

## 2. Übungsblatt

(Statistische Klassifikatoren)

Statt einer Normalverteilung pro Klasse bietet es sich häufig an, mehrere Verteilungen vorzusehen. Dafür ist es notwendig, geeignete Verteilungen im Datenraum zu ermitteln.

Folgen Sie den Kommentaren im Quelltext der Module *blatt2.aufg03* und *blatt2.aufg04*.

**3. Aufgabe:** Vektorquantisierung zur Initialisierung eines Verteilungsmodells.

1. Implementieren Sie einen Vektorquantisierer nach Lloyd und visualisieren Sie die Ergebnisse auf der Trainingsmenge des data2d Datensatz (klassenunabhängig). Wie lassen sich diese zur Initialisierung eines Gauß'schen Mischverteilungsmodells verwenden?
2. (*optional*) Implementieren Sie auch den  $k$ -means-Algorithmus (nach McQueen, siehe bspw. Skript Spracherkennung, S. 53 ff.). Vergleichen Sie die Laufzeit und die Qualität des erzeugten Modells mit dem Lloyd-Algorithmus.
3. Auf Grundlage der vorherigen Ergebnisse berechnen Sie nun mehrere Normalverteilungen für den data2d Trainingsdatensatz (klassenunabhängig). Visualisieren Sie die Ergebnisse. Die Verwendung wievieler Normalverteilungsdichten erscheint sinnvoll?
4. (*optional*) Visualisieren Sie für jeden Datenpunkt des data2d Trainingsdatensatz welche Normalverteilung diesen mit größter Likelihood generiert.

**4. Aufgabe:** Nutzen Sie das zuvor geschätzte Gauß'sche Mischverteilungsmodell zur Erstellung eines Mischverteilungsklassifikators.

1. Realisieren Sie einen Mischverteilungsklassifikator mit klassenunabhängigen Komponentendichten:

$$p(\mathbf{c}|\Omega_{\kappa}) = \sum_{l=1}^L p_{\kappa l} \mathcal{N}(\mathbf{c}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \quad \text{mit} \quad \sum_{l=1}^L p_{\kappa l} = 1 \quad , \forall \kappa$$

2. Welche zusätzlichen Parameter werden neben den  $p_{\kappa l}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l$  noch für die Klassifizierung benötigt?
3. Berechnen Sie die Klassifikationsfehlerraten die sich bei Verwendung der einzelnen Konfigurationen für die Teststichprobe des data2d Datensatz ergeben (jeweils Gesamtfehlerraten als auch klassenspezifisch). Verwenden Sie unterschiedliche Anzahlen von Basisdichten und visualisieren Sie diese.
4. Realisieren Sie den Mischverteilungsklassifikator mit klassenabhängigen Komponentendichten  $(\boldsymbol{\mu}_{\kappa l}, \boldsymbol{\Sigma}_{\kappa l})$ .

**Hinweise** zur Schätzung von Verteilungsmodellen:

- Bei der Auswertung und Bewertung von Verteilungsmodellen kann es zu numerischen Problemen kommen. Woran liegt das und wie lassen sich diese für die Klassifikation vermeiden?
- Wieviele Trainingsdaten werden benötigt um eine Normalverteilung zu schätzen? Welche Eigenschaften müssen diese haben? Stellen Sie sicher, dass sie die Mischverteilungsmodelle entsprechend initialisieren.