### 2.3.3 k-Means

Teile jede Klasse $\Omega_\kappa$ in $L_\kappa$ (Teil-)Gebiete, die jeweils durch ihren Schwerpunkt oder Mittelpunktvektor $\vec{\mu}_\kappa^l$ repräsentiert werden.

Im restlichen Abschnitt lassen wir die Klassenindizes $\kappa$ zur Vereinfachung weg:

- sei $\Omega = \{\vec{c}_i | i = 1, \ldots, N\}$ die (einzige) Klasse
- sie wird in $L$ disjunkte Teilgebiete $R^l$ zerlegt:
  $\Omega = R^1 \dot\cup R^2 \dot\cup \ldots \dot\cup R^L$
- jedes Teilgebiet $R^l$ wird durch seinen Schwerpunkt repräsentiert:
  $\vec{\mu}^l = \frac{1}{|R^l|} \sum_{\vec{c}_i \in R^l} \vec{c}_i$
- Ziel ist Minimierung des Quantisierungsfehlers

$$\epsilon = \frac{1}{N} \sum_{l=1}^{L} \sum_{\vec{c}_i \in R^l} d(\vec{c}_i, \vec{\mu}^l)$$

wähle aufgrund von Vorwissen oder zufällig (z.B. die ersten $L$ Vektoren der Stichprobe mit der Größe $N$) initiale Mittelpunkte $\vec{\mu}^l, l = 1, \ldots, L$

$\epsilon^0 := \infty$; der Quantisierungsfehler in der Iteration $0$

$t = 0$; Iterationszähler

- $t := t + 1, \quad \epsilon^{(t)} := 0$
- FOR alle Gebiete $l = 1, \ldots, L$
  - $N^l := 0; \quad \hat{\vec{\mu}}^l := \vec{0}$
- FOR alle Vektoren $\vec{c}_i$ der Stichprobe
  - bestimme $\vec{\mu}^l$ mit minimalem Abstand zu $\vec{c}_i$
  - $\epsilon^{(t)} := \epsilon^{(t)} + d\left(\vec{c}_i, \vec{\mu}^l\right)$
  - berechne neuen Schätzwert für den Mittelpunkt $\hat{\vec{\mu}}^l := \hat{\vec{\mu}}^l + \vec{c}_i$
  - $N^l := N^l + 1$
- $\epsilon^{(t)} := \epsilon^{(t)}/N$
- FOR alle Gebiete $l = 1, \ldots, L$
  - $\vec{\mu}^l := \hat{\vec{\mu}}^l/N^l$

UNTIL $(\epsilon^{(t-1)} - \epsilon^{(t)})/\epsilon^{(t)} \leq \varepsilon$

- k-means konvergiert “fast immer”
- es gibt keine Garantie in eine (lokales) Minimum des Quantisierungsfehlers zu kommen

  starte den -means mit unterschiedlichen Initialisierungen und nimm bestes Ergebnis

  wir werden später noch eine Variante kennenlernen, die garantiert zu einem lokalen Minimum konvergiert
- es wird (praktisch immer) der eulidische Abstand verwendet
- neben der Vektorquantisierung wird das k-means eine Verfahren oft zum Clustern benutzt:
  eine gegebene Menge von Datenpunkten soll in (eine vorgegebene Anzahl von) Cluster/Häufungsgebiete aufgeteilt werden