MACHINE LEARNING, 7. SEMINAR – CLUSTERING

Aufgabe 1. Man betrachte einen „Fisher Klassifikator mit Radial Basis Funktionen“, der die folgende Abbildung $\mathbb{R}^n \to \{1 \dots K\}$ realisiert:

$$y = \arg\min_k (\|x - \mu_k\|^2 - r_k^2), \tag{1}$$

mit den klassenspezifischen Zentren $\mu_k \in \mathbb{R}^n$ und Radien $r_k \in \mathbb{R}$.

a) Transformieren Sie den Input-Raum $\mathscr{X}$ derart, dass diese Entscheidungsregel in dem modifizierten Raum $\tilde{\mathscr{X}}$ einem „gewöhnlichen“ Fisher Klassifikator entspricht, d.h.

$$y = \arg\min_k \langle \tilde{x}, w_k \rangle. \tag{2}$$

Wie ergeben sich die Vektoren w_k aus den bekannten Parametern μ_k und r_k des ursprünglichen Klassifikators?

b) Zum Anlernen der Entscheidungsregel (2) wird der Perceptron Algorithmus verwendet. Somit ergeben sich die Vektoren w_k im modifizierten Raum $\tilde{\mathscr{X}}$. Wie ergeben sich daraus die Parameter μ_k und r_k der ursprünglichen Entscheidungsregel (1)?

Aufgabe 2. Konstruieren Sie ein möglichst einfaches Beispiel (d.h. mit nur zwei Clustern und wenigen Datenpunkten), bei dem der K-Means Algorithmus nicht zum globalen Optimum konvergiert.

Aufgabe 3. Zeigen Sie, dass die Clusterungsaufgabe

$$\sum_k \sum_{ij \in I_k} d(i,j) \to \min_C$$

(siehe Vorlesung) der Aufgabe

$$\sum_{ij : C(i) \neq C(j)} d(i,j) \to \max_C$$

äquivalent ist. Mit anderen Worten, statt die Summe der Abstände innerhalb der Clusters zu minimieren wird die Summe der Abstände zwischen den Clustern maximiert.

Aufgabe 4.

a) Die Zeitkomplexität einer Iteration des K-Means Algorithmus ist $O(nk)$, wobei n die Anzahl der Muster und k die Anzahl der Clusters sind (siehe Vorlesung). Man betrachte Spezialfall, in dem das Merkmal eindimensional ist, d.h. $x^i \in \mathbb{R}$, $i = 1 \ldots n$. Ist es in diesem Fall möglich, die Zeitkomplexität einer Iteration zu verbessern?

Hinweis: Die Muster x_i sowie die Cluster-zentren $y_k \in \mathbb{R}$ können geordnet werden. Dadurch kann die Zeitkomplexität bis zu $O(n+k)$ verbessert werden.

b) Nehmen wir weiterhin an, dass das Merkmal ganzzahlig ist und die Anzahl der Werte die das Merkmal annehmen kann viel kleiner ist, als die Anzahl der Muster (zum Beispiel: die Muster sind Pixel eines Grauwertbildes, das Merkmal kann somit nur 256 Werte annehmen). Ist es möglich, die Zeitkomplexität weiter zu verbessern?

Hinweis: Die Lernstichprobe kann durch das Histogramm repräsentiert werden. Dadurch verbessert sich die Zeitkomplexität einer Iteration bis zu $O(m+k)$, wobei m die Anzahl der Werte ist, die das Merkmal annehmen kann.

Aufgabe 5. Der Abstand $d(i, j)$ zwischen zwei Mustern i und j ist wie folgt definiert. Gegeben sei ein Graph, dessen Knoten den Mustern entsprechen. Die Kanten dieses Graphen sind gewichtet. Der Abstand $d(i, j)$ zwischen zwei Mustern i und j ergibt sich als die Länge des kürzesten Pfades zwischen den Knoten i und j in diesem Graphen. Zeigen Sie, dass ein so definiertes Abstandsmaß eine Metrik ist.