

Aufgabenblatt 1

Abgabe bis zur Übung am 09.11.2007

Neben der schriftlichen Abgabe, sind die Programmieraufgaben **zusätzlich** per E-Mail an den Tutor zu schicken. Eine E-Mail ersetzt nicht die schriftliche Abgabe.

Aufgabe 1 (10 Punkte)

In der Vorlesung wurde der *k-means*-Algorithmus vorgestellt. Implementieren Sie diesen Algorithmus und clustern Sie die Trainingsdaten (*pendigits-training.txt*) für $k = 10, 11, \dots, 30$. Verwenden Sie ein geeignetes Abbruchkriterium.

Ermitteln Sie jeweils die Erkennungsraten und Laufzeiten für die Testdaten und überlegen Sie sich eine Methode, die in Abhängigkeit von k und der Erkennungsraten entscheidet, welches k für den aktuell verwendeten Datensatz das Beste ist. Beschreiben Sie, was Sie für das Beste halten und warum. Das Zahlenformat wird in den Tutorien besprochen. Die verwendeten Datensätze finden Sie auch auf der Veranstaltungshomepage.

Aufgabe 2 (10 + 3 Punkte)

Implementieren Sie das in der Vorlesung vorgestellte Verfahren zur Klassifizierung mit Hilfe von Gaußschen Verteilungen.

Zur einfacheren Berechnung der Kovarianzmatrix und der Determinanten wird erlaubt nur die diagonalen Elemente der Kovarianzmatrix zu verwenden. Bei korrekter Berechnung der Determinanten gibt es Zusatzpunkte.

Für die Berechnung der Parameter der Gaußschen Verteilungen verwenden Sie die Daten aus *pendigits-training.txt*. Ermitteln Sie die Erkennungsraten für die Daten aus *pendigits-testing.txt*.

Aufgabe 3 (3 Punkte)

Welchem Distanzmaß ähnelt der Exponent der allgemeinen (multivariaten) Gleichung der Gaußschen Dichtefunktion? Was ergibt sich, wenn die Kovarianzmatrix die Identitätsmatrix ist? Was, wenn sie eine Diagonalmatrix mit verschiedenen Varianzen ist?

Aufgabe 4 (3 Punkte)

Vergleichen Sie k-Nearest-Neighbor, k-means und die Gauß-basierte Klassifikation bzgl. Erkennungsrate und Laufzeit.