Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Projekt Teil 1 - Abgabe bis 30.11.2017 – 23:59

Ziel des Projektes ist es, praktische Erfahrungen im Maschinellem Lernen zu sammeln. Hierzu sollen mehrere Projektaufgaben mit Hilfe des Machine Learning Frameworks *Weka* (`http://www.cs.waikato.ac.nz/ml/weka/`) gelöst werden. Des Weiteren folgt im späterem Verlauf eine Implementierungsaufgabe. Das Projekt soll in Kleingruppen von Studierenden bearbeitet werden. Die Anmeldung, Bildung von Gruppen und Abgabe finden online in Moodle statt (`https://moodle.informatik.tu-darmstadt.de/course/view.php?id=346`). Die Abgaben erfolgen in Moodle bis zu den jeweiligen Stichtagen. Folgendes ist zu den Abgaben zu beachten:

- **Format:** Die Abgabe soll eine einzelne **Präsentation** im **PDF-Format** sein (z.B. 4:3 Querformat). Der Umfang einer Abgabe sollte 15 Folien nicht überschreiten.
- **Inhalt** Die Präsentation soll **selbsterklärend** sein. Das bedeutet, dass die Lösung ohne mündliche Erklärung nachvollziehbar ist. Das Dokument muss genügend Erläuterungen und Ausführungen enthalten. Eine reine Ansammlung von Grafiken und Tabellen ohne jegliche Begleittexte ist hierfür nicht ausreichend. Die **Bewertung** findet allein anhand der PDF-Datei statt (abgesehen von der Implementationsaufgabe).
- **Abbildungen** Tabellen, Diagramme, etc. müssen **vollständig beschriftet** sein, d.h. sie müssen zumindest direkt an der Abbildung eine kurze Beschreibung enthalten und ausreichend kommentiert sein. Insbesondere sollte bei Abbildungen von Resultaten der verwendete Datensatz angegeben werden.

  Für einen Klausurbonus ist es nicht zwingend nötig alle Aufgaben zu bearbeiten. Bei Teilabgaben kann noch ein entsprechender Teilbonus erreicht werden.

  In den jeweiligen Aufgaben verwenden Sie eine vorgegebene Anzahl von Datensätzen, die Sie der Sammlung von Klassifikations- und Regressionsdatensätzen unter `http://www.ke.tu-darmstadt.de/lehre/ws-17-18/mldm/projekt/datasets.zip` entnehmen, falls die Datensätze nicht in der Aufgabenstellung festgelegt werden. Achten Sie hierbei bitte darauf, möglichst unterschiedliche Datensätze zu wählen und diese in den einzelnen Aufgaben zu variieren. Bei der Auswahl der Datensätze ist weiterhin zu beachten, dass bestimmte Lernverfahren nicht mit allen Datensätze umgehen können. Optional können Sie dieses Problem beheben, indem Sie die Daten vorverarbeiten, z.B. mittels `FilteredClassifier` und einem entsprechendem Preprocessing-Filter. In diesem Fall, bzw. falls Sie die Standardeinstellungen in Weka modifizieren, geben Sie dies bitte in ihrer Lösung an.

Aufgabe 1 Regellernen: Anwendung und Interpretation (3 Punkte)

In dieser einführenden Aufgabe sollen Sie die Verwendung von WEKA erlernen und dabei die Ergebnisse dreier Regellerner auf drei unterschiedlichen Datensätzen vergleichen. Wenden Sie hierzu die Regellerner `ConjunctiveRule`, `JRip` und `Prism` auf 3 Klassifikationsdatensätze an. Benutzen Sie hierfür keine Cross-validation, sondern *training set* als Test-Option.

a) Vergleichen Sie die Anzahl der Regeln, der Bedingungen und der vorhergesagten Klassen der resultierenden Regelmengen jeweils in Bezug auf

   - die einzelnen Datensätze
   - die jeweiligen Regellerner

b) Existiert bei allen Algorithmen eine Default-Rule? Wenn ja:

   - Welche Klasse wird üblicherweise als Default-Rule ausgewählt?
   - Wie interpretieren Sie die Güte dieser Default-Rule?

c) Läßt sich anhand der vorherigen Teilaufgaben eine Aussage treffen, welche der drei Datenmengen am leichtesten bzw. am besten zu lernen ist?

d) Vergleichen Sie die Regelmengen der Algorithmen `JRip` und `Prism` für den Datensatz `contact-lenses.arff`. Wie schätzen Sie die Allgemeinheit der von `JRip` und `Prism` gefundenen Regeln ein? Beachten Sie hierbei, daß `JRip` als Heuristik Information Gain und `Prism` Precision verwendet.

Aufgabe 2 Evaluation von Regellernern (4 Punkte)

In dieser Aufgabe sollen unterschiedliche Evaluierungsmethoden unter Verwendung von WEKA eingesetzt und deren Ergebnisse diskutiert werden. Wenden Sie den Regellerner `JRip` auf 5 Datensätze an. Teilen Sie hierzu jeden Datensatz zunächst in 2 gleich große, stratifizierte Teile, einer Trainingsmenge und Validierungsmenge, auf. Eine stratifizierte Aufteilung kann mit dem Filter `StratifiedRemoveFolds` erreicht werden.

a) Trainieren Sie nun `JRip` auf jeder dieser Trainingsmengen (ggf. auf Teilen dieser Mengen, siehe Cross-Validation usw.) und evaluieren Sie die Genauigkeit (prozentualer Anteil der korrekt klassifizierten Beispiele) der resultierenden Klassifizierer jeweils mittels:

   - 1x5 Cross-Validation
   - 1x10 Cross-Validation
   - 1x20 Cross-Validation
   - Leave-One-Out
   - und auf der Trainingsmenge

   Wie schätzen Sie die Qualität der erhaltenen Genauigkeitsabschätzungen ein?

   Anmerkung: *In dieser Teilaufgabe sollen vorerst keine Veränderungen an weiterführenden Einstellungen, wie z.B. andere Random-Seeds, vorgenommen werden.*

b) Wiederholen Sie Aufgabe a) mit dem Unterschied, daß Sie nun eine 10x10 Cross-Validation zur Evaluation verwenden sollen. Wenden Sie hierzu zehnmal eine 1x10 Cross-Validation mit 10 unterschiedlichen Random-Seeds an und mitteln die erzielten Genauigkeiten.

   Vergleichen Sie die so erzielte Genauigkeitsabschätzung mit den Abschätzungen aus der Aufgabe a).

   Führt ihrer Meinung nach eine geschickte Auswahl von Random-Seeds zu einer besseren Abschätzung?

c) Bestimmen Sie die Genauigkeit auf der Validierungsmenge (d.h. verwenden Sie diese als Testmenge). Wie schätzen Sie nun unter der Annahme, daß es sich bei der Validerungsmenge um einen realen Anwendungsfall handelt, die Abschätzungen der Evaluierungsmethoden aus den Aufgaben a) und b) ein?

Aufgabe 3 ROC-Kurven (2 Punkte)

a) Vergleichen Sie für einen ausgewählten Klassifikationsdatensatz die ROC-Kurven und die Fläche unter diesen Kurven für die Klassifizierer `J48` und `NaiveBayes`. Sie können die ROC-Kurven betrachten, indem Sie mit der rechten Maustaste im Fenster “Result List” den Menü-Punkt “Threshold List” auswählen.

b) Interpretieren Sie die Resultate. Sie können die Werte, die zum Zeichnen der Kurve verwendet wurden, auch mit “Save” in ein ARFF-File exportieren, und dieses (nach Löschen des Headers) in Grafik-Programme importieren. So können Sie z.B. beide Kurven (für `J48` und `NaiveBayes`) übereinander legen.