Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Projekt Teil 2 - Abgabe bis 30.12.2017 – 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 Entscheidungsbäume (3 Punkte)

a) Wählen Sie 2 Klassifikationsdatensätze aus. Vergleichen Sie für diese Datensätze die ROC-Kurven und die Fläche unter diesen Kurven für die Klassifizierer `J48` einmal mit und einmal ohne Pruning (Option 'unpruned') und `ID3`. Bei `J48` verwenden Sie für die anderen Optionen die Default-Werte.

b) Vergleichen Sie die Klassifizierer ebenfalls mit den Accuracy-Werten der Cross-Validation.

c) Betrachten Sie auch die Größe der entstandenen Bäume (Anzahl Knoten und/oder Blätter im Baum) und setzen Sie diese in Zusammenhang mit der Güte der Klassifizierer.

Aufgabe 2 Nearest Neighbor (2 Punkte)

a) Verwenden Sie für diese Aufgabe die gleichen Datensets wie in der vorherigen Aufgabe. Finden Sie heraus, für welches $k \in \{1, 3, 5, 7, 9, 11\}$ der Algorithmus `k-NN` (in `weka` heisst der Algorithmus `IBk`; verwenden Sie auch hier die Default-Optionen) die höchste Cross-Validation-Accuracy bekommt.

Aufgabe 3 Regressionsbäume (3 Punkte)

Benutzen Sie die 5 Regressionsdatensätze für diese Aufgabe (außer dem Datensatz `regression`). Für nominale Attribute beachten Sie bitte, dass der Lerner `M5P` eine Binarisierung der Daten vornimmt ($A = a <= 0.5$ bedeutet also: alle Instanzen, für die A *nicht* den Wert a hat). Die Gesamtanzahl der Instanzen ist n, der tatsächliche Wert einer Instanz j ist y_j und der vorhergesagte Wert einer Instanz j ist r_j (genau wie im Skript).

a) Vergleichen Sie den *Mean Absolute Error* ($\frac{1}{n} \cdot \sum_j |y_j - r_j|$) und den *Root Mean Squared Error* ($\sqrt{Mean\ Squared\ Error}$) (10 CV), sowie die Modelle (Interpretierbarkeit/Größe) jeweils für den Regressionsbaumlerner `M5P`, einmal mit angeschaltetem Pruning und einmal ohne Pruning (Benutzen Sie Regressionsbäume, also setzen Sie die Option 'buildRegressionTree' auf 'True'). Bringt Pruning bei Regressionstasks eine Verbesserung?

b) Verwenden Sie nun Model Trees (Option 'buildRegressionTree' auf 'False' setzen, ansonsten Default Optionen). Vergleichen Sie die Model Trees mit den Regressionsbäumen.