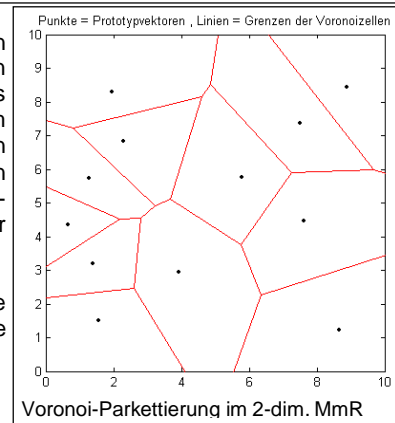


5.5. Unüberwachtes Verfahren: K-Means Klassifikator

- Vektorquantisierung: Approximation der MmV $\mathbf{x} \in \mathfrak{R}^d$ der Lernstichprobe mit deutlich weniger Prototypvektoren $\mathbf{v} \in \mathfrak{R}^d$ (Codebook vectors)
- der nächst benachbarte PrV \mathbf{v}_c repräsentiert näherungsweise den MmV \mathbf{x} : $\|\mathbf{x} - \mathbf{v}_c\| = \min_i \|\mathbf{x} - \mathbf{v}_i\|$ bzw. $c = \arg \min_i \|\mathbf{x} - \mathbf{v}_i\|$ (\mathbf{v}_c = closest prototype vector)
- der PrV \mathbf{v} kann bspw. so gewählt werden, daß für alle Vektoren \mathbf{x} der Quantisierungsfehler zu ihren nächst benachbarten PrV \mathbf{v}_c minimal wird; dieser Fehler entspricht der mittleren quadratischen Abweichung zwischen \mathbf{x} und \mathbf{v} und wird oft auch distortion measure genannt:
- $p(\mathbf{x})$ ist die Wahrscheinlichkeitsdichtefunktion der MmV; die Integration erfolgt im gesamten MmR

$$E = \int \|\mathbf{x} - \mathbf{v}_c\|^2 p(\mathbf{x}) d\mathbf{x} \approx \frac{1}{n} \sum_{c=1}^k \sum_{\mathbf{x} \in G_c} \|\mathbf{x} - \mathbf{v}_c\|^2$$

- Voronoi-Zerlegung (Voronoi-Parkettierung): wenn eine Menge von PrV \mathbf{v} gegeben ist, dann findet man an jedem Ort im MmR einen nächst benachbarten PrV \mathbf{v}_c ; bei kleiner Ortsveränderung kann es zu abrupter Änderung des zugehörigen \mathbf{v}_c kommen; alle Orte im MmR, die zu ein und demselben PrV \mathbf{v} gehören, nennt man Voronoi-zelle; ihre Grenzen lassen sich aus den mittelsenkrechten Hyperebenen zwischen zwei PrV konstruieren (Abb.). Alle Voronoi-zellen füllen den MmR lückenlos aus. Alle MmV \mathbf{x} , die in einer Voronoi-zelle liegen, bilden ihre Voronoi-menge G_c .



- für die Berechnung der PrV \mathbf{v} gibt es keine allgemeine Lösung, die den Quantisierungsfehler minimiert. Es existieren jedoch iterative Verfahren, wie bspw. der K-Means-Algorithmus: [MacQueen, 1965] [Linde, Buzo, Gray, 1980]

1.) Initialisierung von Prototypvektoren

- Anzahl der Prototypvektoren k muß festgelegt werden
- zufällige oder datengetriebene Initialisierung: $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_k\}$

2.) Merkmalsvektoren mit Nearest-Neighbour-Verfahren klassifizieren

- $\mathbf{x} \in G_c$, falls $\|\mathbf{x} - \mathbf{v}_c\| = \min_i \|\mathbf{x} - \mathbf{v}_i\|$ ($i = 1, 2, \dots, k$)
- G_c ist die Menge aller MmV \mathbf{x} , die nächste Nachbarn von \mathbf{v}_c sind (Voronoi-menge von \mathbf{v}_c)

3.) Adaptation der Prototypvektoren

- alle Prototypvektoren \mathbf{v}_i werden neu berechnet mit dem Mittelwert der MmV $\mathbf{x}_j \in G_i$:
- Fortsetzung mit Schritt 2
- Abbruch, falls sich die Mengen G_i nicht mehr verändern oder
- falls sich der Quantisierungsfehler nicht mehr wesentlich verringert
- jeder PrV liegt im Schwerpunkt der Verteilungsdichte in seinem Voronoi-gebiet

$$\mathbf{v}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in G_i} \mathbf{x}_j$$

(n_i = Anz. MmV aus G_i)

5.6. Unüberwachte Verfahren: Nonlinear Mapping (Sammons Algorithmus) [Sammon, 1969]

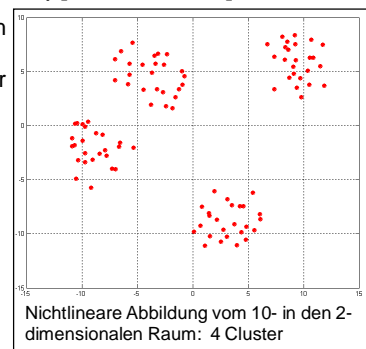
- Visualisierung der MmV-Verteilung mit einer nichtlinearen Abbildung vom hoch- in den niederdimensionalen Raum
- zu jedem MmV \mathbf{x}_i im hochdimensionalen Raum korrespondiert ein Vektor im niederdimensionalen Raum \mathbf{y}_i

- Verfahren:

- 1.) Initialisiere zufällig: $\mathbf{y}_i \forall i=1, \dots, n$
- 2.) Berechne \mathbf{y}_p und \mathbf{y}_q , die zu zwei MmV \mathbf{x}_p und \mathbf{x}_q korrespondieren, wobei p und q zufällig gewählt werden:

$$\Delta y_p = -\frac{1}{2} \lambda (y_p - y_q) \quad , \quad \Delta y_q = +\frac{1}{2} \lambda (y_p - y_q) \quad \lambda = \frac{1 - \sqrt{\frac{\|\mathbf{x}_p - \mathbf{x}_q\|}{\|y_p - y_q\|}}}{1 + \|\mathbf{x}_p - \mathbf{x}_q\|}$$
- 3.) Wiederhole 2.), falls der Fehler E einen Schwellwert

$$\text{überschreitet: } E = \frac{1}{\sum_{i < j} \|\mathbf{x}_i - \mathbf{x}_j\|} \sum_{i < j} \frac{(\|\mathbf{x}_i - \mathbf{x}_j\| - \|y_i - y_j\|)^2}{\|\mathbf{x}_i - \mathbf{x}_j\|}$$



- zu allen möglichen Paaren $\mathbf{x}_i, \mathbf{x}_j$ wird proportional distanzerhaltend versucht, Paare $\mathbf{y}_i, \mathbf{y}_j$ zu finden
- funktioniert, falls die Daten in einem Unterraum (auf einer Mannigfaltigkeit) geringer Dimensionalität liegen