebiete in der Biologie ist die Analyse von Genexpressionen. Ausgehend von dem Wissen, dass verschiedene Krebsarten mit der Bildung verschiedener Proteine zusammenhängen - einerseits solche, die für den Krebs verantwortlich sind, andererseits solche, die ihn bekämpfen, ist man daran interessiert herauszufinden, welche Proteine verstärkt/vermindert (die Genexpression ist also erhöht oder verringert) gebildet werden.

Dazu wird präparierte und fluoreszierend markierte mRNA aus Gewebsproben auf ein sogenanntes Microarray aufgebracht, das komplementäre Basenfolgen der zu untersuchenden Proteine enthält. Die Proteine der Gewebsprobe binden sich an die zugehörigen Basen - je nach Anzahl der gebundenen Proteine erscheint dieser Ausschnitt auf dem Microarray durch die Fluoreszenz heller oder dunkler. Der Chip wird eingescannt und der Helligkeitswert jedes Proteins wird in einen Zahlenwert umgewandelt.

In diesem Falle wurden mehrere Knochenmark-Proben mit je etwa 35403 Genen (jede Probe ein Chip) von Patienten mit Leukämie (ALL) oder Ewing (Knochenmark-Krebs), sowie gesundem Knochenmark (KM) eingescannt.

Nun soll versucht werden die Daten dieser drei Arten von Gewebsproben zu clustern. Also möglichst ähnliche Proben (Gene mit ähnlichem Experssionsverhalten) in ein Cluster zu bringen. Dabei ist zu beachten, dass man gerade bei der Verwendung von zufälligen Anfangspunkten völlig verschiedene Ergebnisse erhalten kann. Für Tests ist also auch interessant, zwischen welchen Werten der Quantisierungsfehler schwankt. Zur Vereinfachung wird nur eine Auswahl von Genen betrachtet.

Die Daten sind, neben den Übungsblättern, im Web zu finden - in GenDaten.txt die bereits klassifizierten Proben, und in ProbenXYZ.txt 2 noch unbekannten Proben.

Aufgabe 3.1 (6 Punkte) Programmieren Sie den k-means-Algorithmus mit mindestens 2 verschiedenen Distanzfunktionen. (Integrieren Sie den k-means-Algo in das bereits programmierte Gerüst und erweitern Sie das Gerüst so, dass man menu-basiert bereits eingelesene Stichproben mittels k-means klassifizieren kann. Beim Klassifizieren entsteht aus der verwendeten Stichprobe eine neu (mit (neuen) Klassenzuordnungnen)). Diese sollte ueber die bereits programmierten Funktionen ausgewählt und in eine Datei gespeichert werden köennen.) Testen Sie Ihr Programm mit der gegebenen Sammlung von Genen (GenDaten.txt). Welche Cluster werden gebildet und welche Proben werden korrekt bzw. falsch zugeordnet? Zu welchen Clustern werden die zwei unbekannten Proben am ehesten zugeordnet? Geben Sie jeweils die Quantisierungsfehler an. (Der Source-Code sollte ausreichend Kommentare enthalten, die das Verstehen des Programms erleichtern.)

Aufgabe 3.2 (3 Punkte) Gegeben sei folgendes Muster $f(k), k \in Z$ (diskrete Dreiecksfunktion):

$f(-1) = 1, f(0) = 2, f(1) = 1, f(k) = 0$, für alle sonstigen k.

Die Implusantwort $g(j), j \in Z$ eines linearen Systems T sei gegeben als "diskreter Sägezahn":

$g(-1) = 2, g(0) = 1, g(j) = 0$ für alle anderen j.

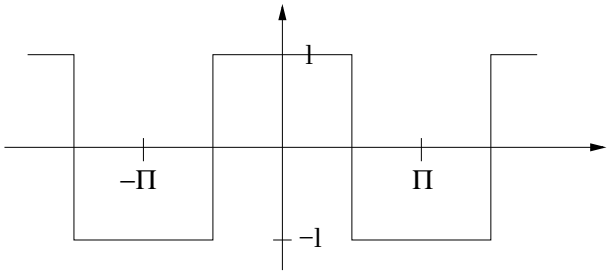
Berechnen Sie $T\{[f]\}$

Versuchen Sie den Berechnungsvorgang graphisch sichtbar zu machen.

Aufgabe 3.3 (3 Punkte) Die Koeffizienten a_ν und b_ν der Fouriertransformation können für periodische Funktionen folgendermaßen ermittelt werden:

$a_\nu = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) cos(\nu t) dt$
$b_\nu = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) sin(\nu t) dt$

Bestimmen Sie a_ν und b_ν für nebenstehende Funktion.



Die Programme bitte rechtzeitig per mail an: birgit.moeller@informatik.uni-halle.de schicken.