SE3

Aufgabenblatt 10

Softwareentwicklung 3: Logikprogrammierung - WS 2019/2020 - W. Menzel

Anwendungen der Listenverarbeitung II

Gesamtpunktzahl: 30 Abgabe der Lösungen bis zum 13.01.2020

Aufgabe 1: Nächster-Nachbar-Klassifikator 8 Punkte

maximale Bearbeitungszeit: 30 Minuten

Ein Nächster-Nachbar-Klassifikator ist ein beispielbasiertes Entscheidungsverfahren, das eine neue Beobachtung (Testfall) auf der Basis der Abstände zu den bereits bekannten und klassifizierten Beispielen (Trainingsdaten) einordnet. Voraussetzung dafür ist ein metrischer Raum, in dem jede Beobachtung durch einen n-dimensionalen Vektor von Messgrößen beschrieben ist. Beispielsweise lässt sich das Wetter zu einem bestimmten Zeitpunkt an einem bestimmten Ort durch die Merkmale Lufttemperatur, Luftdruck, Luftfeuchtigkeit, Windstärke und Windrichtung beschreiben. Eine Klassifikationsaufgabe könnte dann darin bestehen, vorherzusagen, wie groß der Publikumsandrang in einem beliebten Freizeitpark sein wird oder ob die Anlage z.B. wegen eines Unwetters ganz geschlossen werden muss. Die Trainingsbeispiele wären dann durch eine größere Anzahl von Beobachtungen (Merkmalsvektoren und ihre jeweilige Klassenzuordnung) aus der Vergangenheit gegeben, ein Testfall hingegen ist eine Beobachtung für den aktuellen Tag, für den es jedoch noch keine Klassenzuordnung gibt.

Da die Anzahl der Dimensionen von der konkreten Klassifikationsaufgabe abhängig ist, empfiehlt es sich, für eine wiederverwendbare Implementation Listen zur Repräsentation der Merkmalsvektoren zu verwenden. Die Trainingsdaten für eine Klassifikationsaufgabe lassen sich dann z.B. durch eine Sammlung von Fakten eines zweistelligen Prädikats `d(Merkmalsvektor,Klasse)` in der Datenbank des Prolog-Systems speichern:

```
% d(Merkmalsvektor,Klasse)
d([3,4,5],a).
d([2,4,1],a).
d([4,5,2],b).
d([3,2,5],b).
d([1,2,3],c).
d([3,3,2],c).
d([3,3,1],d).
...
```

Jede Beobachtung in den Trainings- und Testdaten kann man sich als den Endpunkt eines (Orts-)Vektors in einem (möglicherweise hochdimensionalen) Merkmalsraum vorstellen. Die Entscheidung fällt dann für die Klasse desjenigen Elements der Trainingsmenge, das in diesem Merkmalsraum den geringsten Abstand zur der neuen, zu klassifizierenden Beobachtung besitzt.

1. Überlegen Sie sich eine nichttriviale Klassifikationsaufgabe (mindestens drei Merkmale, mindestens drei Klassen) und stellen Sie dafür eine Trainingsmenge zusammen (mindestens drei Trainingsbeispiele pro Klasse und überlappende Entscheidungsgebiete im Merkmalsraum).

2. Definieren Sie ein Prädikat, das den Abstand von zwei Punkten in einem rein numerischen Merkmalsraum mit beliebiger Dimensionalität berechnet. Verwenden Sie dafür z.B. die Euklidische Distanz:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

3. Definieren Sie ein Prädikat, das eine Liste erzeugt, die für jedes Trainingsbeispiel eine Struktur mit zwei Angaben enthält: den Abstand zwischen dem Trainingsbeispiel und dem aktuellen Testbeispiel, sowie die Klassenzugehörigkeit des jeweiligen Trainingsbeispiels.

4. Modifizieren Sie das Prädikat aus Aufgabenteil 3 so, dass nur die Klassenzuordnung für dasjenige Trainingsbeispiel berechnet wird, das den geringsten Abstand zur Beobachtung aufweist.

Aufgabe 2: k-Nächste-Nachbarn-Klassifikator 14 Punkte

maximale Bearbeitungszeit: 80 Minuten

Ein Nächster-Nachbar-Klassifikator ist sehr empfindlich gegenüber untypischen Beispielen (Ausreißern) in den Trainingsdaten. Daher erweitert man die Idee zum k-Nächste-Nachbarn-Klassifikator, der als Grundlage für die Entscheidung über die Klassenzuordnung diejenigen k Trainingsbeispiele verwendet, die den geringsten Abstand zum Testbeispiel besitzen. Stimmen die Klassenzuordnungen nicht überein, wird eine Mehrheitsentscheidung getroffen.

1. Definieren Sie ein Prädikat, das die k nächsten Nachbarn einer zu klassifizierenden Beobachtung ermittelt.

2. Implementieren Sie ein Abstimmungsverfahren, das aufgrund der ermittelten Klassenzuordnungen eine Mehrheitsentscheidung herbeiführt. Bei gleicher Stimmenanzahl soll eine beliebige Entscheidung getroffen werden.

3. Untersuchen Sie anhand von geeigneten Beispieldaten, wie die Wahl von k das Entscheidungsverhalten Ihres Klassifikators beeinflusst.

Bonus: Erweitern Sie das Abstimmungsverfahren aus Teilaufgabe 2 so, dass bei gleicher Stimmenanzahl die minimale Gesamtdistanz der jeweiligen Punktmengen zur Entscheidung herangezogen wird. (6 Punkte)

Aufgabe 3: Normalisierung 8 Punkte

maximale Bearbeitungszeit: 40 Minuten

1. Bei Klassifikatoren, die auf einem Abstandsmaß beruhen, muss sichergestellt sein, dass die Werte in den verschiedenen Dimensionen eines Merkmalsvektors in vergleichbaren Größenordnungen liegen. Treten hier extreme Unterschiede auf, wie beispielsweise zwischen Lufttemperatur (–30 ... +50 Grad Celsius) und Luftdruck (900 ... 1100 mbar), besteht die Gefahr, dass das Klassifikationsergebnis vor allem durch die Dimensionen mit den größeren Wertebereichen dominiert wird. Aus diesem Grunde sollten die Merkmalswerte in einem Vorverarbeitungsschritt skaliert werden (Normalisierung bzw. Standardisierung). In der Statistik verwendet man dafür oftmals die Z-Transformation.

   Seien μ der Mittelwert und σ die Varianz einer Zufallsvariablen X, so berechnet sich die Z-Transformierte zu

   $$Z = \frac{X - \mu}{\sigma}$$

   Die resultierende Zufallsvariable hat dann einen Mittelwert von Null und eine Varianz von Eins. Mittelwert und Varianz für n gegebene Zahlenwerte $x_1, \ldots x_n$ berechnen sich zu

   $$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{bzw.} \quad \sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

   Schreiben Sie ein Prädikat, das die Trainingsdaten für Ihren Klassifikator mit Hilfe der Z-Transformation normalisiert, und stellen Sie eine Normalisierungsfunktion für die Vorverarbeitung der Testbeispiele bereit.

Bonus: Untersuchen Sie, ob sich durch die Normalisierung das Verhalten Ihrer Klassifikatoren aus Aufgabe 1 bzw. 2 verändert. (2 Punkte)