Verhaltensklassifikation

Ausgangssituation und Zielsetzung

Bei dieser Aufgabe geht es um die Klassifikation von dynamischen Verhaltensmustern. Wie schon bei Aufgabe 2 tun wir dies auf Basis echter Beobachtungsdaten und wieder ist der Segelfisch im Spiel: Das Verhalten eines Segelfisches beim Angriff auf einen Sardinenschwarm soll vom Angriffsverhalten des gestreiften Marlins, eines verwandten Fisches mit ähnlichem Körperbau, unterschieden werden.

In einem biologischen Forschungsprojekt wurden folgende Aktionen ermittelt, aus denen sich die Angriffssequenzen beider Fischarten zusammensetzen:

|  |  |
| --- | --- |
| *approach* | Annäherung an den Sardinenschwarm |
| *bill\_use* | Einsatz des Rostrums |
|  |  |
|  |  |
| *prey\_contact* | Direkter Kontakt des Rostrums mit der Beute |
| *open\_mouth* | Öffnen des Mauls |
| *ingest* | Fressen einer Sardine |
| *leave* | Beenden des Angriffs |
|  |  |

Jede Angriffssequenz lässt sich mit folgendem regulären Ausdruck beschreiben:

*approach* (*bill\_use* | *prey\_contact* | *open\_mouth* | *ingest*)\* *leave*

Beispiel: *approach  bill\_use  prey\_contact  open\_mouth  bill\_use  prey\_contact  leave*

Für das Training des Klassifikators sind die Dateien *Sailfish\_train.txt* und *Marlin\_train.txt* zu verwenden und für die Evaluation die Dateien *Sailfish\_eval.txt* und *Marlin\_eval.txt*. Alle vier Dateien besitzen denselben Aufbau. Die Zeilen enthalten jeweils eine Angriffssequenz.

***Bitte beachten****: Alle Datensätze enthalten****echte****Daten, die nur für diese Lehrveranstaltung verwendet und nicht weitergegeben werden dürfen!*

Konkrete Aufgabenstellung

Es ist eine Software zu entwickeln, die anhand einer Angriffssequenz, bestehend aus den oben aufgeführten Aktionen, entscheiden kann, ob es sich um den Angriff eines Segelfisches oder den eines Marlins handelt. Für die Lösung ist eine Bayes'sche Klassifikation zu verwenden (s. Skript, Folien 103 -105), bei der die Wahrscheinlichkeit P(b|c) durch eine Markovkette erster Ordnung modelliert wird (s. Skript, Folien 111 -121). Zur Vereinfachung darf davon ausgegangen werden, dass die Vertreter beider Klassen („Segelfisch“ und „Marlin“) gleichhäufig auftreten.

Ermitteln Sie den Prozentsatz der korrekt klassifizierten Sequenzen (insgesamt und separat für jede der beiden Evaluationsdateien).

Die Software ist in Java zu implementieren. Abzugeben sind der Quellcode der Java-Klassen sowie eine kurze schriftliche Darstellung der Ergebnisse als PDF-Dokument.

**Weitere Hinweise**

Denken Sie an eine software- und programmiertechnisch korrekte Form der Implementierung. Denken Sie auch an den angemessenen Einsatz von Datenstrukturen und Algorithmen. Zu einer tadellosen Lösung gehört auch eine aussagekräftige und vollständige Kommentierung des Quelltextes. Die Kommentierung sollte im Idealfall javadoc-fähig sein.

Die Bearbeitung der Aufgabe soll in Zweierteams erfolgen. Es wird vorausgesetzt, dass sich beide Teampartner mit dem Lösungsweg und auch mit der konkreten Implementierung der Lösung bestens auskennen.

Packen Sie alle zu Ihrer Lösung gehörenden Klassen sowie das PDF-Dokument in ein Zip-Archiv, benennen Sie es nach dem Schema vorname1.nachname1-vorname2.nachname2-ISys3.zip (Beispiel: stefan.krause-michael.breuker-ISys3.zip) und laden Sie das Archiv rechtzeitig im Lernraum-Kurs hoch. Verspätete Abgaben werden nicht berücksichtigt..