Anlage 1: Modulkonzept zu k-Nearest-Neighbor

Förderkennzeichen:	01IS17073
Vorhabenbezeichnung:	Verbundprojekt: COSY-Entwicklung von sieben Praxis-Versuchen zum Thema Datenanalyse und Maschinelles Lernen an zwei Hochschulstandorten
Zugehörigkeit:	Modulkonzept zu Versuch 1 - Schallortung

Allgemeines:

Der k-NN Algorithmus ist ein vergleichsweise einfaches Verfahren zur Gruppenbildung aus teilweise gelabelten Daten, weshalb es sich besonders zur Einführung in das Machine Learning eignet. Dabei weist der k-NN Algorithmus ungelabelten Datensätzen das Label zu, das unter seinen k nächsten Nachbarn am häufigsten vorkommt.
Der Algorithmus vergleicht zunächst einen zu klassifizierenden Datensatz $\mathbf{x}_0 \in \mathbb{R}^n$ mit der Menge $\mathcal{M}^{\mathcal{L}} = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_n, l_n)\}$ aller n vorhandenen, mit $l_i \in \{1, \ldots, L\} = \mathcal{L}$ gelabelten Trainingsdatensätze $\mathbf{x}_i \in \mathbb{R}^n$, wobei $i \in \{1, \ldots, n\}$ und ermittelt zunächst die k nächsten, bzw. ähnlichsten Datensätze $\mathcal{N}_k(\mathbf{x}_0) \in \mathcal{M}^{\mathcal{L}}$ zu $\mathbf{x}_0$ anhand eines ausgewählten Kriteriums. Das Kriterium wird dabei dem Anwendungsfall bzw. den zu vergleichenden Daten angepasst. Im Rahmen von Versuch 1 wird die euklidische Distanz $d(\mathbf{x}, \mathbf{y})$ genutzt, welche die Distanz zweier Vektoren beschreibt und der Länge des Verbindungsvektors der beiden Vektoren $\mathbf{x}$ und $\mathbf{y}$ entspricht. [1]

k-Nearest Neighbor Algorithmus:
Algorithmus 1 stellt den für den Versuch verwendeten Pseudocode zum k-Means Clustering dar.[1]

Algorithm 1 k-Nächste-Nachbarn Klassifizierung: kNN($\mathcal{M}^{\mathcal{L}}, \mathbf{x}_0$)
Eingabe: Trainingsdaten $\mathcal{M}^{\mathcal{L}} = \{(\mathbf{x}_1, l_1), \ldots, (\mathbf{x}_n, l_n)\}$ und Testdatensatz $\mathbf{x}_0$
Ausgabe: Klasse l von $\mathbf{x}_0$

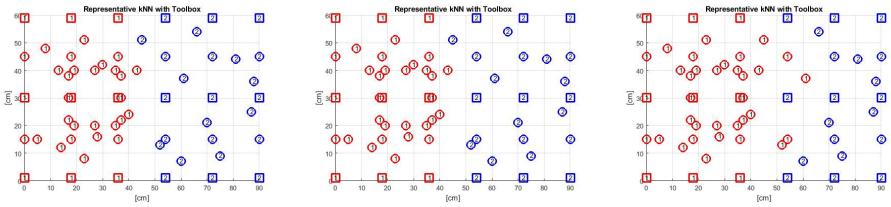
for $j = 1, \ldots, k$ **do**
  $\text{dist}_j = \infty$
  $\text{label}_j = 0$
end for
for $i = 1, \ldots, n$ **do**
  $d = d_e(\mathbf{x}_i, \mathbf{x}_0)$
  $j = 1$;
  while $j < k + 1$ **do**
    if $d < \text{dist}_j$ **then**
      for $p = k, \ldots, j + 1$ **do**
        $\text{dist}_p = \text{dist}_{p-1}$
        $\text{label}_p = \text{label}_{p-1}$
      end for
      $\text{dist}_j = d$
      $\text{label}_j = l_i$
      $j = k + 1$
    else
      $j = j + 1$
    end if
  end while
end for
return $l = \text{argmax}_{l \in \mathcal{L}}\{\ |\{i : \text{label}_i = l,\ i \in \mathcal{K}\}|\ \}$

Versuch 1: Schallortung liefert gute Daten für die Verwendung von k-NN. Aus diesem Grund bildet Versuch 1 den Rahmen für das Modulkonzept k-NN.

Aufgaben:

Die Studenten definieren Trainings- und Testmenge aus den zur Verfügung gestellten Datensätzen. Sie rufen den k-NN Algorithmus mit den Labeln der Trainingsdaten und beiden Datenmengen mit verschiedenen Konfigurationen auf, um die unterschiedlichen Ergebnisse zu beobachten und zu interpretieren. Dabei sollten die unterschiedlichen Ergebnisse bei verschiedenen Kriterien bzw. Werten für k auffallen. Die wesentlichen Aufgaben sind:

1. Aufruf der k-NN Toolbox mit unterschiedlicher Anzahl k an Nachbarn.
2. Visualisierung und Interpretation der Ergebnisse



3. Dabei sollte die Gleichheit der Ergebnisse bei den Werten 1 und 2, bzw. Unterschiedlichkeit anderen Werten für k, und die damit verbundene Fehleranfälligkeit auffallen.

4. Eigenständige Implementierung des Algorithmus mit Euklidischer Distanz.

5. Eigenständige Entwicklung der Ermittlung des Labels der kürzestens Distanz für $k > 2$ von $l = \text{argmax}_{l \in \mathcal{L}}\{\ |\{i : \text{label}_i = l,\ i \in \mathcal{K}\}|\ \}$

Ziele:

- Die Studenten kennen optionale Konfigurationen von k-NN in der Toolbox und sind in der Lage Datensätze zu Analysezwecken zu visualisieren.
- Die Studenten erkennen das Ziel von supervised learning und stellen fest, dass Schallsignale als Trainingsdaten für eine Lokalisierung verwendet werden können indem konkrete Bereiche des Tisches durch die gelabelten Daten definiert werden.
- Die Studenten verstehen die Funktionsweise des Algorithmus und sind in der Lage ihn eigenständig zu reproduzieren.

Besonderheit Durch das Finden des fehlerhaften Datensatzes soll ein Erfolgserlebnis vermittelt werden, welches den häufig trockenen Umgang mit Daten auflockert.

Literatur

[1] Matthias Dziubany, Rüdiger Machhamer, Hendrik Laux, Anke Schmeink, Klaus-Uwe Gollmer, Guido Burger, and Guido Dartmann. Machine learning based indoor localization using a representative k-nearest-neighbor classifier on a low-cost iot-hardware.