

9. Übung zur Vorlesung *Künstliche Intelligenz*

Institut für Informatik, FU Berlin, SoSe 2006
Prof. Dr. Raúl Rojas, Marco Block, Ernesto Tapia

Neben der schriftlichen Abgabe, sind die Programmieraufgaben **zusätzlich** per e-mail an den Tutor zu schicken. **Eine e-mail ersetzt nicht die schriftliche Abgabe!** Zur Erinnerung: Testläufe gehören zur Abgabe und werden ebenfalls bewertet. Die Trainingsdaten finden Sie auf der Webseite.

1. Aufgabe (5 Punkte) *k-nearest neighbors*

Implementieren Sie den in der Vorlesung vorgestellten *k-nearest Neighbors*-Algorithmus. Verwenden Sie als Basismenge **digits.trn** und testen Sie damit **digits.tst**.

TIPP: Wählen Sie eine geschickte Implementierung, damit sie Ihre Lösung für Aufgabe 2 verwenden können.

2. Aufgabe (13 Punkte) *k-means*

In der Vorlesung wurde der Linde-Buzzo-Gray-Algorithmus (LBG, *k-means*) vorgestellt.

a) (6 Punkte) Implementieren Sie den Algorithmus. Ihr Programm sollte die Clusteranzahl l als Input erhalten. Verwenden Sie als Trainingsmenge für die Klassifizierung die Datei **digits.trn**.

b) (6 Punkte) Erweitern Sie Ihr Programm (aus a) um die Funktionalität des *variablen k-means* und geben Sie die Erkennungsraten ($l = 10, 12, \dots, 100$) für die Testdatei **digits.tst** graphisch an.

c) (1 Punkt) Machen Sie sich Gedanken für ein geeignetes Abbruchkriterium. Wann würde Ihr Programm mit diesem Abbruchkriterium stoppen?

3. Aufgabe (8 Zusatzpunkte) *rekursiver k-means*

Sie kennen das Verfahren des rekursiven Ansatzes beim *k-means-Algorithmus*, um die Zugriffszeit auf die gelernten Cluster zu verringern.

a) (4 Punkte) Erweitern Sie Ihr Programm (aus 2.a) um folgende Funktionalität: Die Datenmenge D (**digits.trn**) sollte zunächst in l Cluster getrennt werden. Sie erhalten die Datenmengen D_1, \dots, D_l . Diese Datenmengen sollen dann rekursiv wieder in l Cluster getrennt werden. Das machen Sie bis zu einer Tiefe t .

b) (3 Punkte) Testen Sie Ihre Implementierung mit **digits.tst** und verwenden Sie für $l = 2$ und $t = 5$.

c) (1 Punkt) Angenommen Sie würden l für jeden Knoten im "Clusterbaum" variabel gestalten. Machen Sie sich Gedanken für eine geeignete Heuristik, um zu entscheiden, wie gross l sein sollte, wenn Sie $|D|$ Daten vorliegen haben mit k unterschiedlichen Klassentypen.

Abgabe: (Donnerstag) 06.07.2006, 12:00 Uhr (s.t.) in der Vorlesung
(verspätete Abgaben werden nicht mehr entgegen genommen!)